UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

PROYECTO FIN DE CARRERA

IDENTIFICACIÓN DE HABLANTES A PARTIR DE TRAYECTORIAS TEMPORALES EN UNIDADES LINGÜÍSTICAS SOBRE GRANDES BASES DE DATOS

Ingeniería de Telecomunicación

Fernando Manuel Espinoza Cuadros
Septiembre 2012

IDENTIFICACIÓN DE HABLANTES A PARTIR DE TRAYECTORIAS TEMPORALES EN UNIDADES LINGÜÍSTICAS SOBRE GRANDES BASES DE DATOS

AUTOR: Fernando Manuel Espinoza Cuadros
TUTOR: Joaquín González Rodríguez

Área de Tratamiento de Voz y Señales
Dpto. de Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Septiembre 2012

Resumen

En este proyecto de fin de carrera se presenta un nuevo tipo de parametrización basado en el modelado de contornos temporales en unidades lingüísticas, TCLU (por sus siglas en inglés *Temporal Contours in Linguistic Units)* [González-Rodríguez, 2011] [González-Rodríguez et al., 2012]. Esta técnica ha sido inspirada en los últimos trabajos en identificación forense de locutor, cuya eficiencia ha sido ampliamente demostrada en los últimos años.

En primer lugar, se presenta el estado del arte de los sistemas de reconocimiento de locutor. El trabajo se centra en la técnica de modelado GMM-UBM (Gaussian Mixture Models) [Reynolds et al., 2000], técnica consolidada en el reconocimiento de locutor independiente de texto. Adicionalmente, se presenta las técnicas empleadas en este proyecto para extraer la información distintiva de la identidad presente en ella, y las herramientas necesarias para la evaluación del rendimiento del sistema, como por ejemplo: DETS, Cllr, EER y minDCF.

Posteriormente, se presenta una descripción del sistema que se ha implementado. El sistema de reconocimiento presenta configuraciones en función de parámetros como: modelado, número de mezclas gaussianas, tipos de parametrización de características extraídas de la señal de voz, conjunto de datos de entrenamiento, tipo de unidades lingüísticas, así como, las técnicas de fusión empleadas.

A continuación, se llevan a cabo diversas pruebas con el fin de obtener resultados objetivos y poder evaluar el rendimiento del sistema bajo diferentes condiciones. Mediante la aplicación de herramientas y valores numéricos (DETS, Cllr, minCllr, scores y minDCF) se mide el rendimiento del sistema. También, mediante las técnicas de fusión y normalización de scores o puntuaciones se pretende mejorar el rendimiento del sistema. Los experimentos se llevarán a cabo siguiendo el protocolo de evaluaciones NIST (National Institute of Standards and Techonology). El sistema de reconocimiento de locutor se evaluará sobre el protocolo NIST SRE (Speaker Recognition Evaluation) 2006.

Finalmente, se presentan las conclusiones extraídas a lo largo del trabajo junto con las propuestas del trabajo futuro.

Los resultados obtenidos en este proyecto fin de carrera han sido publicados en un artículo: Franco-Pedroso, J., Espinoza-Cuadros, F., González-Rodríguez, J., "Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition", 2012; el cual ha sido aceptado y está a la espera de ser publicado en el Congreso IberSpeech 2012.

Palabras claves

Reconocimiento de locutor, TCLU, MFCC, formantes, GMM-UBM, *Factor Analysis,* NIST, SRE, DETS, Cllr, contorno temporal, unidad lingüística, calibración.

Abstract

In this M.Sc. Thesis, we present new methods and the state of the art of the existing techniques for speaker recognition. Our work is focused on new approach to automatic speaker recognition based on modeling of Temporal Contours in Linguistic Units (TCLU) [González-Rodríguez, 2011] [González-Rodríguez et al., 2012]. This new approach has been inspired in successful work in forensic speaker identification.

In order to achieve this goal, firstly, we report a study of the state of the art of the techniques regarding automatic speaker recognition. The implementation of the speaker recognition system is performed by means of using the GMM-UBM (Gaussian Mixture Models) [Reynolds et al., 2000], which has largely dominated speaker recognition in text independent. Moreover, we report feature extraction process of the speech signal, system behavior with respect to some variables and tools used for evaluating the system performance: DETS, Cllr, EER and minDCF.

Secondly, we report the system description. The system has different setups and different parameters, which set an explicit system, these parameters are: different number of Gaussian of the model, different features of the parameterized voice, data training set and different linguistic units. Further, we report the fusion techniques.

Along the experimental part of the master's thesis several experimental results showing the system behavior are reported. The evaluation has been performed by means of using tools and numeric values (DETS, Cllr, minCllr, EER and minDCF). Furthermore, the system performance has been improved by means of using some techniques (normalization and fusion of scores).The experiments have been performed using the evaluation protocols proposed by NIST *(National Institute of Standards and Technology).* Speaker recognition system will be evaluated over NIST SRE *(Speaker Recognition Evaluation)* 2006 protocol.

Finally, conclusions are drawn, and future lines of work are proposed.

Results from this master's thesis have been published in the paper: Franco-Pedroso, J., Espinoza-Cuadros, F., González-Rodríguez, J. , "Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition", 2012; which has been accepted and it is waiting for being published in IberSpeech 2012 congress.

Key words

Speaker recognition, TCLU, MFCC, formants, GMM-UBM, *Factor Analysis,* NIST, SRE, DETS, Cllr, temporal contour, linguistic units, calibration.

A mis padres y a mi hermana.

El talento es algo bastante corriente. No escasea la inteligencia, sino la constancia.

Doris Lessing

Agradecimientos

Se va cerrando una etapa muy importante de mi vida. Esta empresa que empezó ya hace cinco años, con grandes sueños y retos, en un principio inimaginables, pero que con el paso del tiempo, mucho sacrificio y persistencia han ido completándose poco a poco.

En primer lugar, me gustaría dar las gracias a mi tutor, Joaquín González Rodríguez, por haberme dado la oportunidad de realizar el presente proyecto en el Área de Tratamiento de Voz y Señal (ATVS), así como su encomiable esfuerzo y disponibilidad durante la realización del proyecto.

En segundo lugar, agradecer a todos los ATVsianos: Sandra, Miriam, Javi González (muchas gracias por tu gran ayuda y tus consejos), Daniel Ramos, Marta, Pedro, Rubén Vera, Sara, Álvaro, Ram, Ruifang, Tousef, Silvia y Rubén Tazo. Soy muy afortunado el haber tenido la oportunidad de trabajar con todos vosotros, sois los mejores. No quiero olvidar de esta lista a Javier Franco, mi segundo mentor y guía en este proyecto. Mi especial agradecimiento por todo tu apoyo y su entera disponibilidad brindada en todo este tiempo. He aprendido mucho, tanto en lo profesional, como en lo personal. Muchas gracias Javi, ¡eres grande!.

De la misma forma, no quiero olvidarme de dos grandes persona, compañeros de batalla en estos cinco años de estudio. Darwin y Fabricio, muchas gracias por dejarme compartir grandes experiencias y anécdotas llenas de alegrías, victorias y derrotas, pero lo hemos logrado, ¡ya estamos aquí!. ¡Beell…!

No quiero terminar esta lista sin olvidar a otra gran persona, Manuel Crisol. Muchos dicen que la verdadera amistad se consigue con el paso de muchos años. Sin embargo, en este corto tiempo he aprendido el valor de la verdadera amistad. No olvidaré las grandes charlas que se hacían interminables en el tiempo. Tus consejos me han servido para poder ver la vida de muchas otras perspectivas. Muchas gracias amigo.

También quiero agradecer a toda mi familia por el apoyo que me han brindado durante todo este tiempo. En especial a mis abuelos, por todas las enseñanzas, consejos y valores que me inculcaron durante todo este tiempo.

Pero esta etapa no llegaría a su fin sin el apoyo de dos personas esenciales en mi vida: mis padres. Quiero expresarles mi gran e interminable agradecimiento por todo lo dado. Sé que todo este agradecimiento es poco por todo lo que han hecho por mí, pero les debo todo. Todo lo que soy lo es por ustedes. Y no quiero olvidarme de Silvana, mi hermana. Gracias por todo tu apoyo, especialmente, por hacerme reír en los días muy difíciles.

Gracias a todos.

Fernando Espinoza Cuadros
Septiembre 2012

Este proyecto ha sido realizado en el Área de Tratamiento de Voz y Señales (ATVS) en la Escuela Politécnica Superior de la Universidad Autónoma de Madrid. El proyecto ha sido financiado parcialmente por el Ministerio de Educación y Ciencia a través del proyecto TEC 2009-14179-C02-01.

Índice de contenidos

Índice de figuras

Índice de tablas

Glosario de términos

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 pero que utiliza únicamente números reales.

DFT *Discrete Fourier Transform* (transformada discreta de Fourier). Función de transformación ampliamente empleada en tratamiento de señales y campos afines para analizar las frecuencias presentes en una señal muestreada.

DET *Detection Error Trade-off* (compensación por error de detección). La curva DET se emplea para representar de forma gráfica el rendimiento de un sistema de reconocimiento biométrico para los distintos puntos de trabajo posibles. Se obtiene mediante un cambio de escala en los ejes de la curva ROC.

EER *Equal Error Rate* (tasa de error igual). Tasa de error, en los sistemas de reconocimiento biométrico, en que se igualan las tasas de falsa aceptación y falso rechazo.

FA *Factor Analysis* (análisis de factores). Técnica empleada en reconocimiento de locutor para el modelado explícito de la variabilidad inter-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 *Fast Fourier Transform* (transformada rápida de Fourier). Algoritmo para la implementación rápida 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.

Locución En el ámbito del reconocimiento de locutor, se emplea con el significado de "señal de audio utilizada como rasgo biométrico para la obtención de un patrón de referencia o como rasgo de test en el proceso de identificación o verificación".

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.

NAP Nuisance Attributte Projection. Técnica de compensación de variabilidad aplicada a los sistemas basados en SVM.

NIST *National Institute of Standards and Technology*(Instituto Nacional de Estándares y Tecnología de los Estados Unidos de América).

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.

Score Puntuación obtenida por un sistema de reconocimiento biométrico en la comparación entre un patrón de referencia y un rasgo biométrico de test.

Scoring Proceso de obtención de scores.

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.

TCLU Temporal Contour in Linguistic Units (Contornos temporales en unidades lingüísticas).

TCLU-MFCC Contornos temporales de coeficientes cepstrales.

Trial Juicio o comparación entre un rasgo de test y un patrón de referencia.

Trial de impostor Comparación entre un patrón de referencia y un rasgo de test cuya identidad NO se corresponde con la de aquél.

Trial genuino Comparación entre un patrón de referencia y un rasgo de test cuya identidad SÍ se corresponde con la de aquél.

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.

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.

1. Introducción

1.1 Motivación del proyecto

Hoy en día, el reconocimiento de voz es una de las tecnologías biométricas muy usadas debido: al gran avance en las Tecnologías de la Información y las Comunicaciones (TIC); una mayor aceptación social dado que se trata de una técnica no invasiva y cuya adquisición no supone grandes costes; a la aparición de aplicaciones móviles que basan su funcionamiento en la voz; y, especialmente, a la gran cantidad de información que contiene la voz como rasgo biométrico: identidad del locutor, idioma, edad, estado de ánimo, aspecto emocional, etc.

Existen dos áreas en el reconocimiento de voz: reconocimiento del habla y reconocimiento de locutor. En el área reconocimiento del habla, el objetivo es extraer el contenido lingüístico de una locución. Por otra parte, en el área de reconocimiento de locutor, el objetivo se centra en la extracción de la identidad del locutor que pronuncia una locución. El presente proyecto se basa en el área de reconocimiento de locutor.

Para el reconocimiento de locutor se usa diversas características extraídas a partir de la voz. Estas características se clasifican en función a su interpretación física: nivel alto, nivel prosódico, nivel prosódico y acústico, y nivel acústico (2.3.1).

En las dos últimas décadas, los sistemas de reconocimiento de locutor se han centrado en el nivel acústico/espectral debido a la precisión y eficiencia en reconocimiento mediante técnicas de modelado: GMM-UBM, SVM, i-vectors y PDLA. Sin embargo, la información lingüística (alto nivel) de las características extraídas no existe en este nivel.

El uso de características de alto nivel (lingüístico y fonético) para el reconocimiento de locutor [Shireberg, 2007], han demostrado propiedades tales como: alto poder discriminativo en el reconocimiento, interpretabilidad, aceptación, y, lo más importante, presenta un alto potencial en la combinación con sistemas espectrales a corto plazo (nivel acústico).

Los últimos trabajos basados en la combinación de niveles características (lingüístico, fonético y acústico) y en el uso de contornos temporales en unidades lingüísticas (Temporal Contours in Linguistic Units, TCLU) han demostrado notables resultados mediante técnicas de modelado y scoring tales como: Funciones multivariantes gaussians (MVN) y Funciones multi-variantes con kernels gaussianos (MVK), usando como características locuciones segmentadas en unidades lingüísticas: fonemas, difonemas, trifonemas, silabas y palabras [González-Rodríguez et al., 2012].

Por otra parte, hoy en día el reconocimiento automático de locutor hace frente a dos problemas: la compensación de sesión de variabilidad (2.6.4.1) mediante el uso de supervectores (2.6.2.3) y el análisis de variables latentes [Dehak et al, 2011]; y la obtención de relaciones de verosimilitud o LRs (Likelihood Ratio) (2.5.1) calibrados por trial independientes de aplicación [Brümmer and Preez, 2006], el cual permite obtener información acerca de la identidad del locutor a reconocer. La obtención de los LRs ha hecho que los sistemas de reconocimiento de locutor en texto independiente funcionen de manera eficiente en condiciones controladas, como en evaluaciones NIST (3.4.1.1 Evaluación NIST). Sin embargo, dado el ruido producido por el canal u otros factores de variabilidad provoca errores en el reconocimiento. Estos errores se encuentran en las locuciones recogidas por las bases de datos que se usarán para el entrenamiento y testeo del sistema. Por tanto, se obtendrán valores de LR mal calibrados en función de dos muestras de locución provenientes de esa base de datos. Esto obliga a que las bases de datos sean muy grandes con el fin de poder recoger mayor variabilidad de las locuciones y así implementar un sistema suficientemente robusto frente a todo tipo de variabilidad, sin embargo, es imposible generar una gran base de datos que recoja todo tipo de variabilidad.

Ante este problema surge el nuevo enfoque de análisis por unidades lingüísticas, el cual es empleado en el análisis acústico forense. Esto permite obtener valores de LR calibrados de locuciones por unidad lingüística analizada. Además, mediante la combinación de diferentes unidades los sistemas de reconociendo adquieren alta capacidad de discriminación.

Por tanto, la motivación de este proyecto se centra en el estudio de trayectorias temporales en unidades lingüísticas, resultado de la combinación resultante de características de alto nivel (nivel lingüístico y fonético) y características espectrales a corto plazo (nivel acústico). De esta forma, se pretende desarrollar un sistema de reconocimiento automático de locutor enfocado hacia el reconocimiento forense.

1.2 Objetivos del proyecto

El presente proyecto presenta un nuevo enfoque en el uso de contornos temporales en unidades lingüísticas (TCLU) [González-Rodríguez, 2011] para el reconocimiento automático de locutor. El estudio se centrará en el análisis y evaluación de un sistema de reconocimiento de locutor a partir de contornos temporales en distintas unidades lingüísticas (fonemas, difonemas y trifonemas) mediante la aplicación de modelos de mezclas de gaussianas (GMMs por sus siglas en inglés, Gaussian Mixture Models).

Por tanto, el proyecto enfoca tres objetivos:

- Estudiar los sistemas de reconocimiento automático de locutor basado en modelos de mezclas de gaussianas (GMMs).

- Desarrollo de un sistema de reconocimiento automático de locutor. El sistema consta de varias etapas: procesado del audio, extracción de características, medidas de similitud entre las entidades en dos fragmentos de voz, cálculo de puntuaciones (*scores*).

- Evaluación del sistema mediante la realización de diferentes pruebas. Las pruebas consisten en evaluar el rendimiento del sistema usando la técnica GMM-UBM. También se evaluará el rendimiento del sistema usando distintos tipos de parámetros que caracterizan la señal de voz, tales como: parámetros cepstrales, trayectorias temporales de parámetros cepstrales y trayectorias temporales de formantes. Para el caso de unidades lingüísticas, se evaluará fonemas, difonemas y trifonemas. En función de los resultados finales se intentará mejorar el sistema mediante la aplicación de técnicas de fusión de sistemas, normalización y calibración de puntuaciones.

1.3 Metodología

El desarrollo del proyecto se divide en las siguientes fases:

- **Documentación**: En la primera fase del proyecto, el alumno ha estudiado la literatura sobre el estado del arte actual en biometría, técnicas de reconocimiento de locutor independiente de texto, GMM-UBM, así como documentación sobre la base de datos que se utilizará (NIST SRE 2006, NIST SRE 2005 y NIST SRE 2004).

- **Estudio del software:** En un principio, el alumno se ha familiarizado con el software desarrollado por el grupo ATVS y los toolkits Matlab necesarios para el desarrollo de experimentos.

- **Experimentos y desarrollo de software:** Posteriormente, se ha realizado experimentos sobre la bases de datos NIST SRE 2006. 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 a partir de las pruebas realizadas así como una comparativa entre los distintos parámetros o características de locutor empleados en el reconocimiento. Con los resultados obtenidos y los respectivos análisis realizados, se ha procedido a redactar la presente memoria.

1.4 Estructura 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.

- **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. Se presenta la descripción del modo de operación de estos sistemas, como también la descripción de la información presente en la señal de voz. Posteriormente, se describe los tipos de parametrización de las características extraídas del rasgo biométrico, voz. Adicionalmente, se presenta las herramientas que permiten la medida del rendimiento del sistema, así como, las distintas técnicas para mejorar el rendimiento de éste. Finalmente, se describe las diferentes técnicas empleadas en el estado del arte del reconocimiento de locutor independiente de texto, especialmente aquellas que modelan las características acústicas de la señal de voz mediante modelos de mezclas de gaussianas y que son empleadas por los sistemas objeto de estudio en este proyecto.

- **Capítulo 3: Descripción del sistema.** En este capítulo se presenta la descripción del sistema implementado. Se detalla los tipos de parametrización de características empleados, como también, las técnicas empleadas en el reconocimiento de locutor. Finalmente, se describe el entorno experimental del sistema, tales como: protocolos de evaluación y las bases de datos usados en el proyecto.

- **Capítulo 4: Descripción de resultados.** En este capítulo se presenta los resultados obtenidos a lo largo de la evaluación del sistema implementado. Así mismo, se detalla la secuencia de experimentos realizados para dar con la configuración idónea de estos esquemas y validar su funcionamiento en distintos entornos.

- **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.

2. Estado del arte en reconocimiento biométrico de locutor

2.1 Introducción

En este capítulo se presentará, en las dos primeras secciones, el estado del arte en los sistemas de reconocimiento biométrico, dedicando mayor atención en aquellos basados en la voz. En el resto de las secciones, se presentará el estado actual de las técnicas empleadas en los sistemas de reconocimiento de locutor, así como las técnicas empleadas en la extracción de características de la señal de voz. Finalmente, se describirá las herramientas usadas en la evaluación del sistema implementado

2.2 Sistemas de reconocimiento biométrico

Un sistema de reconocimiento biométrico se basa en el reconocimiento de patrones. Estos patrones son datos biométricos o *rasgos biométricos* extraídos a partir de rasgos conductuales o rasgos físicos intrínsecos de una persona. Mediante la comparación entre los rasgos biométricos extraídos y los rasgos biométricos registrados en un sistema se reconoce a una persona.

2.2.1 Características de los rasgos biométricos

Cualquier característica física o conductual puede ser usada como identificador biométrico mientras satisfaga los requisitos siguientes [Maltoni et al., 2003]:

- **Universalidad:** todo el mundo debe poseer esa características
- **Distintivo:** los individuos deberán ser suficientemente diferentes en términos de ese rasgo.
- **Estabilidad:** la característica debe permanecer invariable a lo largo de un periodo de tiempo aceptable.
- **Evaluable:** el rasgo deber ser medido cuantitativamente.

En los sistemas de reconocimiento biométrico se debe cumplir también los siguientes requisitos:

- **Rendimiento:** el rasgo debe permitir alcanzar una precisión de reconocimiento de velocidad y robustez aceptables.
- **Aceptabilidad:** las personas deben estar dispuestas a aceptar el uso de ese rasgo como identificador biométrico.
- **Seguridad:** los sistemas basados en ese rasgo deben ser suficientemente robustos frente a posibles intentos de acceso fraudulentos.

Existen una gran variedad de características que cumplen estos requisitos y que son usados por sistemas reales de reconocimiento como rasgos biométricos.

A continuación se enumeran, por orden alfabético, algunos de los más empleados actualmente:

- ADN
- Cara
- Dinámica de tecleo
- Escáner de retina
- Firma manuscrita
- Forma de andar
- Geométrica de la mano
- Huella dactilar
- Iris
- Venas del dorso de la mano
- Voz
- …

El nivel de cumplimiento de estas características varía en función de la naturaleza misma del rasgo. Cada uno presenta ciertas ventajas y desventajas que favorecen su uso en ciertas aplicaciones y lo imposibilitan en otras. Por tanto, la elección del identificador biométrico a emplear por el sistema de reconocimiento estará condicionada por el tipo de aplicación final (necesidad de seguridad, aceptabilidad del rasgo, etc.).

En función del tipo de característica, física o conductual, empleada por el sistema biométrico, puede dividirse a éstos en dos grandes categorías:

- Biometría estática: Engloba todas 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 todas 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 de características estáticas (contenido espectral de la voz, o características acústicas). 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. Por tal motivo, asignar una de las dos categorías a la voz, definidas anteriormente, resulta difícil dado al tipo de características que emplean los sistemas de reconocimiento.

Respecto al nivel de cumplimiento de la voz como rasgo biométrico, ésta presenta un alto grado de ***aceptabilidad***, dado que no requiere un método de adquisición intrusivo.

Por otra parte, aunque no se trate de un rasgo suficientemente distintivo como el ADN, iris o la huella dactilar, sí proporciona un alto grado de seguridad en aplicaciones de verificación y permite obtener una relación de individuos más probables en sistemas de identificación menos exigentes. Como principal desventaja resalta la baja ***estabilidad***, dado que las características de la voz pueden variar en función de diversos factores como la edad, enfermedad, estado de ánimo, etc.

La Tabla 1 muestra el nivel de cumplimiento de las características citadas anteriormente por parte de la voz.

	Universalidad	Distintividad	Estabilidad	Evaluabilidad	Rendimiento	Aceptabilidad	Seguridad
Voz	M	B	B	M	B	A	A

Tabla 1. Propiedad de la voz. A, M y B denotan niveles Alto, Medio y Bajo respectivamente. Tabla adaptada de [Maltoni et al., 2003]

2.2.2 Funcionamiento de un sistema de reconocimiento biométrico

Un sistema de reconocimiento biométrico basa su funcionamiento en el reconocimiento de patrones. De esta forma, clasifica a los usuarios en función de unos rasgos biométricos preestablecidos en el sistema.

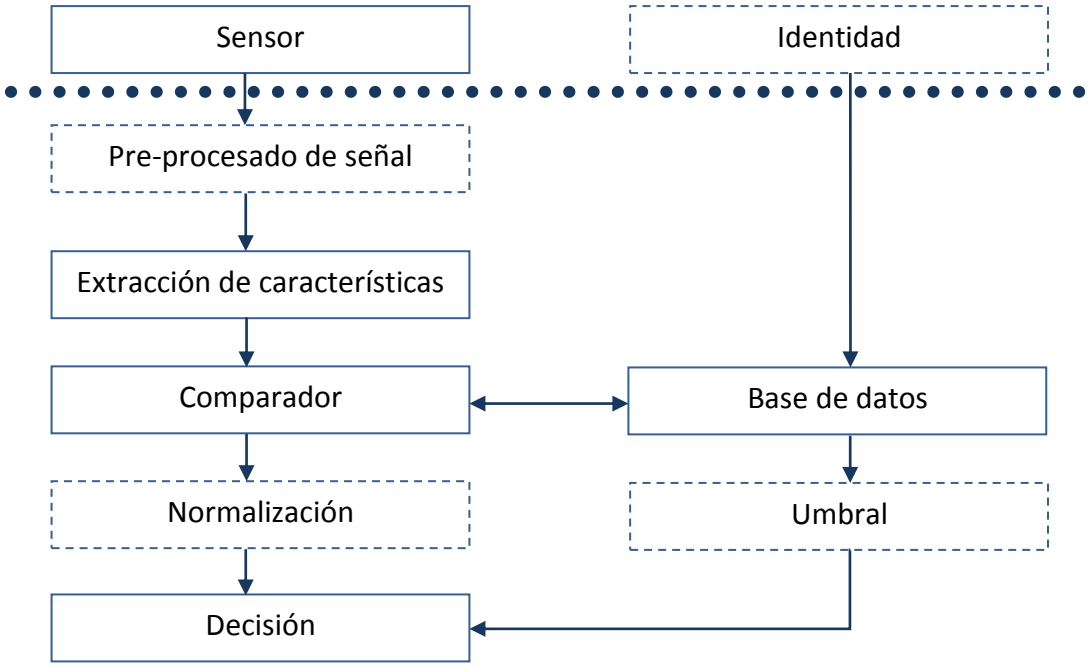

Figura 1. Esquema de funcionamiento de un sistema de reconocimiento biométrico.

La Figura 1 muestra el esquema de funcionamiento de un sistema de reconocimiento biométrico. Los módulos dibujados en línea continua son parte indispensable del sistema, mientras que los dibujados en línea discontinua son opcionales, dependiendo del tipo de sistema: identificación o verificación (términos que se explicarán en el siguiente apartado). La línea horizontal dibujada en puntos marca la separación entre la interfaz de usuario y el propio sistema. En esa interfaz se encuentra un **sensor** cuya función será de recoger el rasgo biométrico preestablecido en el sistema. En caso de un sistema de verificación, el módulo de **identidad** permitirá identificarse al usuario (mediante PIN o tarjeta de identificación).

A continuación, se presenta la etapa de **pre-procesado**. La ejecución de esta etapa será en los casos que exista degradación de la señal capturada y necesite una compensación por posibles degradaciones o facilitar la **extracción de características**. Este último modulo, **extracción de características**, será el encargado de extraer las características de la señal capturada. Estas características serán las más distintivas de la identidad para ese rasgo biométrico en concreto.

Una vez extraídas las características, se procede a la etapa de **comparación**, en este módulo las características extraídas son comparadas con los patrones de referencia previamente registrados en la **base de datos** (comparación con la identidad reclamada en el caso de un ***sistema de verificación*** o comparación con todos los almacenados en caso de un ***sistema de identificación***). Cada comparación o enfrentamiento (trial) que el sistema realiza, obtiene un valor de similitud entre las características y el patrón. Este valor de similitud es llamado también ***puntuación*** o ***score***. La puntuación obtenida puede ser sometida a una **normalización**. La normalización de puntuaciones o scores permiten la transformación a un rango de valores donde es más fácil identificar si una puntuación pertenece a un usuario genuino o a un impostor.

Finalmente, en función del valor de la **puntuación** y el **umbral**, el sistema tomará una **decisión:** en caso de un ***sistema de verificación*** el sistema decidirá si las características extraídas se corresponden con las de la identidad reclamada; en caso de un ***sistema de identificación*** el sistema decidirá si las características extraídas se corresponden o no con algunas de la identidades registradas.

2.2.3 Modos de operación

Dependiendo del contexto de la aplicación, un sistema biométrico opera en dos modos: modo verificación o modo identificación. Previo a estos dos modos, el sistema requiere un modo adicional en el que los patrones de referencia de las identidades a reconocer son guardados en el sistema: modo registro.

- **Modo registro**

En este modo se registra todos los patrones de referencia junto con la información del usuario a reconocer**.** Los rasgos biométricos son capturados por el sensor. A partir de estos rasgos capturados se procede a la extracción de las características identificativas. Dependiendo de la técnica a emplear en el reconocimiento en que se basa el sistema, estas características pueden constituir en sí mismas el patrón de referencia de la identidad a reconocer o puede ser necesario un proceso de "entrenamiento" a partir de ellas para generar un modelo estocástico de la identidad a reconocer. Es por ello que a este modo también se conoce como **fase de entrenamiento**. Finalmente, estos patrones de referencia o modelos, se almacenan en una base de datos.

Terminado la fase de registro de todos los usuarios que debe contemplar el sistema, éste entrará en funcionamiento en uno de estos dos modos:

- **Modo verificación**

En este modo el sistema valida la identidad de una persona mediante la comparación de los rasgos biométricos de la identidad reclamada, los cuales están registrados en la base de datos, y los rasgos biométricos de la persona que reclama la identidad. En este modo el sistema hace una comparación 1:1, en donde sólo se tiene una identidad reclamada como modelo o patrón de referencia, y las características de un solo usuario.

- **Modo identificación**

En este modo el sistema reconoce a un individuo mediante la búsqueda, en la base de datos, de rasgos biométricos de todos los usuarios que coincidan con las características biométricas del individuo y así asignarle una identidad. En este modo el sistema hace una comparación 1:N.

La Figura 2 muestra el esquema de funcionamiento de los tres modos basados en características extraídas de la voz. La ejecución de la fase de **pre-procesado** que se muestra en la figura dependerá si se requiere compensación de la señal o facilitar la extracción de características.

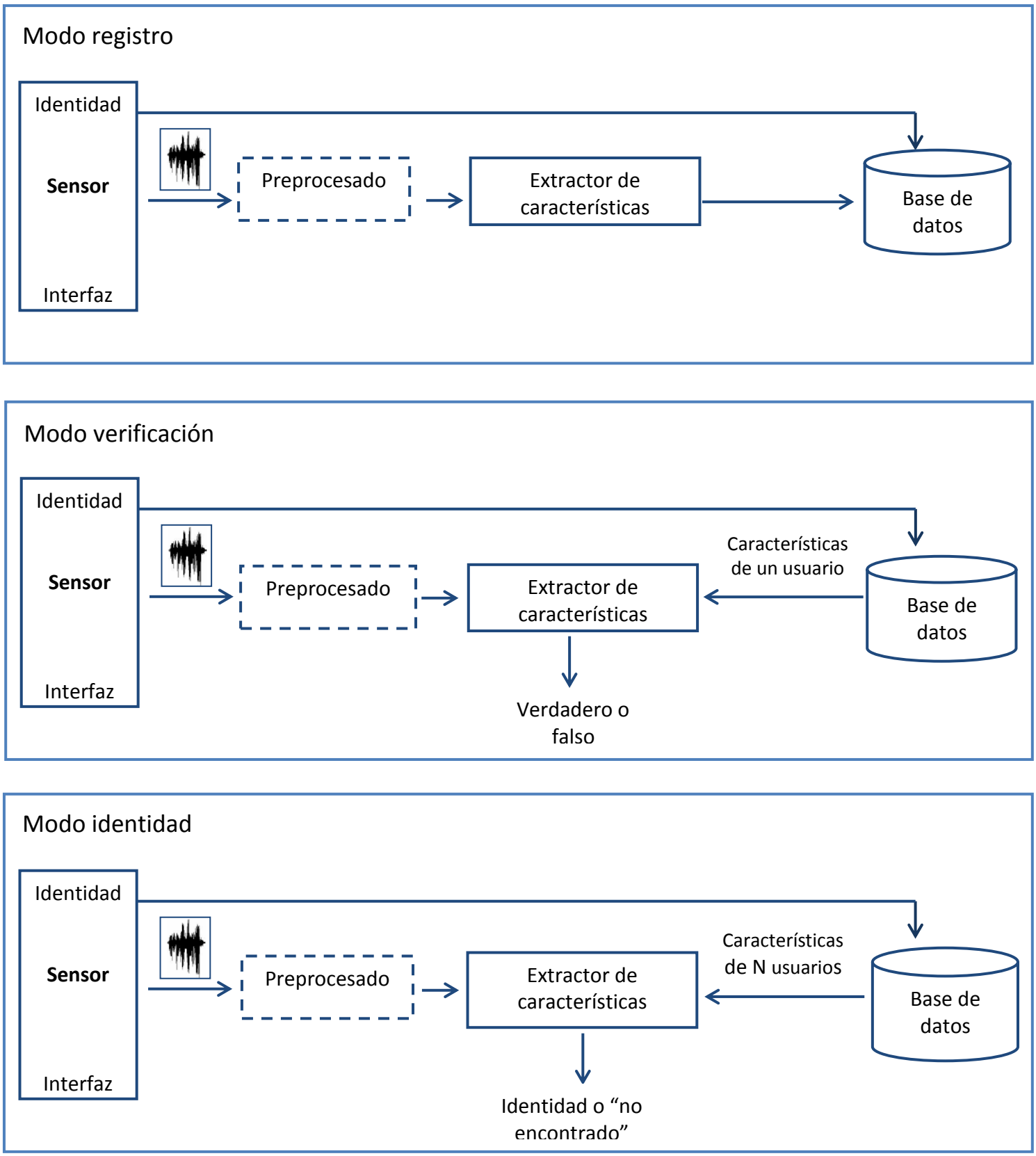

Figura 2. Modos de funcionamiento de un sistema de reconocimiento biométrico

2.3 Información del hablante en la señal de voz

La comunicación oral es la forma más habitual de transmitir información entre las personas. La transmisión de información se logra mediante el uso de mecanismos complejos, los cuales nos permiten construir un mensaje lingüísticamente correcto y articular los distintos órganos que entran en la producción de la señal que transporta el mensaje. Adicionalmente, la señal generada debe tener una clara información de los fonemas y estar en concordancia con una serie de normas gramaticales para poder ser entendida por los receptores que hablan la misma lengua.

Las distintas señales generadas por una sola persona no presentan uniformidad, dado que existen infinitas realización acústicas posibles para la trasmisión de un mismo mensaje. Las infinitas realizaciones acústicas se vinculan directamente con la variabilidad, siendo uno de los grandes problemas que se afronta en las tecnologías del habla. La variabilidad depende de factores como el estado de ánimo, edad, estado de salud, patologías fonatorias y fisiológicas.

La identidad de los locutores está directamente relacionada con las características fisiológicas (longitud y forma de tracto vocal, configuración de los órganos articulatorios), factores sociolingüísticos (nivel de educación, contexto lingüístico y diferencia dialectales) y de comportamiento (hábitos lingüísticos, características en la entonación). En función de estas características cada locutor introduce particularidades en la señal de voz que nos permite diferenciarlo de los demás.

Estas características particulares de la señal de voz serán la base del reconocimiento de la identidad de locutor. Además, esta información no se encuentra aislada dentro de la señal de voz, sino enlazado con el resto de informaciones a distintos niveles de información. Por tanto, se debe centrar en la extracción eficiente de características individualizadoras.

2.3.1 Niveles de información

Existen multitud de estudios en los que se muestran los mecanismos de las personas para reconocer la identidad de los distintos locutores. Todos ellos parecen apuntar a que la clave está en la combinación de los distintos niveles de información, así como en el peso que se le da a cada uno de ellos. Los sistemas de reconocimiento automático de locutor trataran de asemejarse al comportamiento humano, combinando las distintas fuentes de información de la mejor manera posible [Reynolds, 2003].

Las particularidades de la voz pueden dividirse en cuatro grandes grupos según el nivel en que se den. Se presenta desde el nivel más alto hasta el más bajo: nivel lingüístico, nivel fonético, nivel prosódico y nivel acústico.

- **Nivel lingüístico**

  En este nivel se encuentran las características idiolectales. Estas características describen la forma en que el locutor hace uso del sistema lingüístico, y se verán influenciadas por aspectos relacionados a la educación, origen y las condiciones sociológicas del locutor. En función de estas características se puede tener sistemas que modelen locutores por la frecuencia de uso de las palabras o secuencias de palabras.

- **Nivel fonético**

  Está compuesto por las características fonotácticas [Carr, 1999], es decir fonemas y sus respectivas secuencias. El uso de estas características conforma un patrón único para cada locutor. Por tanto, los sistemas fonotácticos intentan modelar el uso que lo locutores hacen de estas unidades léxicas.

- **Nivel prosódico**

  La prosodia se define como la combinación de energía, duración y tono de los fonemas, y es la principal de dotar a la voz de sentido y naturalidad. La prosodia consta de elementos comunes para todos los hablantes, permitiendo distinguir el tipo de mensaje: declarativo, interrogativo, imperativo; sin embargo, cada locutor emplea dichos elementos prosódicos de manera distinta. Dos de los elementos prosódicos más representativos del hablante son el tono y la energía. En función de estas características se puede implementar sistemas de reconocimiento de locutor.

- **Nivel acústico**

  Es el nivel más bajo de la clasificación. Está compuesto por las características espectrales a corto plazo de la señal de voz y su evolución a lo largo del tiempo. Estas características están directamente relacionadas con las acciones articulatorias de cada individuo, la forma en que se produce cada sonido y la configuración fisiológica del aparato fonatorio en el mecanismo de producción de voz. La información espectral extrae las particularidades del tracto vocal de cada locutor así como su dinámica de articulación. Esta información puede dividirse en dos grupos: estático y dinámico. La información estática es la extraída del análisis de cada trama individual. Por el contrario, la información dinámica se extrae del análisis de las tramas de forma conjunta, de esa forma es posible recoger el cambio de posiciones de articulaciones a otras.

Cada vez es más la tendencia de la tecnología actual de alcanzar los niveles de precisión humana. Para ello, deberá ser capaz de procesar la mayor cantidad de información posible de cada uno de los niveles e integrarla de manera muy eficaz.

2.3.2 Variabilidad de parámetros determinantes de la identidad

Como se comentó al inicio de la sección, la variabilidad es una de las mayores dificultades al momento de trabajar con la señal de voz. Muchos de los factores que influyen en la variabilidad son controlables voluntariamente por el locutor, mientras que otras no lo son.

Si realizamos un análisis más concreto de los factores de variabilidad que están presentes en las características determinantes de la identidad de locutor, se puede clasificar en dos grupos principales: factores que se corresponden con la variabilidad propia de la señal de voz y la variabilidad debida al paso del tiempo.

- **Variabilidad propia de la señal de voz**

  Dentro de este grupo existen factores clasificados en dos subgrupos:

  Factores intrínsecos: Entre estos factores destacan: edad del locutor, estado emocional, estado físico, velocidad de articulación, etc. Algunos de estos factores son controlables por el locutor; sin embargo, la mayoría no se puede controlar voluntariamente, por lo cual las características de la señal de voz no permanecen fijas.

  Factores extrínsecos: Están relacionados con características externas al locutor, tales como: entorno acústico y los dispositivos de adquisición y transmisión. Cada uno de estos factores introduce una serie de características propias como el ancho de banda, margen dinámico, reverberación, etc.

- **Variabilidad debida al paso del tiempo**

  Es uno de los factores más perjudiciales a la hora de determinar automáticamente la identidad de locutor.

Ambos factores pertenecen a la variabilidad intra-locutor o inter-sesión (variabilidad para un mismo locutor entre distintas sesiones de captura). Por ello, es deseable que la variabilidad sea la menor posible a nivel intra-locutor. Mediante técnicas tales como Factor Analysis [Kenny, 2006], se pretende modelar la variabilidad y compensarla. Por el contrario, se intentará maximizar la variabilidad existente entre distintos locutores (inter-locutor) de forma que sea más fácil distinguirlos.

2.4 Extracción de características de locutor

La extracción de características de la señal de voz, también conocido como parametrización, es el paso previo a cualquier sistema de reconocimiento automático. Estas características, idealmente, deben cumplir las siguientes condiciones [Rose, 2002] [Wolf, 1972]:

- Presentar una menor variabilidad para un mismo locutor (intra-locutor) y mayor entre locutores distintos (inter-locutor).
- Ser robustas frente al ruido y la distorsión.
- Ocurrir de forma frecuente y natural en el habla.
- Ser fáciles de medir a partir de la señal de voz.
- Ser difíciles de imitar.
- Ser independiente de la salud de locutor o variaciones a largo plazo en la voz.

El tipo de característica a emplear en el reconocimiento dependerá del nivel de información al que la identidad del locutor quiera reconocerse. Es decir, existen características que son más aptas para trabajar a nivel prosódico que a nivel acústico, por ejemplo. En el caso de los sistemas de reconocimiento de alto nivel como los basados en fonemas, necesitan reconocer la secuencia de fonemas de la locución, por tanto, deben apoyarse en sistemas de reconocimiento de habla (*Automatic Speech Recognition, ASR*), los cuales hacen uso de las características acústicas y espectro-temporales.

A continuación se presenta una posible clasificación de las características en la que atiende su interpretación física [Kinnunen and Li, 2010]. Se ha incluido en esta clasificación los niveles de información que trata de capturar cada tipo de características:

- **Características espectrales a corto plazo (nivel acústico)**

  Se obtienen sobre segmentos de voz entre 20 y 30 milisegundos de duración. La información espectral a corto plazo está acústicamente relacionada con el timbre, así como de las propiedades de resonancia del tracto vocal.

- **Características de la fuente de voz (nivel acústico)**

  Caracterizan la forma en que se origina la voz en el tracto vocal, es decir, modelan el flujo glotal.

- **Características prosódicas espectro-temporales (nivel acústico y prosódico)**

  Recogen información de la entonación y el timbre sobre segmentos de voz de duración entre decenas y centenas de milisegundos.

- **Características de alto nivel (nivel lingüístico y fonético)**

  Recogen información particular a nivel de conversación de los locutores, como por ejemplo el uso característico de fonemas o palabras determinadas, o secuencia de éstos.

2.4.1 Coeficientes MFCC (Mel-Frecuency Cepstral Coefficients)

Los coeficientes MFCC [Davis and Mermelstein, 1980] fueron introducidos originalmente para el reconocimiento de habla y posteriormente adaptados para el reconocimiento de locutor. El proceso de obtención de estos coeficientes consiste en una serie de etapas que a continuación se describirá.

En primer lugar, se divide el flujo de la señal de voz en tramas de duración igual a 20 o 30 milisegundos, que son procesadas individualmente. La división en tramas permite que dentro de ese intervalo la señal pueda considerarse estacionaria (véase Figura 3). Además, la división en tramas se realiza con solapamiento del 50 % (10 ms), a través de ventanas tipo Hamming, con el objetivo de no perder información existente entre ellas debido al procesado posterior ("enventanado").

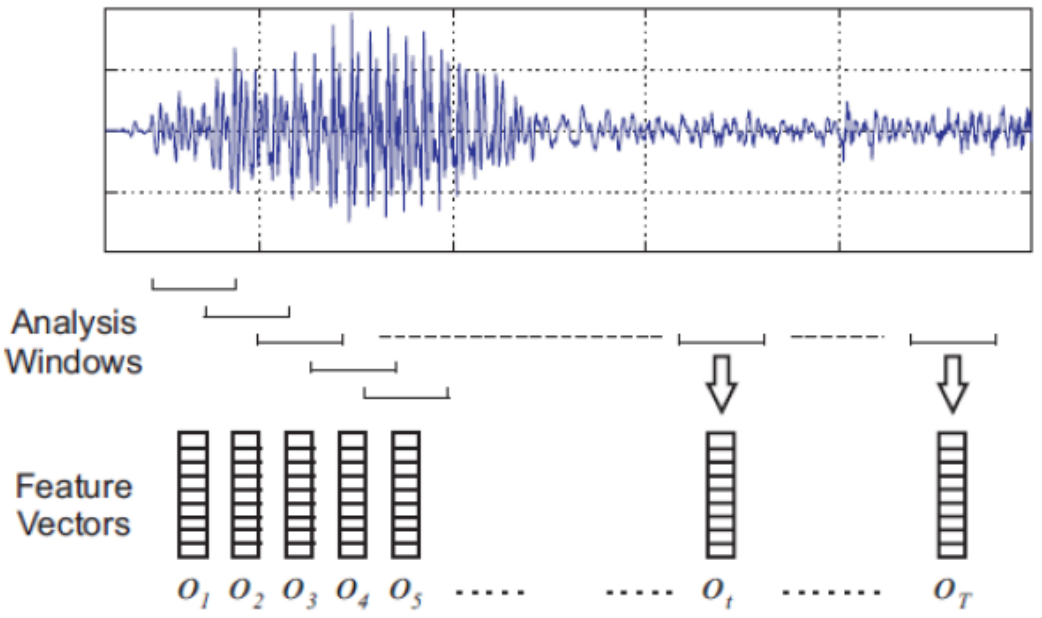

Figura 3. División de la señal de voz en tramas para la extracción de características

A continuación, se procede a la etapa de "enventanado" tipo Hamming (véase Figura 4), necesaria para realizar posteriormente un correcto análisis espectral a través de la transformada discreta de Fourier (DFT) [Oppenheim et al., 1999] o su implementación rápida, la FFT (Fast Fourier Transform) que es el que se emplea normalmente para mayor eficiencia.

El proceso de "enventanado" atenúa la señal en los extremos de la trama con el fin de que no aparezcan componentes de alta frecuencia inexistentes en la señal original.

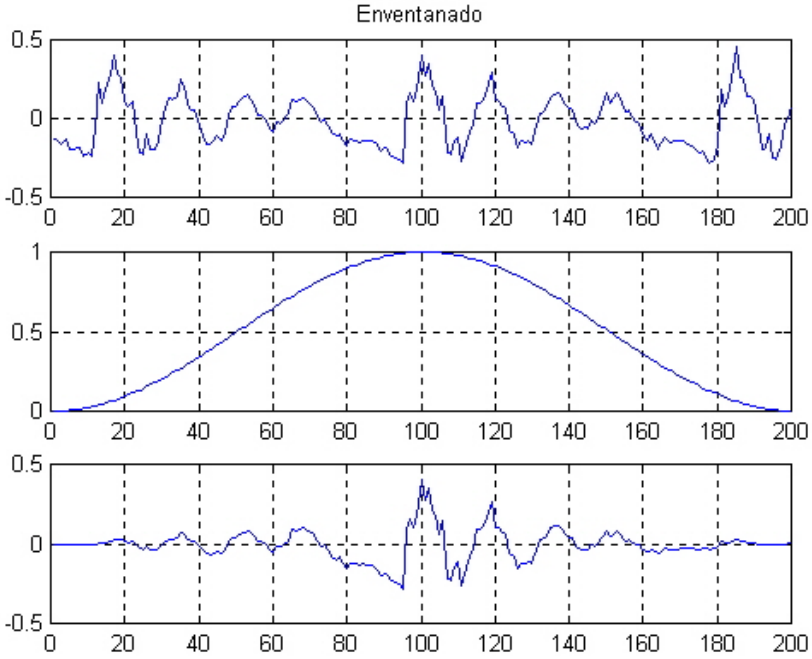

Figura 4. Enventanado de la señal con ventana tipo Hamming [López et al., 2003]

Luego del enventanado, a cada trama se puede aplicar un preénfasis para resaltar las altas frecuencias del espectro. Posteriormente, se procede al cálculo del FFT (Fast Fourier Transform) de la señal obtenida. Normalmente, sólo se guarda la amplitud del espectro obtenido. La información de dicha envolvente se recoge mediante un banco de filtros perceptual de tipo Mel. El objetivo de este filtrado es aproximar la resolución espectral a la respuesta del oído humano mediante la siguiente transformación de frecuencia:

$$Mel(f) = 2595.\log_{10}\left(1 + \frac{f}{700}\right) \tag{2.1}$$

A la salida de los filtros, que integran la energía dentro de su ancho de banda, se aplica el logaritmo natural a dichas energías y luego la transformada discreta del coseno (Discrete Cosine Transform, DCT) con el objetivo de comprimir la información, reduciendo así el número de dimensiones del vector de características. De las salidas de los filtros, denotadas mediante Y(m), m = 1,…,M , los coeficientes MFCC se obtienen mediante a siguiente transformación:

$$C_n = \sum_{m=1}^{M} [lnY(m)]\cos[\frac{\pi.n}{M}(m - \frac{1}{2})] \tag{2.2}$$

donde n es el índice del coeficiente cepstral. El vector de características final se forma con 12 o 20 primeros coeficientes C_n. La Figura 5 representa el proceso de obtención de los coeficientes MFCC.

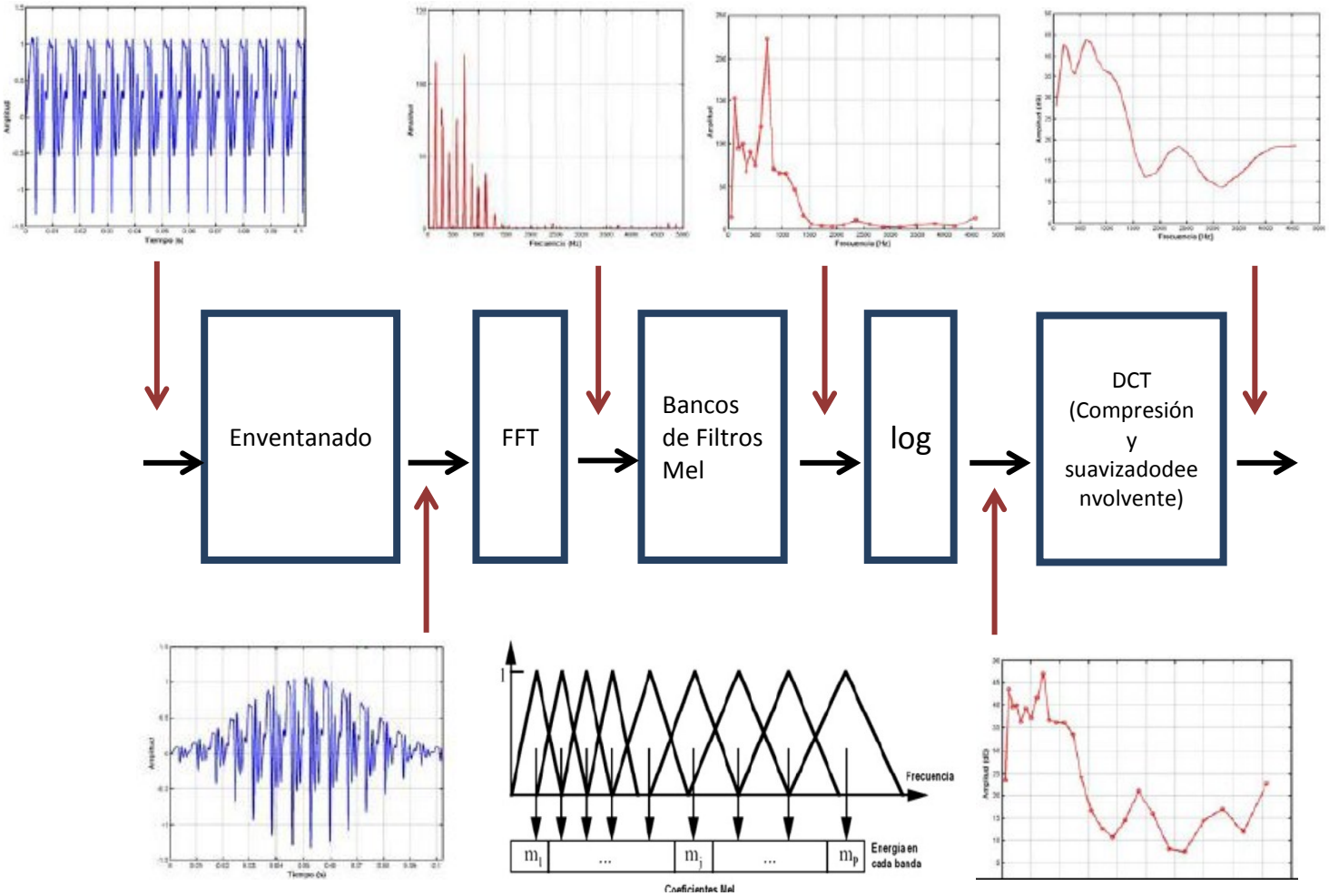

Figura 5. Proceso de obtención de los coeficientes MFCC

La ventaja de los coeficientes MFCC es su naturaleza ortogonal. Esto permite trabajar con matrices de covarianza diagonales en los sistemas basados en GMM 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. Mediante técnicas como la normalización con respecto a la media cepstral (Cepstral Mean Normalization, CMN) o el filtrado RASTA (RelAtiveSpecTrAl) es posible reducir la influencia de esta componente aditiva.

2.4.2 Normalización de características

En principio, es posible usar técnicas supresoras de ruido para mejorar la calidad de la señal previa a la extracción de características. Sin embargo, la mejora de la señal dentro del proceso de reconocimiento incrementaría la carga computacional. Por tanto, es deseable diseñar un extractor lo suficientemente robusto o usar técnicas de compensación sobre las características. A continuación se presentarán distintas técnicas que tratarán de compensar la influencia de ruido y otros efectos perturbadores de la señal.

- **Normalización por media cepstral (CMN, Cepstral Mean Normalization)**

  El efecto de canal sobre la voz se modela en el dominio espectral mediante el producto de la señal S(z), y la función de transferencia del canal, G(z) [Atal, 1974] [Furui, 1981].

$$T(z) = S(z).G(z) \tag{2.3}$$

  En el dominio *cepstral*, el efecto del canal se convierte en aditivo, al utilizarse el logaritmo de las componentes espectrales (véase la expresión (2.2)):

$$\mathrm{DFT}^{-1}(\log|T(z)|) = \mathrm{DFT}^{-1}(\log|S(z)|) + \mathrm{DFT}^{-1}(\log|G(z)|) \tag{2.4}$$

  Asumiendo que el canal es invariante y que la media de los coeficientes cepstrales es nula, el canal puede ser estimado como la media temporal de la señal filtrada T(z), por lo que sustrayendo la media temporal de los coeficientes cepstrales, el efecto aditivo del canal será minimizada en cierta medida.

$$\mathrm{y[n]} = t[n] - \frac{1}{N}\sum_{n=1}^{N} t[n] \tag{2.5}$$

  donde t[n] representa el vector de característica en el instante n de la señal de voz afectada por el canal g[n], e y[n] representa el vector de características normalizadas.

- **Filtrado RASTA (RelAtiveSpecTrAl)**

  El filtrado RASTA [Hermanski and Morgan, 1994] [Malayath et al., 2000] realiza un filtrado paso banda sobre la trayectoria temporal de cada característica (dimensión del vector) en el dominio cepstral para suprimir frecuencias de modulación que estén fuera del rango típico de la señal de voz.
  Su funcionamiento se basa en que cualquier constante o componente que varíe muy rápido o muy lento, no se considera habla.

Cabe destacar que el filtrado RASTA es independiente de la señal, mientras que la normalización CMN son adaptativos dado que utilizan estadísticas obtenidas de la propia señal.

- **Feature warping**

  Su objetivo es modificar la distribución de las características a corto plazo para ajustarse a una distribución final gaussiana de media nula y varianza unidad, ya que la distorsión de canal modifica la distribución real de los coeficientes cepstrales en cortos periodos de tiempo [Pelecanos y Sridharan, 2001].

- **Feature mapping**

  Esta técnica de normalización requiere modelar, mediante GMM-UBM, la influencia de cada canal por separado y así aplicar una compensación particular a las características provenientes de un mismo canal [Reynolds et al., 2003].

2.4.3 Coeficientes Δ-Cepstrales

En el apartado 2.4.1 se hizo un enfoque sobre los coeficientes MFCC, también llamados parámetros estáticos. Estos parámetros estáticos modelan las características de la señal de voz a nivel segmental. Para incorporar características supra-segmentales a los vectores de características debemos analizar la información transicional que aparece en éstos. Mediante la primera y segunda derivada temporales de los coeficientes cepstrales se obtienen coeficientes que tratan de representar la información de coarticulación entre fonemas, por ello miden velocidades y aceleraciones alrededor de un tiempo. Estos coeficientes se les conoce como: coeficiente de velocidad, o Δ, y aceleración, o Δ^2 [Furui, 1981] [Huang et al., 2001] [Soong and Rosenberg, 1988]. Estos coeficientes se obtienen a partir de las siguientes expresiones:

$$C_m^{(0)}[n] = \frac{\sum_{k=-K}^{K} h_k . y_m[n+k]}{\sum_{k=-K}^{K} h_k} \tag{2.6}$$

$$C_m^{(1)}[n] = \frac{\sum_{k=-K}^{K} k . h_k . y_m[n+k]}{\sum_{k=-K}^{K} k^2 h_k} \tag{2.7}$$

Donde h_k es una ventana temporal, normalmente rectangular y simétrica, de longitud 2K+1, y m indica la dimensión del vector característica sobre el que se aplica la derivada. Los coeficientes Δ y Δ^2 se calculan sobre el vector de características estáticas previamente normalizadas, y posteriormente se añaden al mismo vector. Por ejemplo, la Figura 6 muestra el proceso para el caso de 7 coeficientes MFCC (estáticos) con sus primeras derivadas para K=1 en el instante t_n. El vector final de características estará conformado por catorce coeficientes, $\vec{x} = \{C_0, C_1, \dots, C_6, \Delta C_0, \Delta C_0, \Delta C_1, \dots, \Delta C_6\}$.

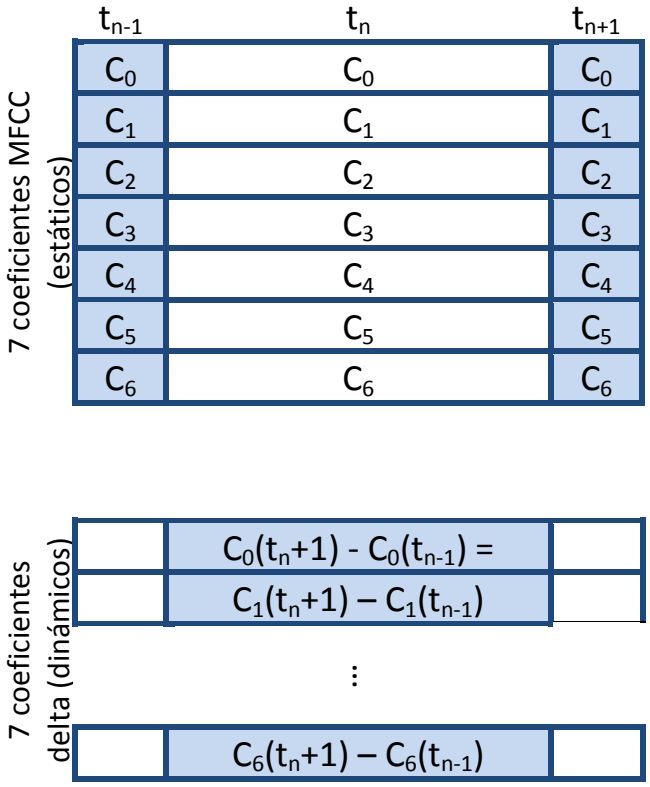

Figura 6. Inserción de los coeficientes derivados (información dinámica) a continuación de los coeficientes cepstrales (información estática)

Para los sistemas empleados en la parte experimental de este proyecto, se trabaja con 19 coeficientes cepstrales y 19 coeficientes delta (primera derivada, K=1), normalizados mediante CMN, filtrado RASTA y *feauture warping*.

2.5 Rendimiento de los sistemas de reconocimiento de locutor

En este apartado se hará una descripción del método y herramientas para la evaluación del sistema de verificación de locutor a implementar, y además, técnicas para para mejorar su rendimiento.

2.5.1 Relación de Verosimilitud (LR, Likelihood Ratio)

En la literatura, la relación de verosimilitud o LR se define, en el campo forense acústica, como la relación de las proposiciones del *fiscal* y la *defensa*:

- H_p (hipótesis del fiscal): el segmento de audio recuperado en la escena del crimen proviene del sospechoso.

- H_d (hipótesis de la defensa): el segmento de audio recuperado en la escena del crimen no proviene del sospechoso.

El ratio o relación se define:

$$LR = \frac{P(E|H_p, I)}{P(E|H_d, I)} \tag{2.8}$$

donde **E** es la evidencia disponible, la cual está formada por una muestra de origen desconocido y una muestra controlada cuyo origen sí se conoce; **I** es la información relevante para el caso. Mediante el teorema de Bayes se puede calcular la relación de probabilidades a posteriori tomando en cuenta el valor de los LR y la información a priori de acuerdo con la fórmula:

$$\frac{P(H_p|E, I)}{P(H_d|E, I)} = LR\,\frac{P(H_p|I)}{P(H_d|I)} = \frac{P(E|H_p, I)}{P(E|H_d, I)}\,\frac{P(H_p|I)}{P(H_d|I)} \tag{2.9}$$

Aplicando la relación de verosimilitud en el reconocimiento de locutor, dado un segmento de habla, ***Y***, y un locutor hipotético, ***S***, el objetivo en los sistemas de verificación de locutor es determinar si el segmento de habla o locución ***Y*** fue dicho por el locutor ***S***. Por tanto, nos encontramos con dos hipótesis:

- H_0: ***Y*** pertenece al hipotético locutor ***S***.
- H_1: ***Y*** no pertenece al hipotético locutor ***S***.

La decisión entre las dos hipótesis se basa en el test de la relación de verosimilitud. Este test será idealmente óptimo si se conocen las funciones de verosimilitud que contempla cada hipótesis.

$$\frac{p(Y|H_0)}{p(Y|H_1)}\begin{cases}\geq\ \theta\ \ se\ acepta\ \ H_0\\ <\ \theta\ se\ acepta\ H_1\end{cases} \tag{2.10}$$

donde $p(Y|H_i)$, i =0, 1 es la función de densidad de probabilidad de la hipótesis H_i evaluada para el segmento de habla Y. Además, tal como se presenta en la expresión (2.10), si la razón de verosimilitud es mayor o igual que el umbral de decisión θ, la hipótesis H_0 se acepta y si es menor se acepta la hipótesis H_1.

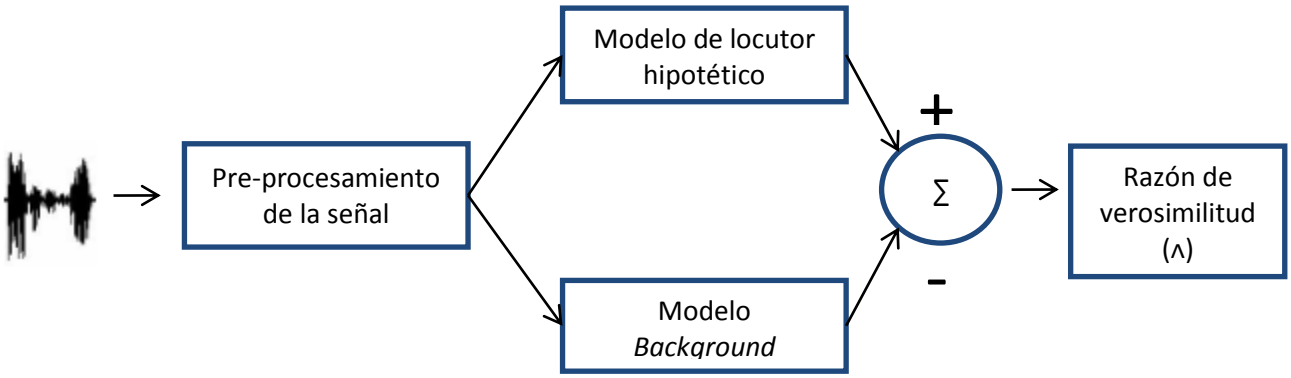

Figura 7. Sistema de verificación de locutor basado en relación de verosimilitud.

En la Figura 7 se muestra el esquema de funcionamiento de un sistema de verificación de locutor basado en la razón de verosimilitud. En la parte de pre-procesamiento se extrae las características X = {x_1,x_2, …, x_T} de la señal que transmite información dependiente de locutor. Posteriormente, estas características son usadas para calcular las funciones de verosimilitud de las hipótesis H_0 y H_1. Finalmente, se calcula la razón de verosimilitud. Adicionalmente, se debe tener en cuenta que las hipótesis se representan mediante modelados estadísticos. Por tanto, el modelo λ_{hyp} representa a H_0 y el modelo $\overline{\lambda_{\mathrm{hyp}}}$ representa la hipótesis alternativa, H_1. La nueva razón de verosimilitud, aplicado posteriormente el logaritmo, es:

$$\Lambda = \log p\left(X\middle|\lambda_{hyp}\right) - \log p\ (X|\overline{\lambda_{\mathrm{hyp}}}) \tag{2.11}$$

Dentro del contexto de los sistemas de evaluación de locutor usando GMM-UBM, el modelo λ_{hyp} caracteriza al locutor hipotético ***S*** en el espacio de características de x, mientras que, $\overline{\lambda_{\mathrm{hyp}}}$ caracteriza al locutor hipotético dentro del espacio UBM. Es decir, se compara la probabilidad de que las características extraídas provengan del modelo del locutor ($\lambda_{hyp} = \lambda_{target}$) entre la probabilidad que provengan del modelo UBM ($\overline{\lambda_{\mathrm{hyp}}} = \lambda_{UBM}$).

2.5.2 Evaluación del rendimiento

El buen diseño e implementación de un sistema de reconocimiento conlleva también a su respectiva evaluación. El objetivo de la evaluación es comprobar las capacidades del sistema, bondad del sistema desarrollado y mejorar su rendimiento. Para ello, mediante un conjunto de pruebas de reconocimiento y con ayuda de herramientas (serie de valores, curvas, etc.) se evalúa las diferentes técnicas empleadas para el reconocimiento. Las pruebas deben realizarse en condiciones lo más parecidas posibles a aquellas en donde el sistema trabaja en un entorno más real y donde se conoce a priori las entidades. Esto permitirá evaluar el rendimiento de forma más objetiva.

Los sistemas de verificación funcionan normalmente en dos pasos: En primer lugar, se calcula un valor de similitud, también llamado puntuación (score), entre las características de un rasgo biométrico capturado por el sistema y el patrón de referencia de la identidad reclamada. Idealmente, cuanto mayor sea la ***puntuación*** o ***score***, mayor será el apoyo a la hipótesis de que la identidad de ambos coincida. En segundo lugar, mediante el proceso de calibración (2.5.2.1) se obtiene una relación de verosimilitudes (2.5.1) para luego ser comparado con un umbral. Si el valor de la relación de verosimilitud es mayor que el umbral, el sistema aceptará el rasgo biométrico relacionado a la identidad reclamada. En caso contrario, si el valor de verosimilitud es menor que el umbral, el sistema lo rechazará. En las decisiones que toma el sistema de verificación pueden darse dos tipos de errores:

- **Error de falso rechazo**: se produce cuando el sistema rechaza el rasgo biométrico de un usuario genuino, es decir, aquel que reclama su verdadera identidad.
- **Error de falsa aceptación**: se produce cuando el sistema acepta el rasgo biométrico de un impostor, es decir, aquel que reclama una identidad no propia.

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

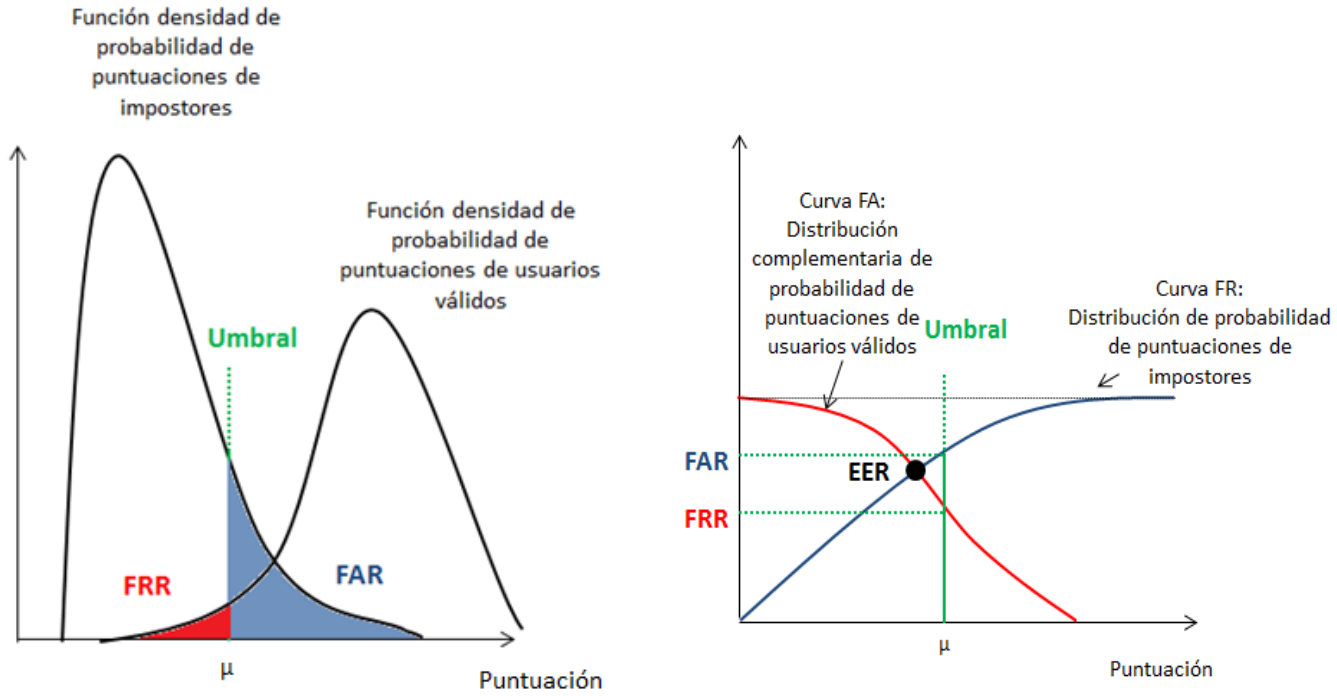

Figura 8. Funciones de densidad y distribuciones de probabilidad de usuarios e impostores

En la Figura 8 se muestra dos formas de representar las tasas de falsa aceptación y de falso rechazo. La gráfica de la izquierda representa la función de densidad de probabilidad de las puntuaciones obtenida por los usuarios válidos e impostores. Para un valor fijo del umbral (eje de abscisas), la tasa de falso rechazo (FRR, False Rejection Ratio) o probabilidad de que un usuario válido sea rechazado es igual al área bajo la curva de densidad de probabilidad de puntuaciones de usuarios genuinos (a la izquierda del umbral).
Por el contrario, la tasa de falsa aceptación (FAR, False Acceptance Ratio) o probabilidad de que un usuario impostor sea aceptado se corresponde con el área bajo la curva de densidad de probabilidad de puntuaciones de usuarios impostores (a la derecha del umbral).

En la gráfica de la derecha se representa la función de distribución de probabilidad de puntuaciones de usuarios válidos e impostores. En esta curva el valor de la tasa de falso rechazo (FRR) y el valor de la tasa de falsa aceptación (FAR) es igual al valor del eje de ordenas correspondiente al valor del umbral en el eje de abscisas. El punto de intersección de las dos curvas se corresponde con el valor de la tasa de error igual (EER, Equal Error Rate). Existe una relación directa entre las tasas de error y el valor del umbral. Para un umbral muy bajo, un mayor número de impostores podrían ser aceptados como válidos pero disminuiría el número de usuarios válidos rechazados, mientras que para un valor muy alto, muchos usuarios válidos serían rechazados pero el número de aceptación de usuarios impostores disminuiría. Por tanto, el valor que se fije en el umbral dependerá de las especificaciones del punto de trabajo deseado para el sistema.

2.5.2.1 Calibración

En los sistemas de verificación de locutor, las puntuaciones (scores) que se obtienen miden la similitud entre las entidades dada una evidencia ***E***, por ejemplo, mientras el valor de la puntuación sea más alto, más se apoya la hipótesis de que un segmento del habla ***Y*** pertenezca a un hipotético locutor ***S***; sin embargo, no se interpreta como una razón de verosimilitud (LR, Likelihood Ratio). Es decir, una puntuación o *score* no tiene una interpretación probabilística, el cual es requerido en un contexto forense. Por otra parte, en el contexto forense, en la ecuación 2.9 se observa que mediante el valor de LR y la información a priori es posible obtener la relación de las probabilidades a posteriori, sin embargo, la labor del forense o científico sólo se limita a brindar el valor de LR sin considerar la información a priori.

Por tanto, el objetivo es calcular el valor de LR para saber el grado de apoyo a una determinada hipótesis. Este proceso de transformación es referido como calibración. Existen métodos para la calibración de puntuaciones, pero el más extendido es la transformación lineal de scores [Brümmer and Preez, 2006] mediante regresión logística (*FoCaltooikt http://sites.google.com/site/nikobrummer/focal*).

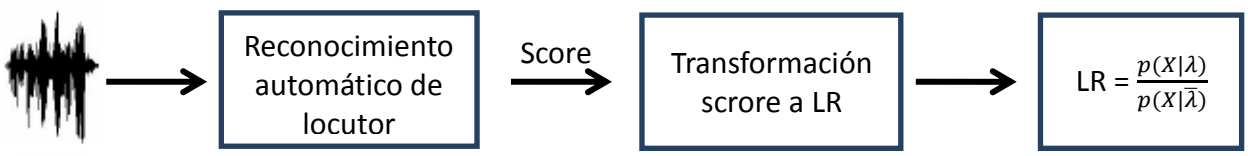

Figura 9. Esquema de la transformación de un score a LR.

2.5.2.2 Curvas Tippett

Una forma de valorar los LR obtenidos, en análisis bayesiano de evidencias forenses, es mediante las curvas Tippett. En esta representación, las distribuciones de los valores de LR de las hipótesis (H_p y H_d) son dibujadas juntas. De esta forma, se observa las distribuciones y las tasas de fallo. Esta tasa de fallos se define como la proporción de valores de LR que apoyan a la hipótesis incorrecta (LR > 1 si H_d es verdadero y LR < 1 si H_p es verdadero). En la siguiente figura se muestra un ejemplo de esta curva.

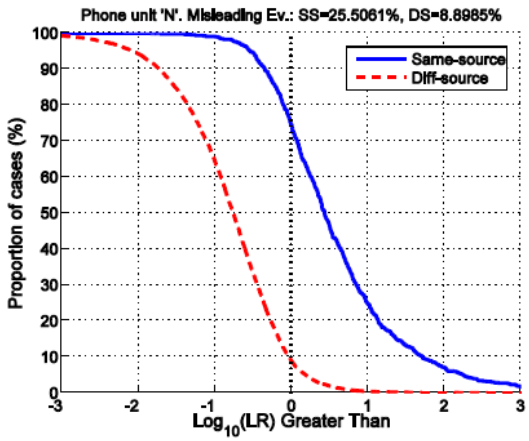

Figura 10. Representación de curvas Tippett [Franco-Pedroso et al, 2012].

2.5.2.3 Función de coste C_{llr}

Esta función de coste de log-LR (Logaritmic Likelihood Ratio Cost, C_{llr}) es una estimación del rendimiento del sistema sobre un conjunto de valores de LR, cuanto mayor sea el valor de C_{llr} peor será el rendimiento del sistema que genere dichos valores. Esta función se detalla en profundidad en [Leeuwen and Brümmer, 2007].

$$C_{llr} = \frac{1}{N_{H_p}} \sum_{i=1}^{N_{H_p}} \log_2 \left(1 + \frac{1}{LR_i}\right) + \frac{1}{N_{H_d}} \sum_{j=1}^{N_{H_d}} \log_2 (1 + LR_j) \qquad (2.12)$$

donde N_{H_p} y N_{H_d} son respectivamente los números de LR en el conjunto de evaluación cuando H_p o H_d es verdadero. Por otra parte, la función de coste C_{llr} se puede descomponer en: pérdida de discriminación, C_{llr}^{min}, y pérdida de calibración, C_{llr}^{cal}.

$$C_{llr} = C_{llr}^{min} + C_{llr}^{cal} \qquad (2.13)$$

- C_{llr}^{min} es la pérdida debida a la limitación del poder discriminativo del conjunto experimental sobre el que se trabaja, cuanto menor sea éste, mayor poder discriminativo tendrá el sistema sobre el conjunto experimental. Por tanto, C_{llr}^{min} mide el nivel de discriminación; además, representa el valor mínimo de C_{llr} que puede alcanzar el sistema sin alterar el poder discriminativo del conjunto experimental.

- Por otra parte, para la medida de la calibración se tiene C_{llr}^{cal}, el cual es la diferencia entre el coste de Log-LR y la pérdida de discriminación. Su valor es mayor que cero, alcanza valores cercanos al cero para sistemas muy bien calibrados, caso contrario, los valores de C_{llr}^{cal} crecerán sin límites y los sistemas calibrados empezarán a descalibrarse.

2.5.2.4 Curvas DET (Detection Error Tradeoff)

Otra forma de representar el rendimiento de los sistemas es mediante la curvas DET. Estas curvas permiten medir la discriminación de un conjunto de puntuaciones y valores de LR. Sobre el eje de ordenadas se representa la probabilidad de falso rechazo y sobre el eje de abscisas la probabilidad de falsa aceptación.

Mediante esta curva resulta más sencillo apreciar el balance 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 diagonal de los ejes de la gráfica.

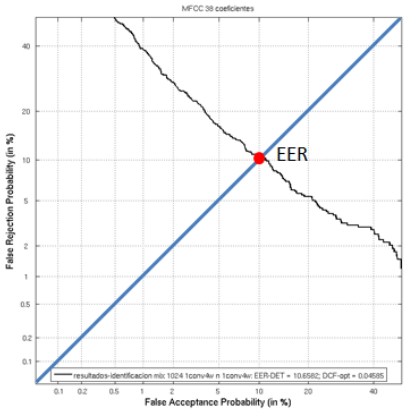

Figura 11. Ejemplo de curva DET.

2.5.2.5 Función de detección de coste (DCT, Detection Cost Function)

La función de detección de coste se define de la siguiente manera:

$$C_{DET} = C_{FR} . P_{FR} . P_{Tar} + C_{FA} . P_{FA} . (1 - P_{Tar}) \quad (2.14)$$

Los parámetros que definen la función C_{DET} son los costes relativos de errores de detección, C_{FR} y C_{FA}, y la probabilidad a priori (P_{Tar}) 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 8).

2.5.3 Normalización de puntuaciones o *scores*

Tal como se explicó en el apartado 2.2.2, en un sistema de reconocimiento biométrico puede existir una etapa de normalización de *scores* tras la comparación del patrón de referencia con las características extraídas del rasgo biométrico y posteriormente, decidir en función del umbral el origen de las características.

Dado que en muchos sistemas el umbral es común a todos los usuarios, así una puntuación determinada puede resultar en el rechazo correcto de un impostor pero suponer la falsa aceptación de un impostor. Por tanto, se produce un "*desalinamiento*" en las puntuaciones.

El objetivo de la normalización es proyectar la distribución de las puntuaciones, bien de usuarios o impostores, sobre una función de densidad de probabilidad de media igual a cero y varianza unidad, de forma que las puntuaciones de uno u otro tipo queden localizadas [Auckenthaler et al., 2000].

Las puntuaciones de impostor son muy usadas para la normalización a partir de la distribución de este tipo de puntuaciones y se conoce como normalización centrada en el impostor [Bimbot et al., 2004] [Fierrez-Aguilar et al., 2005]. La razón de usar las puntuaciones de impostor es que se dispone de mayor número de comparaciones de este tipo a comparación de las puntuaciones basadas en usuarios genuinos. A continuación se presentará dos tipos de normalización centradas en el impostor.

2.5.3.1 Z-Norm (Zero Normalization)

En la normalización Z-Norm [Li and Porter, 1988] se aplica sobre la puntuación o *score raw* obtenida tras la comparación entre un rasgo biométrico de test y un patrón de referencia. Para ello, una cohorte de puntuaciones de impostor de test se enfrenta con cada patrón de referencia, de esa cohorte se obtiene la distribución de puntuaciones cuya media μ_{Znorm} se resta a cada puntuación y se divide entre la raíz cuadrada de la varianza σ_{Znorm}.
Esta distribución de puntuaciones es dependiente del patrón de referencia, de modo que, al aplicar a todos los patrones de referencia, se alinean sus respectivas distribuciones de impostor para cualquier comparación realizada por el sistema.

$$s_{Z_{norm}} = \frac{s_{raw} - \mu_{Z_{norm}}}{\sigma_{Z_{norm}}} \tag{2.15}$$

2.5.3.2 T-Norm (Test Normalization)

Se basa en la misma idea que Z-Norm [Auckenthaler et al., 2000], pero los valores de μ_{Tnorm} y varianza σ_{Tnorm} se obtienen de la distribución de puntuaciones de impostor entre una cohorte de patrones de referencia y un rasgo biométrico de test. Esta transformación es dependiente del rasgo de test que hace que se alineen las distribuciones de impostor de todos los rasgo de test.

$$s_{T_{norm}} = \frac{s_{raw} - \mu_{T_{norm}}}{\sigma_{T_{norm}}} \tag{2.16}$$

2.5.3.3 ZT-Norm (Zero and Test Normalization)

Z-Norm y T-Norm pueden ser empleadas de forma conjunta y obtener una nueva normalización:

$$s_{ZT_{norm}} = \frac{\frac{s_{raw} - \mu_{t_{Znorm}}}{\sigma_{Z_{norm}}} - \mu_{T_{norm}}}{\sigma_{T_{norm}}} \tag{2.17}$$

donde las puntuaciones con normalización Z-Norm, previamente, se les ha aplicado normalización T-Norm y a las puntuaciones o *scores raw*, previamente, se les ha aplicada Z-Norm.

El conjunto de rasgos biométricos de test, en el caso de Z-Norm, y de patrones de referencia, en el caso de T-Norm, se denominan *cohortes.* A la hora de aplicar T-Norm la selección de cohortes de modelos es un elemento importante y sujeto a la investigación. Estos modelos han de ser lo más parecidos posible a los modelos de usuario, y su número ha de ser elevado, dado que se debe estimar una gaussiana a partir de las puntuaciones obtenidas. De la misma forma, para Z-Norm la selección de ficheros de cohorte es necesaria.

2.5.4 Fusión de sistemas

La fusión de sistemas mediante la combinación de resultados de dos o más sistemas de reconocimiento se llega a un sistema más robusto y con mejores prestaciones que los sistemas individuales por separado [Brümmer et al., 2007] [López-Moreno et al., 2008]. De esta forma se puede aprovechar información a distinto nivel de información procedentes de un mismo rasgo [Reynolds et al., 2003], además de ser complementarias y pseudo-ortogonales. La combinación de resultados puede realizarse desde dos perspectivas:

- ***Fusión basada en reglas fijas***
  Combina directamente las puntuaciones obtenidas por los sistemas individuales mediante un operador simple: suma, producto, máximo, mínimo, etc. Como requisito es necesario que las puntuaciones se encuentren en un margen de valores homogéneo. Los tipos de normalización más empleados son: normalización *min-max* y *z-score*. La normalización *min-max* transforma el rango de puntuaciones al intervalo [0,1] sin modificar la distribución original de las puntuaciones, mientras que la normalización *z-score* transforma la distribución de las puntuaciones en una distribución con media cero y varianza unidad.

- ***Fusión basada en reglas entrenadas***
  Emplea las decisiones (aceptación o rechazo) de los sistemas individuales como patrones de entrada a un nuevo sistema, tratando la fusión como un problema de clasificación de patrones. Existen técnicas tales como las redes neuronales, SVMs [Fierrez-Aguilar et al., 2003] o regresión logística [Brümmer and Preez, 2006].

2.6 Técnicas de reconocimiento de locutor independiente de texto

En los sistemas independientes de texto no existen limitaciones en cuanto al contenido de las frases empleadas para el modelado de locutores (fase de enteramiento) y las utilizadas como test, por tanto, el contenido léxico de la fase entrenamiento y la fase de test difieren en su totalidad. Esto hace del reconocimiento independiente de texto una tarea muy desafiante, pero es la vez un sistema idóneo para usuarios no cooperativos.
Durante las dos últimas décadas, estos sistemas han dominado el reconocimiento de locutor, especialmente los basados en características espectrales a corto plazo (o sistemas acústicos), ya que a través del modelado de observaciones acústicas es posible recoger más variabilidad intra-locutor.

A continuación, se presentará las técnicas más empleadas por los sistemas acústicos, prestando mayor atención a la técnica GMM-UBM o GMM-MAP [Reynolds et al., 2000].

2.6.1 Cuantificación vectorial (Vector Quantization VQ)

El modelo de cuantificación vectorial, también conocido como modelo centroide, tiene sus orígenes en la compresión de datos [Gersho and Gray, 1991] y fue introducida para el reconocimiento de locutor en la década de los 80 [Burton, 1987] [Soong et al., 1987]. Es uno de los métodos más simples para el reconocimiento de locutor independiente de texto. También se usa con propósitos de acelerar el proceso computacional [Louradour and Daoudi, 2005] [Kinnunen et al., 2006] [Roch, 2006] e implementación de prácticas ligeras [Saastamoinen et al., 2005].

Los sistemas basados en cuantificación vectorial hacen uso de las ventajas teóricas formuladas en la Teoría de la distorsión y del régimen binario de Shannon, el cual una de las conclusiones fundamentales es que siempre es posible obtener un mejor rendimiento codificando mediante vectores, que con escalares; por consiguiente, se trata de cuantificar una señal de entrada, que pude tomar valores infinitos, mediante la asignación de un vector representativo de entre un conjunto finito posible. Por ejemplo, representando un conjunto de vectores de características próximos entre sí por su vector promedio. Así, un espacio de características se puede dividir en un número de regiones determinado, cada una de las cuales tendrá su vector representativo o centroide (*codeword*). Al conjunto de vectores representativos se le denomina libro de códigos (*codebook*). La división de espacio de características se lleva a cabo mediante algoritmos de agrupamiento (clustering) como K-*means* [Linde et al., 1980] o *binary splitting*.

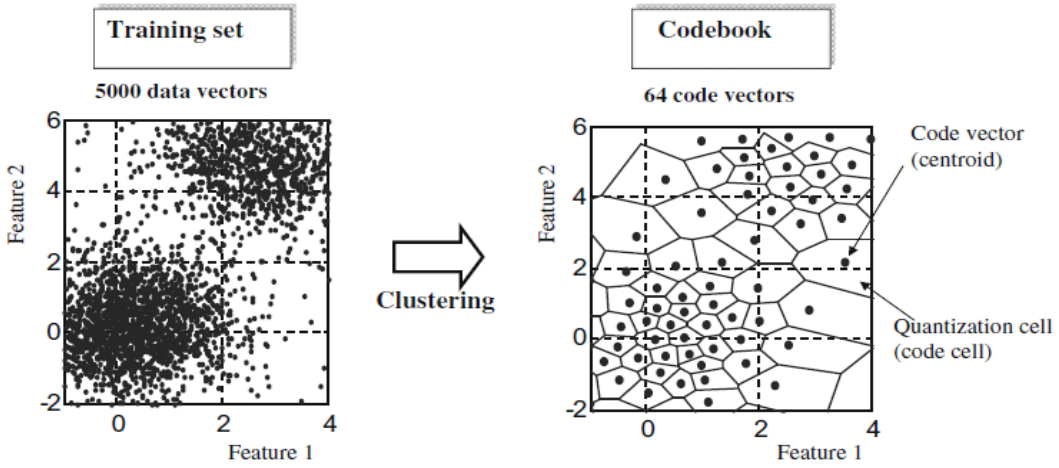

Figura 12. Construcción de un *codebook* mediante cuantificación vectorial usando el algoritmo *k-means*[Kinnunen and Li, 2010]

En el caso de reconocimiento de locutor, la identidad de locutor se puede representar mediante su correspondiente *codebook.*, lo que constituye el modelo de plantilla del locutor en la técnica de VQ. El número de regiones es potencia de 2 para facilitar la representación de los centroides en notación binaria; así un *codebook* de *b* bits tendrá $N = 2^b$ centroides. Por tanto, sólo se almacena los centroides del *codebook* para cada modelo de locutor. Esto supone una reducción drástica de la información espectral.

Con los *codebook* (vectores representativos de la identidad de locutor, centroides) $R = \{r_1, r_2, r_3, \ldots r_N\}$ se procede a la comparación con una locución de test con vectores de características $O = \{o_1, o_2, o_3, \ldots, o_T\}$. La comparación se lleva a cabo mediante la distorsión de cuantificación promedio, que se define como:

$$D_Q(O,R) = \frac{1}{T}\sum_{t=1}^{T} min\, d(o_t, r_n); \quad 1 \le n \le N \tag{2.18}$$

donde d(.,.) es la medida de la distancia entre el centroide y un vector de características. La medida de distancia puede ser la distancia euclídea.
Cuanto menor sea la distorsión de cuantificación promedio mayor será la probabilidad de que el conjunto de vectores de *O* de la locución de test pertenezcan al locutor representado por *R*.

La siguiente figura ilustra el proceso de reconocimiento mediante cuantificación vectorial para un modelo de locutor.

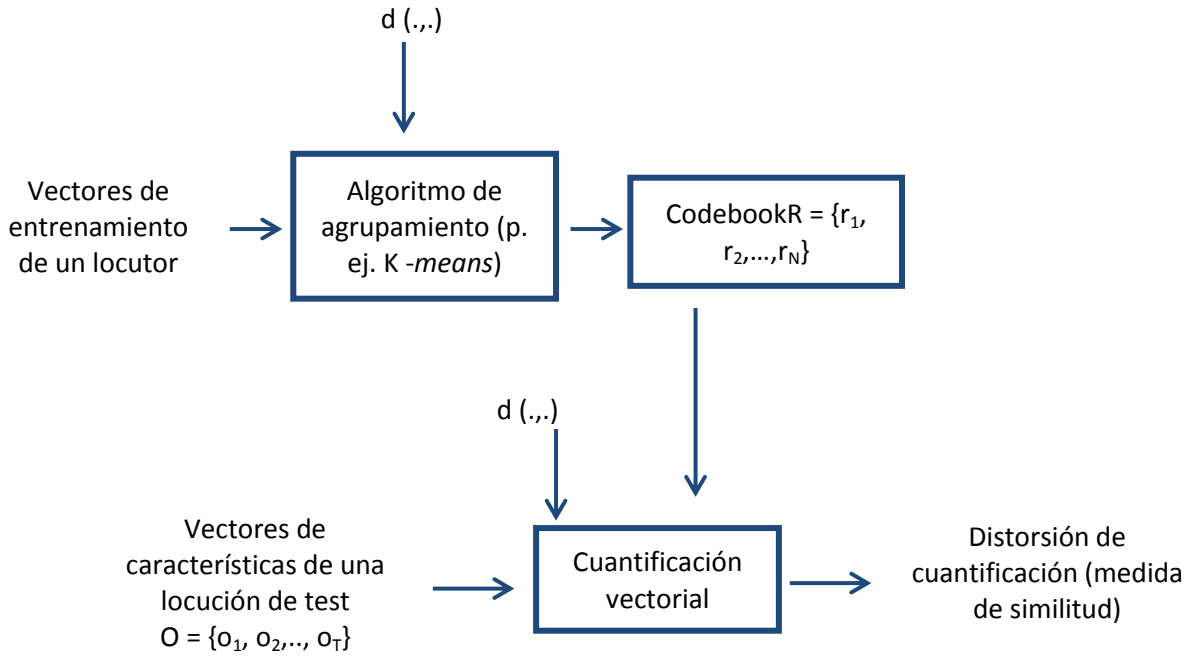

Figura 13. Proceso de entrenamiento del *codebook* de un locutory su comparación con una locución de test mediante cuantificación vectorial

2.6.2 Sistemas basados en Modelos de Mezclas de Gaussianas (GMMs, Gaussian Mixture Models)

Los GMMs [Reynolds and Rose, 1995] [Reynolds et al., 2000] es un *modelo estocástico*[1] el cuál ha sido la durante muchos años el método de referencia para el reconocimiento de locutor independiente de texto. Los GMMs pueden considerarse como una extensión del modelo VQ, en los que existe solapamiento entre las regiones que divide el espacio de características. Así un vector de característica no se asigna a un único centroide, sino que tiene una probabilidad no nula de pertenecer a cualquiera de las regiones.

Dentro del contexto de reconocimiento de locutor independiente de texto, las características son continuas y por tanto no se tienen conocimiento a priori de lo que el locutor va a decir, por tanto, para implementar un detector de la razón de verosimilitud, se requiere una función de verosimilitud *p(X|λ)*, siendo en este caso un GMM.

Un GMM, se denota por λ, está compuesto por un conjunto finito de mezclas de gaussianas multivariadas que se mezclan en el espacio de características. Esta mezcla de gaussianas se representa mediante su función de densidad de probabilidad:

$$\mathrm{p}(x|\lambda) = \sum_{k=1}^{K} w_k N(x|\mu_k, \sum_k) \qquad (2.19)$$

1.Cada locutor se modela como una fuente probabilística con una función de densidad de probabilidad desconocida pero fija.

donde K es el número de gaussianas del modelo, w_k es la probabilidad a priori (peso de la mezcla) de la *k-ésima* gaussiana, y

$$\mathrm{N}(x|\mu_k,\sum_k) = \frac{1}{(2\pi)^{-D/2}|\sum_k|^{-1/2}}\exp\{-\frac{1}{2}(x-\mu_k)^T\sum(x-\mu_k)_k^{-1}\} \qquad (2.20)$$

es la función de densidad gaussiana D-variada (D dimensiones), una combinación de densidades gaussianas uni-modales con vector de media μ_k y matriz de covarianza $\sum_k$. Las probabilidades a priori están restringidas a $\sum_{k=1}^{K} w_k = 1$ con $w_k \geq 0$. La dimensión del vector media es Dx1 y, por razones de carga computacional, las matrices de covarianza de los GMM son usualmente diagonales, el cual restringe los ejes de las elipses gaussianas a la dirección de los ejes de coordenadas. Por otra parte, la naturaleza ortogonal de los coeficientes cepstrales MFCC, hace que la alta independencia entre dimensiones permita usar este tipo de matrices. La dimensión de estas matrices de covarianza es DxD.

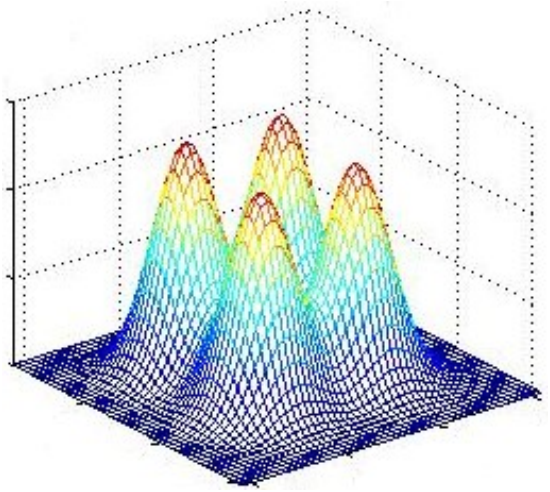

Figura 14. Función de densidad de probabilidad de un GMM de 4 gaussianas sobre un espacio bidimensional.

El entrenamiento de un GMM consiste en estimar los parámetros de un modelo $\lambda = \{w_k, \mu_k, \sum_k\}_{k=1}^{K}$ a partir de datos de entrenamiento X = {x_1, …, x_T}, y así, intentar ajustar la distribución a los vectores características de entrenamiento. Por tanto, mediante el método de estimación de máxima verosimilitud (Maximum Likelihood, ML) se pretende buscar los parámetros que maximicen la verosimilitud del GMM dado los datos de entrenamiento. Para una secuencia de datos X = {x_1, …, x_T}, y asumiendo independencia entre los vectores características:

$$\mathrm{p}(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \qquad (2.21)$$

La expresión 2.21 no es una función lineal de los parámetros λ y por tanto no se puede aplicar máxima probabilidad (ML), sin embargo, se puede aproximar mediante el algoritmo de Expectation Maximization (EM).

El algoritmo EM va cambiando iterativamente los parámetros del GMM. Así el algoritmo empieza con un modelo inicial λ y estima un nuevo modelo $\bar{\lambda} = \{\overline{w_k}, \overline{\mu_k}, \overline{\sum_k}\}_{k=1}^{K}$, de modo que $p(X|\bar{\lambda}) \geq p(X|\lambda)$. El nuevo modelo se convierte en el modelo inicial en la siguiente iteración y el proceso sigue hasta que el valor de la probabilidad converge o alcance un valor máximo de iteraciones. Los parámetros del nuevo modelo serán:

$$\overline{w_k} = \frac{1}{T}\sum_{t=1}^{T} P_r(k|x_t, \lambda) \quad \text{peso,} \tag{2.22}$$

$$\overline{\mu_k} = \frac{\sum_{t=1}^{T} P_r(k|x_t,\lambda)x_t}{\sum_{t=1}^{T} P_r(k|x_t,\lambda)} \quad \text{media,} \tag{2.23}$$

$$\overline{\sigma^2} = \frac{\sum_{t=1}^{T} P_r(k|x_t,\lambda){x_t}^2}{\sum_{t=1}^{T} P_r(k|x_t,\lambda)} - \overline{{\mu_k}^2} \quad \text{varianza,} \tag{2.24}$$

$$P_r(k|x_t, \lambda) = \frac{\overline{w_k} N(x_t|\mu_k,\sum_k)}{\sum_{i=1}^{K} w_i N(x_t|\mu_i,\sum_i)} \quad \text{probabilidad a posteriori} \tag{2.25}$$

Por otra parte, puede usarse el método *K*-means para estimar el modelo inicial λ, de forma que se necesite menos iteraciones del EM. De esta forma, los centroides calculados determinarían los vectores de medias del GMM; las matrices de covarianza de cada gaussiana estarian determinadas por la covarianza de los vectores del conjunto X asignados a cada centroide; y los pesos como el porcentaje de vectores del conjunto X asignados a cada centroide.

Con el modelo, mediante el mismo procedimiento, se calcula la probabilidad del conjunto de vectores de una locución frente al modelo.

2.6.2.1 GMM – UBM

La técnica GMM-UBM ó GMM-MAP [Reynolds et al., 2000] hace frente a dos problemas de la técnica GMM clásica:

- Solventa la escasez de datos que existe en muchas ocasiones a la hora de entrenar un modelo de locutor. De este modo, permite entrenar el modelo de forma que recoja una variabilidad acústica no contemplada en sus datos de entrenamiento.
- Proporciona un mecanismo que permite ponderar la puntuación de una locución de test en función de lo representativas de la identidad que sean esas características en cuestión.

En los sistemas de reconocimiento basados en GMM - UBM se entrena primero un modelo universal (Universal Background Model, UBM), por medio del algoritmo EM.

Este modelo representa la distribución independiente de locutor de los vectores de características, es decir, modela las caracterísitcas comunes a todos los locutores. El entrenamiento se realiza a partir de una gran cantidad de audio procedente de un gran número de locutores y de muy diversas condiciones acústicas. Cuando se registra un nuevo locutor en sistema, los parámetros del UBM se adaptan a la distribución de características del locutor, de forma que el modelo universal o UBM adaptado es el modelo del locutor.

2.6.2.2 Adaptación MAP

En la adaptación de un modelo de locutor, los pámetros del UBM que se adaptan son: pesos, vectores de medias y matrices de covarianza o tan sólo alguno de ellos. En [Reynolds et al., 2000] se demuestran que sólo adaptando los vectores de medias se consiguen buenos resultados. Dado el conjunto de vectores de locutor a registrar X = {x_1, …, x_T} y el modelo UBM $\lambda = \{w_k, \mu_k, \sum_k\}_{k=1}^{K}$, los nuevos vectores de medias adaptados (μ_k') por el método *maximum a posteriori* (MAP) se obtienen como sumas ponderadas de los datos de entranamiento de locutor, X, y las medias,μ_k,del modelo UBM.

$$\mu_k' = \alpha_k \frac{1}{n_k} f_k + (1 - \alpha_k)\mu_k \tag{2.26}$$

donde

$$\alpha_k = \frac{n_k}{n_k + \tau} \tag{2.27}$$

$$n_k = \sum_t P(k|x_t) \tag{2.28}$$

$$f_k = \sum_t P(k|x_t)x_t \tag{2.29}$$

$$P(k|x_t) = \frac{w_k N(x_t|\mu_k, \sum_k)}{\sum_{m=1}^{K} w_m N(x_t|\mu_m, \sum_m)} \tag{2.30}$$

donde n_k y f_k son los estadísticos de orden cero y primer orden respectivamente, $P(k|x_t)$ la probabilidad a posteriori de ocupación de la gaussiana. De acuerdo con la ecución 2.27, se observa que los parámetros τ(factor de relevancia)y α_k(coeficiente de adaptación) controlan la influencia de los datos de entrenamiento sobre el modelo del locutor adaptado respecto al UBM dentro del proceso de adaptación.

El modelo adaptado mediante esta técnica se conoce como GMM-MAP, y en el caso de sólo adaptar de los vectores de media, las matrices de covarianza y el vector de pesos son los mismos que el del modelo UBM. La siguiente figura muetra el resultado de adaptación de los vectores de media del modelo UBM a los datos de entrenamiento de locutor.

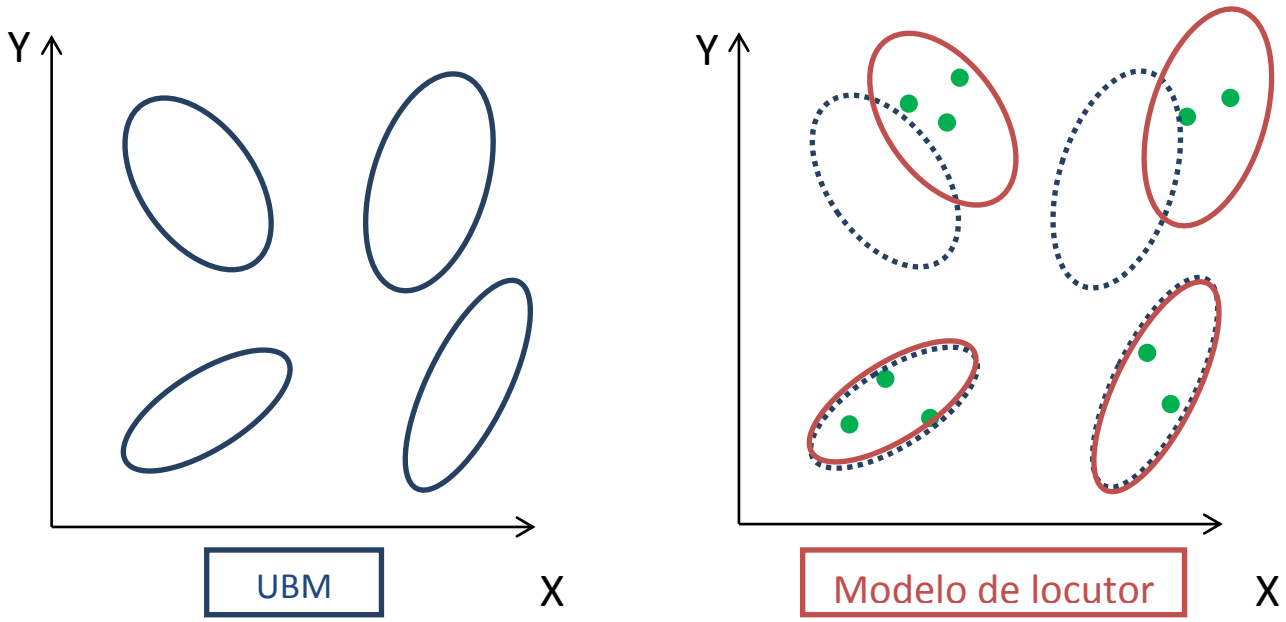

Figura 15. Proceso de adaptación MAP de medias del UBM a los datos del locutor.

En la etapa de reconocimiento, se compara la probabilidad de que los datos de test provengan del modelo del locutor adaptado, λ_{target}, entre la probabilidad de que los datos de test provengan del modelo UBM,λ_{UBM}. La puntuación o valor de verosimilitud se obtiene mediante la relación de las log probabilidades promedio:

$$LLR_{avg}(X, \lambda_{target}, \lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T}\{logp(x_t|\lambda_{target}) - logp(x_t|\lambda_{UBM})\} \quad (2.31)$$

De acuerdo con la ecuación 2.31, cuanto mayor sea el valor de la verosimilitud mayor será la probabilidad de que la identidad de la locución se corresponda con la del modelo. Adicionalmente, el empleo de un UBM en común para todos los locutores hace que las puntuaciones de los distintos locutores se encuentren en márgenes comparables, resultando una primera normalización de puntuaciones.

2.6.2.3 Supervectores

Un supervector representa de forma compacta la información de locutor presente en un GMM [Campbell et al., 2006a]. Este nuevo método de representación ha dado lugar a nuevos tipos de sistemas híbridos GMM-SVM (2.6.3.1) y técnicas de compensación de variabilidad JFA y NAP.

Un supervector consiste en la concatenación de los vectores de medias, de dimensión *1xd*, de las *K* gaussianas de un GMM, uno a continuación del otro, obteniéndose un vector de dimensión *1xKd.*La creación de los GMMs debe partir de la adaptación de un mismo modelo UBM, de forma que los supervectores estén relativamente alineados y sean comparables cuando se realizan operaciones en el espacio *Kd*-dimensional.

Esta forma de representación de una locución mediante un único punto en el espacio de supervectores cuantifica y elimina directamente del supervector la variabilidad no deseada. Este proceso se llama *compensación explícita de la variabilidad inter-sesión* y se puede realizar por varias técnicas [Burget et al., 2007] [Kenny et al., 2008] [Vogt and Sridharan, 2008]. La ventaja de estas técnicas es que no se necesita conjuntos de entrenamiento que contemplan cada tipo de canal o entorno para cada locutor, sino que se entrena un modelo de variabilidad inter-sesión independiente de locutor que luego puede aplicarse al supervector de cualquier locutor. Las técnicas de *Factor Analysis* (FA) aplican esta compensación sobre sistemas basados en GMMs.

2.6.3 Máquinas de vectores soporte (Support Vector Machines, SVMs)

Los SVMs [Cortes and Vapnik, 1995] son clasificadores discriminativos muy potentes que han comenzado a usarse en reconocimiento de locutor con características espectrales [Campbell et al., 2006a] [Campbell et al., 2006b] como prosódicas y de alto nivel [Campbell et al. 2004].

Un SVM es un clasificador binario que modela la frontera la decisión entre dos clases mediante un *hiperplano separador*. El hiperplano puede ser lineal y no lineal. En el caso de datos que no se pueden separarse por una frontera lineal en el espacio de características, dichos datos se transforman a otro espacio de características de mayor dimensión, por medio de una función *kernel.* En el nuevo espacio, las dos clases sí pueden ser separadas por una frontera lineal.

En un sistema de verificación, una clase consiste en los vectores de características para el entrenamiento del locutor a verificar (clase target, etiquetados como +1), y la otra clase consiste en los vectores de características para el entrenamiento de la población de impostores o *background* (clase non-target, etiquetados como -1). Con los vectores de entrenamiento etiquetados, el SVM modelará un hiperplano que maximice el *margen* de separación entre las dos clases, dado por la distancia entre los *vectores soporte* de cada clase.

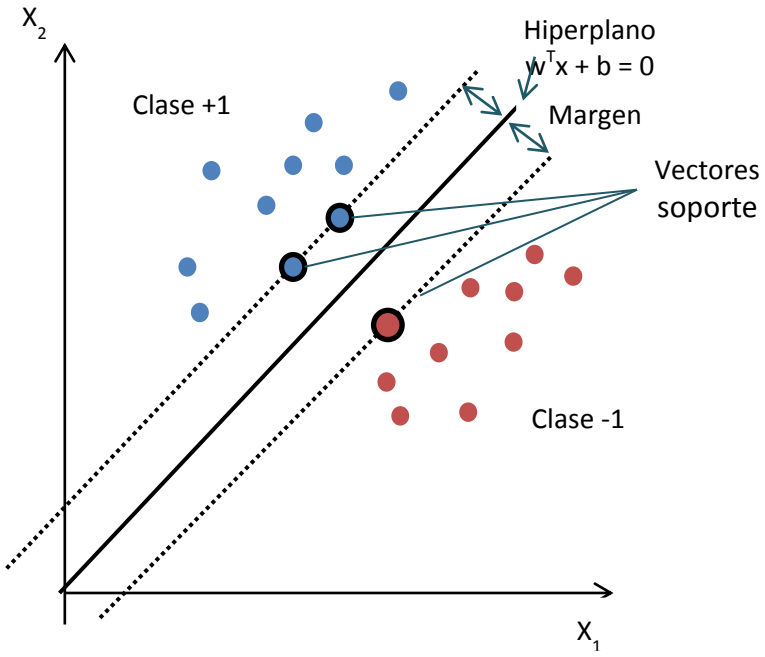

Figura 16. Representación de los elementos de un SVM.

En un espacio n-dimensional la frontera está representada por un plano denotado por {$\vec{w}$, b}, donde $\vec{w}$ es un vector de *n* coeficientes que determina la orientación del plano y *b* es el término independiente de la ecuación paramétrica del plano (véase Figura 16).

En un sistema de verificación, el modelo SVM de locutor es el plano {$\vec{w}$, b} que separa de forma óptima los vectores de características del locutor a verificar y los vectores del resto. La función de puntuación para un SVM respecto a un vector de características $\vec{x}$ se define:

$$f(\vec{x}) = w^T x + b \tag{2.32}$$

En la fase de entrenamiento del SVM se busca el mejor ajuste del plano {$\vec{w}$, b} de forma que, para el conjunto de datos de entrenamiento, se cumplan que $f(\vec{x}) \geq 1$ para los vectores de clase *target* y $f(\vec{x}) \leq -1$ para los de clase *non-target*. En la fase de evaluación, la clasificación se basará los valores que tome la función de puntuación $f(\vec{x})$, los vectores que cumplan $f(\vec{x}) \geq 0$ pertenecerán a la clase target y $f(\vec{x}) < 0$ pertenecerán a la clase non-target.

2.6.3.1 Sistema hibrido GMM-SVM

Está técnica da origen al concepto de *SuperVectors* [Campbell et al., 2006a][Krause and Gazit, 2006]. Fue propuesta por RanGazit, y en los últimos años ha demostrado su gran capacidad para el reconocimiento de patrones. Aprovecha los dos puntos fuertes de cada sistema: el modelado generativo de los sistemas GMM-UBM y el modelado discriminativo de los sistemas SVM.

El funcionamiento de este sistema híbrido se basa en el uso supervectores, procedente de los modelos de locutor GMM-MAP, como vectores de entrada para la clases *target* y *non-target* de entrenamiento de un SVM. Para ello, es necesario disponer de un modelo de la locución involucrada tanto para el entrenamiento como para test, y así obtener el supervector correspondiente.

2.6.4 Técnicas de Factor Analysis

En los últimos años, los sistemas de verificación de locutor independiente de texto se han centrado en el uso de técnicas basadas en *Factor Analysis* gracias a su habilidad para tratar con la variabilidad de sesión. A continuación se presentará una descripción de las técnicas enmarcadas dentro de este marco.

2.6.4.1 Joint Factor Analysis

La técnica Joint Factor Analysis (JFA) se basa en el modelado conjunto tanto de la variabilidad intra-locutor como la debida al canal [Kenny, 2006]. El modelo de locutor se puede representar por un supervector. Este supervector es la concatenación de las medias de los GMM del locutor, previa adaptación MAP únicamente de las medias a partir un único modelo UBM. Por tanto, los modelos adaptados comparten el vector de pesos y las matrices de covarianzas.

Para un locutor determinado, los supervectores obtenidos de distintas locuciones de entrenamiento pueden diferir debido a la variabilidad de canal de transmisión. Por tanto, se requiere una compensación de esa variabilidad de forma que datos procedentes de un canal distinto al del entrenamiento puedan ser comparados correctamente con el del modelo de locutor. Para ello, es necesario modelar la variabilidad de forma explícita. El modelado JFA asume que hay una variabilidad no deseada dentro de un subespacio de baja dimensionalidad que modifica el supervector de locutor *s* para una locución *h*.

$$\mu_{sh} = \mu_s + Ux_h \tag{2.33}$$

donde **U** representa el subespacio de variabilidad de sesión. La componente de x_h son los factores de canal o *channel factors* y dependen de la locución ***h***, estos se estiman a partir de los datos de entrenamiento del locutor y determinan la importancia de cada dirección de variabilidad en **U**. Las columnas de la matriz **U** se denominan *eigenchannels* y son estimadas a partir de un conjunto de datos entrenamiento con gran variabilidad de canal

Respecto al supervector de locutor $\boldsymbol{\mu_s}$, éste se puede descomponer de la siguiente forma:

$$\mu_s = \mu + Vy_s + Dz_s \tag{2.34}$$

donde:

- $\boldsymbol{\mu}$: es el supervector de medias del UBM
- ***V***: contiene la varianza de locutor y es una matriz rectangular cuyas columnas se denominan *eigenvoices*.
- $\boldsymbol{y_s}$: son los pesos que representan al locutor ***s*** en el subespacio de variabilidad de locutor expandido por ***V***. Cada componente del vector $\boldsymbol{y_s}$ se les denomina *speaker factors.*
- ***D***: representa el desplazamiento offset del supervector***m*** como resultados de la adaptación MAP. Lo conforma una matriz diagonal de (*Kd* x *Kd*) y $\boldsymbol{z_s}$ es un vector columna de *Kd* x 1.

De acuerdo con la expresión 2.34, si $\boldsymbol{y_s}$ = 0, $\mu_s = \mu + Dz_s$, describe el proceso de adaptación MAP. Por tanto, la técnica JFA es una expansión de la técnica MAP con la inclusión del modelado *eigenvoice* Vy_s, el cual resulta útil en el caso de tener poco datos de entrenamiento del modelo. Dicho producto restringe la adaptación de medias en el entrenamiento del modelo del locutor a las direcciones dadas por y_s y dentro del espacio determinado por V.

En el proceso de identificación, se obtienen los denominados estadísticos de orden cero (vector de dimensión *1*x*K*) y primer orden (vector de dimensión *1*x*Kd*) de la locución test frente al UBM, dados por las ecuaciones 2.28 y 2.29 respectivamente.

La probabilidad de que la locución de test corresponda al locutor se obtiene mediante el producto escalar del supervector del modelo (dimensión *Kd*) por el estadístico de primer orden, normalizado por la suma de los elementos de los estadísticos de orden cero.

2.6.4.2 i-vectors

Los sistemas basados en i-vectors se han convertido en una técnica del estado del arte en los sistemas de verificación de locutor debido a la capacidad de reducción de la gran dimensión de los datos de entrada hacia una menor dimensión de los vectores de características reteniendo la información más relevante.

$$s = m + Tw \tag{2.35}$$

De acuerdo con la expresión 2.35, los sistemas i-vector usan un conjunto de factores de Total Variability ***w*** para representar la locución de un determinado locutor, es decir, un supervector ***s***. Cada factor controla una eigen-dimension de la matriz de Total Variability ***T.***

3. Descripción del sistema

3.1 Introducción

En este capítulo se presenta una descripción del sistema de verificación de locutor implementado. Se describirá los tipos de parametrización de las características usados en el sistema; así como, los tipos de modelado basados en GMM-UBM. Para los tipos de modelado, previamente se debe disponer de un modelo universal UBM. Para ello, es necesario disponer una gran base de datos, orientadas al reconocimiento de locutor, para recoger la mayor variabilidad de locutores posible y mayor número de factores de variabilidad relacionados a condiciones acústicas y tipo de habla (múltiples dialectos, habla conversacional, espontánea, etc.).

Por otra parte, para medir el rendimiento del sistema de verificación es necesario establecer unas pautas y medidas objetivas normalizadas que permitan a la comunidad científica analizar y estudiar la forma en que se ha llegado a los resultados finales (bases de datos, protocolos de evaluación, función de coste); así como, replicar las mismas pruebas por terceros y poder establecer comparaciones con sus propios sistemas implementados.

3.2 Extracción de características

En este apartado se describirá los tipos de parametrización y proceso de extracción usados en el sistema reconocimiento de locutor implementado.

3.2.1 Segmentación con descifrador SRI's

En la parte experimental del proyecto, el reconocimiento se realizará sobre unidades lingüísticas (fonemas, difonemas y trifonemas) del léxico inglés, por tanto, es necesaria la transcripción cada unidad para realizar posteriormente la segmentación. La transcripción se obtiene del descifrador SRI's (Decipher conversational telephone speech recognition system) [Kajarekar et al, 2009]. Esta transcripción contiene el contenido fonético y el intervalo de duración dentro de una región de la locución.

Se ha empleado 41 fonemas del léxico inglés, donde cada fonema se representa por su respectivo código Arpabet [Arpabet]. Las unidades de difonemas se crean a partir de la combinación de estos fonemas, obteniéndose unos 98 difonemas, los cuales son los de mayor frecuencia de ocurrencia en las locuciones. De la misma forma, para las unidades de trifonemas se obtiene 23 trifonemas.

3.2.2 Extracción de coeficientes cepstrales por unidad

En el apartado 2.4.1 se explicó el proceso de extracción de coeficientes cepstrales para locuciones de audio. En este proyecto se ha usado locuciones de duración aproximada de 5 minutos. Esta extracción se realiza mediante enventanado de tramas de 20 milisegundos solapadas al 50 % o cada 10 milisegundos, obteniéndose así vectores de características que contienen los coeficientes cepstrales. Dado que el proyecto pretender evaluar el rendimiento del sistema con vectores de características por unidad lingüística (fonema, difonema y trifonema), es necesario realizar la segmentación con las transcripciones obtenidas a partir del descifrador SRI's. Adicionalmente, se ha empleado los 19 primeros coeficientes cepstrales (información estática) más sus respectivas derivadas temporales (información dinámica). Estos coeficientes están mediante CMN, filtrado RASTA y *feauture warping*.

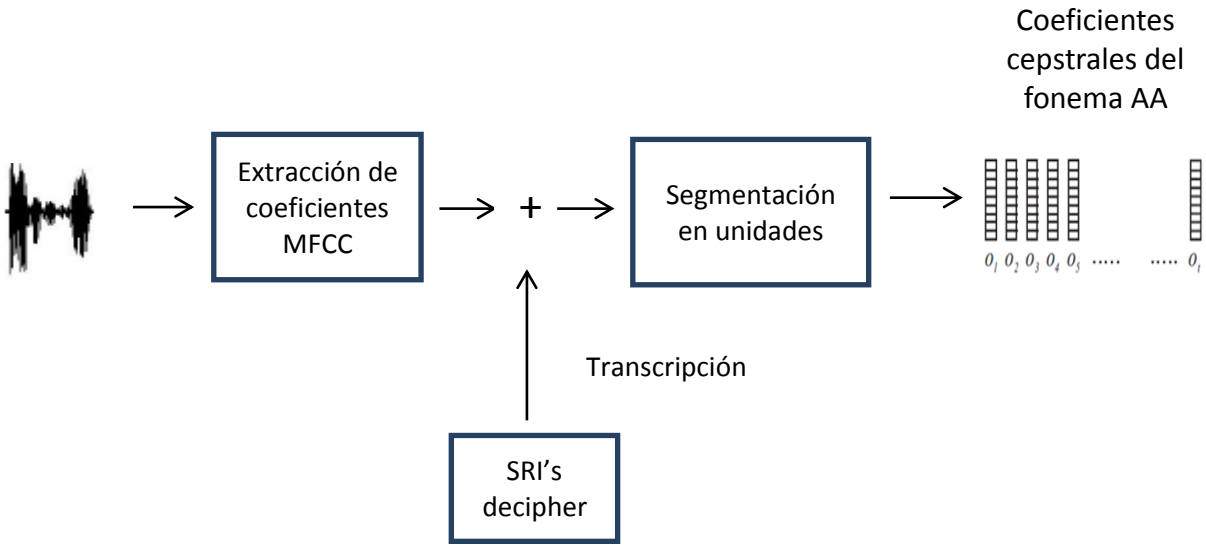

Figura 17. Esquema de extracción de coeficientes cepstrales por unidad.

3.2.3 Extracción de formantes y anchos de banda por unidad

En la producción de sonidos del habla, dentro del tracto vocal las cavidades acústicas son excitadas por la laringe produciendo resonancias denominadas formantes. Éstos permiten estimar y modelar eficientemente las resonancias espectrales del tracto vocal de una persona mediante un conjunto limitado de parámetros: frecuencia de resonancia o formante, ancho de banda y energía. La frecuencia de resonancia se puede visualizar en la envolvente espectral de la señal de voz, siendo las frecuencias de formantes los máximos relativos de dicha envolvente.

Debido a las singularidades de cada tracto vocal, cada individuo presenta unos valores característicos de frecuencias formánticas y contornos temporales de dichas frecuencias (3.2.4.2).

Es decir, cada sonido del habla tendrá una envolvente espectral característica y diferenciadora respecto al resto de sonidos. Por tanto, los sistemas de reconocimiento automático de locutor tienen como objetivo paramétrizar adecuadamente la envolvente espectral para posteriormente alcanzar un grado de discriminación entre los distintos sonidos del habla.

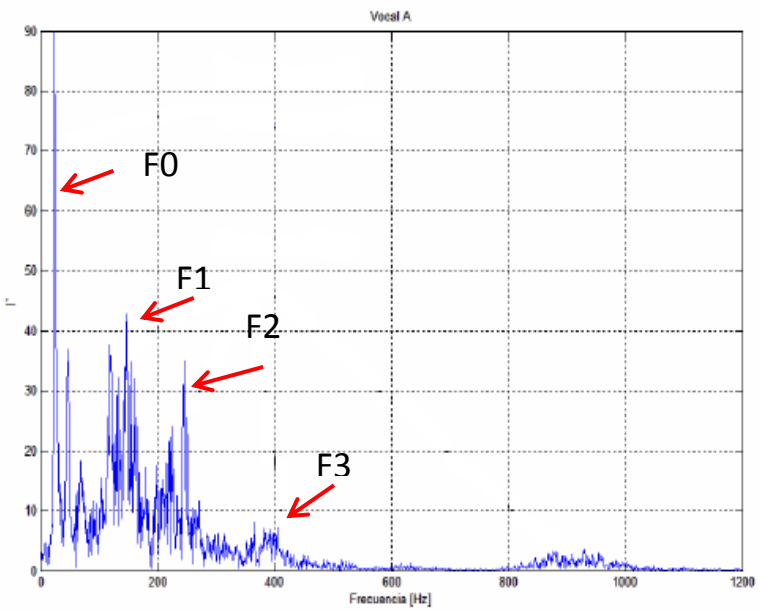

Figura 18. Envolvente espectral de la vocal "A".

A la frecuencia de resonancia más baja se designa *F1*, las frecuencias de resonancias mayores se designan *F2*, *F3*, *F4*, *F5*, *F6*, etc. Adicionalmente, existe la frecuencia fundamental, *F0*, o también denominada *pitch*. Esta frecuencia brinda información sobre la velocidad de vibración de las cuerdas vocales al producir un sonido. En la Figura 18 se muestra la frecuencia fundamental, *F0*, y las tres primeras frecuencias de formantes correspondientes a los picos máximos de la envolvente espectral de la señal.

Otra forma de poder visualizar las frecuencias de los formantes es a través de un espectrograma, el cual representa la frecuencia en función del tiempo. En la Figura 21 se puede apreciar que las zonas más oscuras (*zonas de mayor concentración de energía*) corresponden con las frecuencias de los formantes. La zona de mayor energía o la más oscura corresponderán con la primera frecuencia de resonancia *F1*, la segunda zona con la segunda frecuencia *F2* y así sucesivamente.

3.2.3.1 Proceso de extracción

El proceso extracción de formantes no es una tarea trivial. Los mejores resultados se obtienen con la extracción y etiquetado manual hecha por expertos fonetistas, no obstante, al tener un gran número de grabaciones a examinar, la tarea de extracción demandaría un gran consumo de tiempo. No obstante, se han desarrollado aplicaciones que permiten la extracción automática de las frecuencias y anchos de banda de formantes.

Para este proyecto se ha empleado el *Wavesurfer* [Sjolander and J. Beskow, 2000]. Esta herramienta de código abierto permite extraer los formantes y sus respectivas trayectorias. Una vez extraído las frecuencias y anchos de banda de formantes, y con la transcripción de las unidades, obtenida a partir del descifrador SRI's, se procede a la segmentación. Esta tarea de extracción por unidad se realizó con scripts desarrollados por el grupo ATVS.

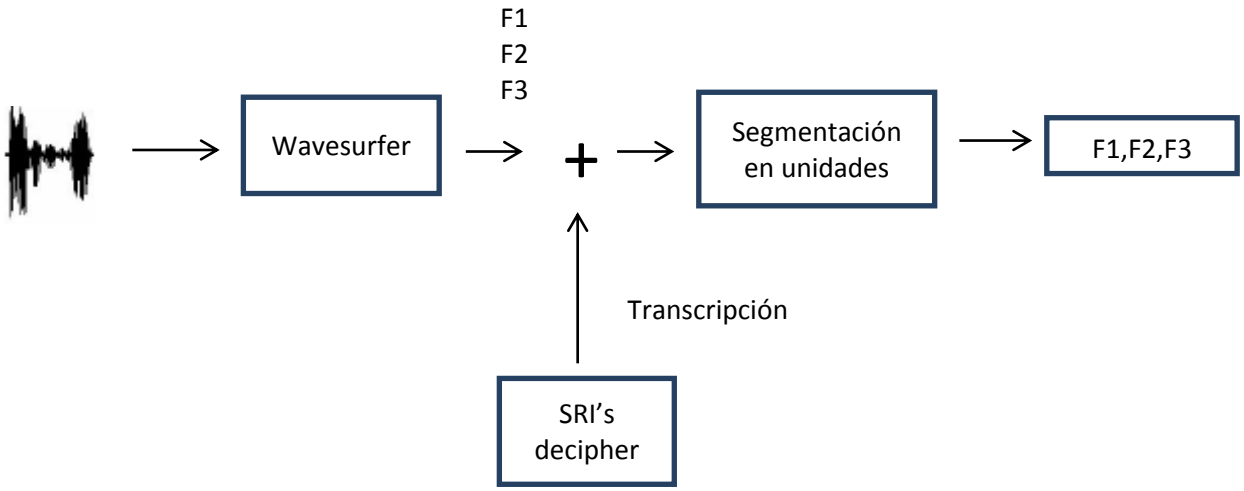

Figura 19. Esquena de extracción de formantes por unidad.

Para la parte experimental del proyecto se ha empleado las tres primeras frecuencias de formantes (F1, F2 y F3) más sus respectivos anchos de banda (Bw1, Bw2 y Bw3).

3.2.4 Contorno Temporal en Unidades Lingüísticas (Temporal Countour in Linguistic Units, TCLU)

Tanto para coeficientes cepstrales como para los formantes extraídos se puede añadir información dinámica (primera o segunda derivada temporal para los coeficientes cepstrales y anchos de banda para los formantes) de la locución en los vectores de características. La incorporación de esta información recoge mayor variabilidad del segmento de locución. Sin embargo, una vez extraído los vectores de características, la información entre los mismos vectores y su información lingüística no se usa. Es a partir de este punto donde nace este nuevo enfoque basado en el modelado explícito de la evolución temporal de cada coeficiente del vector de características a nivel de unidad. Con este tipo de parametrización se enriquece la información recogida por cada vector de características, ya que se recoge la información estática, dinámica y evolución temporal de cada coeficiente MFCC o frecuencia de formante por unidad lingüística.

3.2.4.1 Parametrización de contornos temporales de coeficientes cepstrales por unidad (TCLU-MFCC)

Una vez extraídos los coeficientes cepstrales por unidad, se obtienen vectores de características correspondientes a tramas de longitud variable por unidad. Esta variabilidad de la duración se debe a que cada locutor produce distintos patrones de duración para cada unidad lingüística, citando como ejemplo, los fonemas "AA" dentro de una misma locución pueden tener distintas duraciones.

Dado que la parametrización del contorno temporal debe ser independiente de la duración de la unidad lingüística, como paso previo, se aplica una fase *ecualización*; es decir, la longitud variable de cada trama es *ecualizada* mediante operaciones de diezmado e interpolación a un número de tramas equivalente a una duración de 250 milisegundos. La elección de esta duración se basa en los estudios realizados [Morrinson, 2009] [Castro et al, 2009]. Este proceso permite obtener trayectorias de los coeficientes cepstrales correspondientes a tramas de longitud fija. Con las tramas ecualizadas se extrae el contorno temporal o trayectorias de los coeficientes.

Llegado a este paso se procede a la parametrización de los contornos temporales mediante ajuste polinomial o la transformada discreta del coseno, DCT. Mediante la DCT se lo logra aproximar o paramétrizar los contornos temporales por medio de funciones de cosenos que oscilan a distintas frecuencias. Esto permite que la señal de la información "codificada" o parametrizada se concentre en las bajas frecuencias, es decir, esta información es compactada en pocos coeficientes eliminando así las componentes que generan ruido en las trayectorias originales; por otra parte, los coeficientes DCT presentan propiedades pseudo-ortogonales.

Para este proyecto, la parametrización de los contornos temporales se realiza con el DCT de orden 5. La elección de este número de coeficientes se basa en trabajos previos [González-Rodríguez, 2011] y pruebas realizadas en este proyecto usando DCT de orden 7 y 9. Finalmente, se obtiene vectores de características por unidad de dimensión igual a 95 (19 MFCC o delta MFCC x 5 DCT coefs/trayectoria) o 190 (19 MFCC + 19 delta MFCC = 38 x 5 DCT coefs/trayectoria).

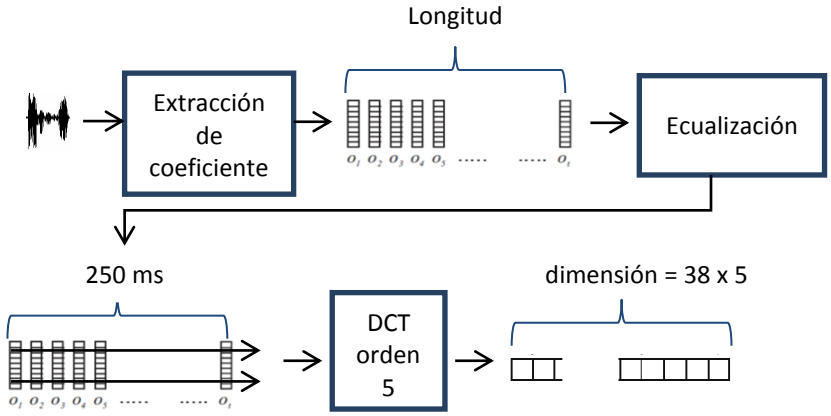

Figura 20. Parametrización de contornos temporales de coeficientes cepstrales en unidades lingüísticas.

3.2.4.2 Parametrización de los contornos temporales de formantes por unidad (TCLU-Formantes)

La parametrización de los contornos temporales de formantes sigue un esquema paralelo al procedimiento de parametrización de contornos temporales de coeficientes cepstrales. Una vez obtenidos los formantes por unidad (3.2.3.1) se procede a la extracción de los contornos temporales a través de la herramienta de extracción *Wavesurfer*. La frecuencia de formantes se selecciona entre los candidatos propuestos por la solución de las raíces del polinomio generado a partir de la predicción lineal periódica de coeficientes. La estimación de los formantes mediante programación dinámica permite optimizar las trayectorias estimadas mediante restricción continua de frecuencias. Las trayectorias de las frecuencias de formantes se puede observar en un espectrograma (Figura 21). Éste representa dichas frecuencias en función del tiempo, es decir su evolución temporal. Las zonas más oscuras, zonas de mayor concentración de energía, corresponden a las frecuencias de formantes.

Como paso previo a la parametrización de los contornos temporales, éstos son ecualizados a una duración de 250 milisegundos. Este procedimiento es similar al explicado en el apartado anterior. Posteriormente, se aplica la DCT de orden 5 y se obtiene un vector de características por unidad de dimensión 15 (3 primeras frecuencias o anchos de banda de formantes x 5 DCT coefs/trayectoria) o 30 (3 primeras frecuencias + respectivos achos de banda = 6 x 5 DCT coefs/trayectoria).

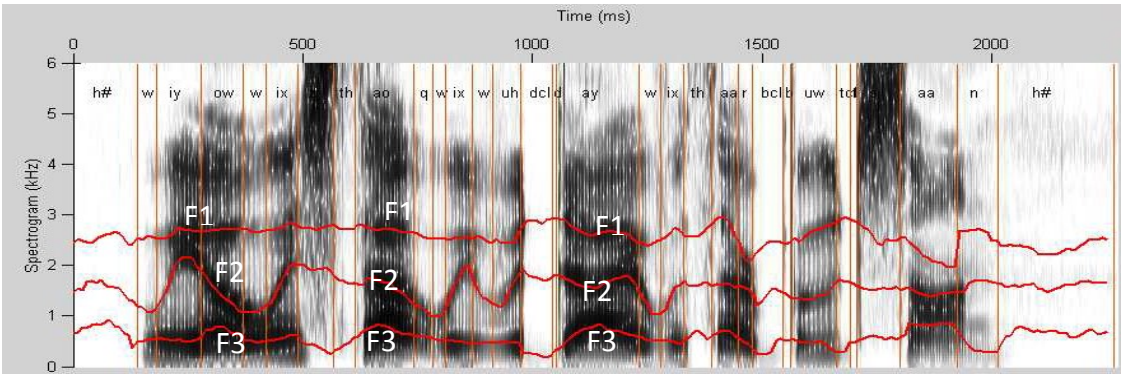

Figura 21. Espectrograma de una señal con las frecuencias de formantes estimadas con Wavesurfer.

3.3 Modelado

Los sistemas GMM-UBM han sido el estado del arte en el reconocimiento de locutor en texto independiente por muchos años hasta la aparición de técnicas como JFA (Joint Factor Analysis) [Kenny et al., 2008] y total variability [Kinnunen and Li, 2010]. Los cuales ha desplazado a los sistemas GMM-UBM debido a la capacidad de modelar la variabilidad el espacio de supervectores.

Para este proyecto, el sistema de reconocimiento de locutor propuesto usa el método GMM-UBM para el modelado estadístico de un conjunto de características extraídas. Este método ha sido seleccionado por tres razones:

- Se usa por primera vez los contornos temporales como vectores de características sobre los sistemas GMM-UBM. Por ello, se busca la configuración óptima sobre estos sistemas, los cuales son la base para los sistemas Supervectors.
- Se pretende evaluar el rendimiento del sistema con dependencia de unidad, el cual es unos de los principales motivos de este proyecto.
- Dado que el conjunto de datos de entrenamiento no es suficiente para poder modelar la variabilidad existente entre el espacio de características de cada unidad lingüística.

En la parte experimental se ha usado dos tipos de configuraciones del modelado GMM-UBM: Global GMM-UBM y Constrained GMM-UBM. La diferencia entre cada configuración es el tipo de datos empleados en el sistema, es decir, datos dependientes y/o independientes de unidad usados en: entrenamiento de los UBMs, entrenamiento de modelos de locutor y datos de test.

Por otra parte, teniendo en cuenta el tipo de parametrización usada, se tiene dos conjuntos de datos:

- Coeficientes cepstrales y frecuencias/anchos de banda de formantes por unidad.
- Contornos temporales de coeficientes cepstrales y frecuencias/anchos de banda de formantes por unidad.

Una vez entrenado el modelo UBM se procede a la adaptación MAP de los modelos de locutor.

En la parte final de este apartado, se comentará sobre los métodos de fusión de sistemas usados en la parte experimental del proyecto.

3.3.1 GMM-UBM Global

En esta configuración, el modelo UBM y el modelo de locutor serán independientes de unidad; de la misma forma los datos de test. Con este tipo de modelado se pretende evaluar el rendimiento del sistema mediante el uso de contornos temporales de las características extraídas de cada locutor a nivel de locución. Para la creación del modelo universal UBM se usa los datos llamados ***development***, los cuales se corresponden con las bases de datos de las evaluaciones NIST SRE 2004 y 2005. De la misma forma, para el entrenamiento del modelo de locutor se usa todas las locuciones segmentadas por unidad agrupadas en una sola locución; estos datos se denomina ***train*** y se corresponden con la base de datos de la evaluación NIST SRE 2006 (3.4.1.1).

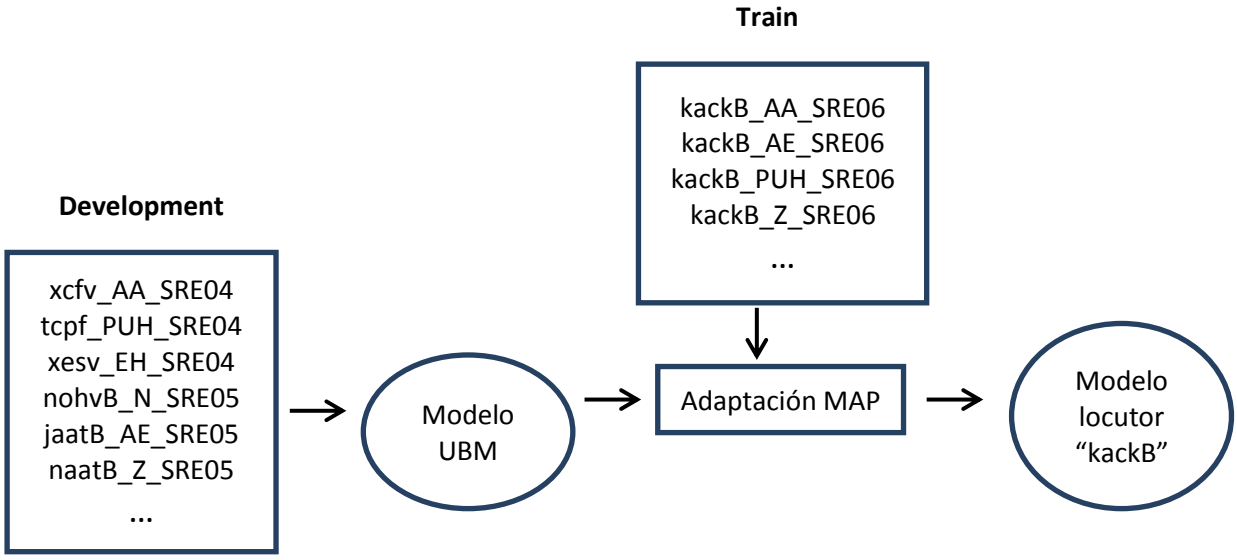

Figura 22. Esquema de entrenamiento de modelos UBM y del locutor "kackB" para sistema Global GMM-UBM.

La Figura 22 representa el esquema de entrenamiento de los modelos UBM y de locutor. Dado que todas las locuciones están segmentadas en unidades, para el entrenamiento del modelo UBM agrupamos todas las locuciones correspondientes a la base de datos de NIST SRE 2004 y 2005. Para el entrenamiento del modelo de locutor, agrupamos las locuciones segmentadas correspondientes al locutor que se desea adaptar desde el modelo UBM.

Para esta configuración se ha probado distintos números de mezcla gaussianas, desde 8 hasta 1024 mezclas con incrementos de potencia de 2.

3.3.2 GMM-UBM Constrained

En esta configuración, el modelo UBM y el modelo de locutor serán dependientes de unidad. Siguiendo el esquema de la Figura 23, para el entrenamiento del modelo UBM y modelo de locutor se emplearán las locuciones segmentadas que correspondan a la misma unidad en la que se esté trabajando.

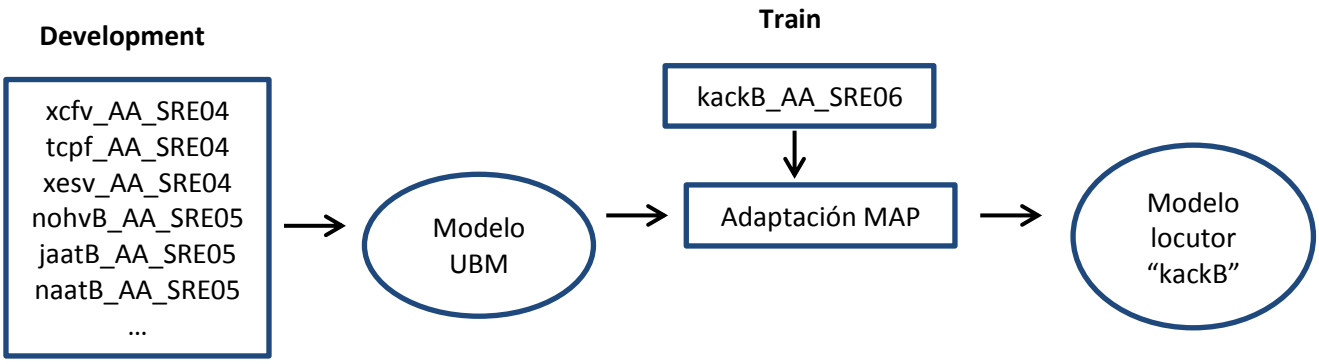

Figura 23. Esquema de entrenamiento de modelos UBM y del locutor "kackB" dependientes de fonema "AA" para sistema Constrained GMM-UBM.

Dentro de este conjunto se ha probado tres configuraciones de GMM-UBM dependiente de unidad usando contornos temporales de coeficientes cepstrales.

- Modelo UBM y modelo de locutor entrenados a partir de datos independientes de unidad; datos test dependientes de unidad.
- Modelo UBM entrenado a partir de datos independientes de unidad; modelo de lo locutor adaptado con datos dependientes de unidad; datos test dependientes de unidad.
- Modelo UBM y modelo de locutor entrenados a partir de datos dependientes de unidad; datos test dependientes de unidad.

Para cada configuración se ha probado distintos número de mezclas gaussianas, desde 2 hasta 1024 mezclas con incrementos de potencia de 2.

3.3.3 Fusión de sistemas y combinación de unidades lingüísticas

El enfoque de proyecto se centra en la evaluación del rendimiento del sistema a nivel de unidad lingüística. Esto nos permite obtener razones de verosimilitud LR para cada muestra de locución a corto plazo (20 ms), en donde actualmente no es aplicable directamente en el estado del arte de los sistemas de reconocimiento de locutor. Adicionalmente, se ha probado la combinación de unidades lingüísticas con el fin de obtener mejor capacidad de discriminación en el sistema. Esta combinación, o también llamado fusión, se ha realizado de dos modos: modo **intra-unidad** y modo inter-unidad. En el primero, la fusión sólo se realiza con unidades del mismo tipo, como por ejemplo la fusión de unidades de fonemas.

En el segundo modo, la fusión se realiza a nivel de varios tipos de unidades o tipos de características, como ejemplo, la fusión de fonemas y difonemas, o fusión de contornos de formantes y contornos de coeficientes cepstrales.

Las unidades a fusionar se seleccionan de modo que se agrupen aquellas que mejor capacidad de discriminación ofrecen al sistema. En este proyecto se ha seleccionado las unidades mediante dos métodos:

- **Método 1:** este método selecciona aquellas unidades que estén por debajo de un umbral de EER dejando afuera aquellas unidades que no cumplen o superan dicho umbral. Sin embargo, este procedimiento de selección no garantiza alcanzar la mejor fusión ya que las unidades con valores de EER muy cercanos al umbral podrían contribuir al sistema fusionado si sus valores de LR's presenta poca correlación con los demás.

- **Método 2:** este método se basa en un algoritmo de selección que permite encontrar complementariedad entra las unidades. Los pasos de describen a continuación:

    1) Seleccionar la mejor unidad que presente el valor de EER muy por debajo del umbral. Esta unidad será el conjunto de las unidades a fusionar.
    2) Seleccionar la siguiente mejor unidad y fusionar con el conjunto de las unidades a fusionar. Si la fusión mejora el rendimiento del conjunto esta unidad se añade al conjunto, en caso contrario se descarta.
    3) El paso 2 se repite para todas las unidades en orden ascendente respecto a sus valores de EER.

Respecto a las técnicas de fusión se han empleado dos técnicas: fusión suma promedio y fusión mediante regresión logística lineal:

- **Fusión suma promedio**: esta técnica se basa en la suma promedio de los scores obtenidos de la evaluación del sistema. Como paso previo a la fusión, se calibra los scores mediante regresión logística lineal.

  Obtenidos los valores de LR, se procede a la suma entre cada LR pertenecientes al mismo *trial* y finalmente el resultado se divide entre el número de LR's sumados.

- **Fusión mediante regresión logística lineal:** está técnica aplica calibración y fusión en un mismo paso. Al igual que la fusión suma, la calibración se realiza mediante regresión logística lineal [Brümmer and Preez, 2006].

3.4 Entorno experimental

3.4.1Protocolos de evaluación

Un protocolo de evaluación se define como el conjunto condiciones que se imponen a los sistemas a implementar. Estas condiciones determinan la base de datos a utilizar para entrenar los modelos de locutor reconocer y testear, como también la duración del audio usado en la evaluación. Además, determina el número de enfrentamientos realizados por cada identidad a reconocer y la proporción de locutores por género. De esta forma se mide de forma objetiva el rendimiento del sistema a evaluar.

El desarrollo de este proyecto ha seguido el protocolo de evaluación del *National Institute of Standards and Technology (*NIST Speaker Recognition Evaluation, NIST SRE) [NIST SRE].

3.4.1.1 Evaluación NIST

El objetivo de las evaluaciones NIST es impulsar el desarrollo tecnológico, medir el estado del arte y encontrar técnicas novedosas que hagan frente a los nuevos desafíos que se presentan en las tareas de reconocimiento de locutor independiente de texto. Para ello establece unas condiciones competitivas protocolares que permiten determinar el rendimiento de los sistemas participantes y así compara las distintas técnicas y configuraciones empleadas.

Las evaluaciones NIST se han venido organizando anualmente desde 1996 hasta 2006, a partir de ese año las evaluaciones han pasado a ser bianuales. A lo largo de los años las condiciones de la evaluación han crecido, pasando de tener datos sólo de canal telefónico a incorporar también del tipo micrófono, de haber estilo de habla conversacional a incluir habla de tipo entrevista, de tener locutores de un único idioma a ampliarlos a varios idiomas, etc. El incremento de condiciones ha descubierto un nuevo frente de desafíos en las tareas de reconocimiento, provocando su evolución. Además, el las evaluaciones han pasado a ser de carácter abierto, permitiendo la participación de grupos de investigación, empresa o entidad, con la obligación de presentar sistema desarrollado. De esta forma se consigue una mayor competitividad e impulso en la investigación en el desarrollo de sistemas de reconocimiento de locutor.

El gran impacto de estas evaluaciones en la comunidad de reconocimiento de locutor e idioma ha provocado que sus conjuntos de datos y protocolos de evaluación se hayan convertido en un estándar al momento de publicar resultados en publicaciones de este ámbito.

El protocolo de evaluación define la medida del rendimiento (función de coste) y los datos sobre los que realizar las decisiones de evaluación: datos de entrenamiento (segmentos de train) para modelar la identidad a reconocer y datos de test (locución de prueba, o segmentos de test) para cotejar con los modelos de locutor generados.

Por ello, existen diversas condiciones en función de la cantidad y tipo de datos que se dispone para el entrenamiento y test:

- *Datos de entrenamiento*: desde 10 segundos (10s) de habla hasta 8 conversaciones (8c) de 5 minutos cada una (2,5 minutos de habla neta por locutor).

- *Datos de test*: desde 10 segundos (10s) de habla hasta 1 conversación (1c) de 5 minutos (2,5 minutos de habla neta por locutor).

La combinación entra una determinada cantidad de habla de entrenamiento y una cantidad determinada de habla de test se denomina *condición* de prueba. Además, en la mayoría de las condiciones se proporciona los ficheros de audio de ambos locutores presentes en la conversación por separado (condiciones *2-channels* o *4-wire*), sin embargo existen otras en que las conversaciones para entrenamiento o test pueden estar mezcladas en un mismo canal, éstas son las condiciones *summed-channel*.

Respecto a la medida del rendimiento del sistema, se emplea la función de detección de coste DCF, definido en el apartado 2.5.2.5. En cada evaluación NIST se proporcionan el coste de falsa aceptación C_{FA} y falso rechazo C_{FR}, y se establece la probabilidad a priori de que una locución de test dada pertenezca al locutor en cuestión, P_{Tar}. Para la evaluación NIST 2006, los valores fijados a priori fueron: $C_{FR} = 1$,$C_{FA} = 10$ y $P_{Tar} = 0{,}01$.

3.4.2 Bases de datos para reconocimiento de locutor

El rendimiento de los sistemas se ve afectado en gran medida por la disponibilidad de la base de datos, ya que a partir de ésta se entrenan los UBMs, matrices de compensación para la aplicación de FA, las cohortes para Z-normalización y T-normalización, etc.

Es deseable que la base de datos tenga una gran población de locutores y recoja la mayor cantidad de factores de variabilidad posible de la señal de voz: multi-sesión (grabaciones obtenidas en distintos momentos de tiempo); múltiples lenguajes; múltiples condiciones ambientales (con y sin ruido de fondo); múltiples canales de grabación (telefónico, microfónico,…); etc.

A continuación, se describe las bases de datos empleadas en el proyecto tanto para el desarrollo del sistema como para su evaluación del rendimiento [Ramos-Castro, 2007].

- **Switchboard 1:** contiene habla conversacional en inglés americano grabada sobre la línea telefónica convencional. Recoge variabilidad producida por el uso de diversas líneas de teléfono y los distintos tipos de terminales telefónicos (micrófono tipo electrect, de carbón, etc). Esta base de datos fue empleada en la evaluación NIST 2001.

- **Switchboard 2:** es muy similar que el Switchboard 1 con la diferencia que recoge mayor variabilidad entre distintas líneas y terminales telefónicos. Además recoge variabilidad dialectal en cada una de las fases en que está grabada: Fase 1 (inglés americano de la mitad Atlántica), Fase 2 (inglés americano de la mitad oeste) y Fase 3 (inglés americano del sur). La Fase 3 fue empleada en las evaluaciones NIST 1996 a 1999, así como en las de 2002 y 2003 junto con la Fase 2; en la evaluación NIST 2000 se empleó la base completa.

- **Switchboard 3:** conocida como *Switchboard cellular*, contiene habla en inglés americano y recoge variabilidad dialectal. Está grabado sobre redes móviles. Fue grabada en dos fases, conteniendo cada una distintos tipos de canal: Fase 1, canal de transmisión GSM; y Fase 2, canal de transmisión CDMA. La Fase 1 fue empleada en la evaluación NIST 2001 y la Fase 2 fue empleada en las evalauciones NIST 2002 y 2003.

- **Mixer y datos adicionales multilenguaje:** recoge mayor variabilidad de canal y terminales que los Switchboard, también incluye habla grabada sobre teléfonos inalámbricos de líneas telefónicas convencionales y redes móviles. Por otra parte, contiene información multi-lenguaje: inglés americano, español, árabe, chino mandarín y ruso. Esta base de datos se empleó en la evaluación NIST 2004. En las evaluaciones NIST de 2005 y 2006 se añadieron nuevas grabaciones incluyendo los idiomas anteriores y siguiendo el mismo protocolo, incluyendo variaciones de dialecto locutores no nativos, esta extensión del Mixer se llama **Mixer 3**.

En resumen, para la realización de la parte experimental del proyecto se ha empleado el conjunto de habla conversacional en inglés correspondiente a las bases de datos de las evaluaciones NIST SRE. El protocolo usado corresponde al de la evaluación NIST SRE 2006. Para el entrenamiento de los UBMs se usan los datos procedentes de las evaluaciones SRE 2004 y 2005 (*corpus* MIXER). Esta base de datos consiste en 367 locutores de género masculino provenientes de 1808 conversaciones. Estos datos también se utilizaron en el entrenamiento de las matrices de *eigenvoices* y *eigenvectors* para la aplicación del *Factor Analysis*. Así mismo, las cohortes de normalización para Z-Norm y T-Norm también se obtuvieron de la base de datos MIXER.

Para el entrenamiento de los modelos de locutor y testeo del sistema se ha empleado la base de datos y protocolos de evaluación NIST SRE 2006 (*corpus MIXER3*) con la tarea *1 conversation 2-channel frente 1 conversation 2-channel* o también llamado *1conv4wire-1conv4wire*. En esta tarea se dispone grabaciones de conversaciones telefónicas (locutor aislado) de duración de 5 minutos y aproximadamente 2,5 minutos de habla neta por locutor. Adicionalmente, la base de datos y protocolos consta de locutores nativos y no nativos agrupados en 9720 trials, los cuales contiene 290 trials genuinos y 9430 trials de impostores. Todos los trials contienen locuciones provenientes de 298 locutores del género masculino provenientes de distintas líneas telefónicas.

El conjunto de datos de la evaluación SRE 2005 también ha sido empleado para la obtención de *puntuaciones* o *scores* para entrenar la calibración (calibración mediante regresión logística lineal) y obtener LR (Likelihood Ratios) calibrados.

Como medida del rendimiento del sistema se emplea la tasa de error igual EER (Equal Error Rate) y la función de detección de coste DCT (Detection Cost Function) definido en la evaluación NIST SRE 2006 [SRE 2006]. Los valores de Cllr, minCllr y $Cllr^{cal}$ (2.5.2.3) han sido empleados para medir la bondad del sistema después del proceso de calibración.

4. Descripción de resultados

4.1 Introducción

En este capítulo recoge los resultados obtenidos a lo largo de la evaluación del sistema. Se ha implementado distintas configuraciones del sistema, las cuales están en función del tipo de modelado de GMM-UBM (Gloabal GMM-UBM y Constrained GMM-UBM), del tipo de parametrización usada (contornos de coeficientes cepstrales, contornos de formantes, coeficientes cepstrales, coeficientes cepstrales compensados de sesión de variabilidad, etc) y el tipo de unidad lingüística (fonema, difonema y trifonema).

4.2 Sistema de referencia

Para la realización de las pruebas de los sistemas basados en GMM-UBM, se empleó como sistema de referencia un sistema GMM-UBM con 1024 mezclas de gaussianas, matrices de covarianza diagonal y coeficientes cepstrales. Estos coeficientes son del tipo 2 (Tabla 2), es decir, 19 coeficientes estáticos más sus respectivas derivadas de primer orden. Además, estos coeficientes han sido previamente normalizados por media cepstral, filtrado RASTA y feature warping. Este sistema obtiene un rendimiento en términos de **EER = 10,21 %** y **minDCF = 0,0457**.

4.3 Experimentos del sistema GMM-UBM Global

En este apartado, se presenta los resultados obtenidos mediante el modelado Global GMM-UBM (3.3.1) con la parametrización de contornos temporales de coeficientes cepstrales y formantes en unidades lingusiticas.

4.3.1 Tipos coeficientes cesptrales y formantes

Primeramente, se ha realizado experimentos para determinar el tipo de coeficientes cepstrales y formantes cuyas trayectorias temporales otorgen mejor rendimiento al sistema GMM-UBM Global. En la siguiente tabla se presenta los tipos de coeficientes y formantes usados en la parametrización de las características.

Coeficientes cepstrales	
Tipo 1	19 coefs. estáticos
Tipo 2	19 coefs. Estáticos más sus primeras derivadas
Tipo 3	Derivadas de los 19 coefs. estáticos
Formantes	
Tipo 1	3 primeras frecuencias: F1, F2 y F3
Tipo 2	3 primeras frecuencias: F1, F2 y F3 más sus respectivos anchos de banda : FBW1, FBW2 y FBW3
Tipo 3	3 anchos de banda: FBW1, FBW2 y FBW3

Tabla 2. Tipos de coeficientes-cepstrales/formantes.

Esta prueba se ha realizado sobre unidades de **fonemas**. Por otra parte, los UBMs y los modelos de locutor se han entrenado con 128, 64 y 32 mezclas para observar mejor el comportamiento de las trayectorias en la verificación de locutor.

De acuerdo con las curvas DET en la Figura 24 el mejor tipo de coeficientes-cepstrales/formantes que mejor rendimiento otorga al sistema en término de EER es del tipo1 para ambos casos.

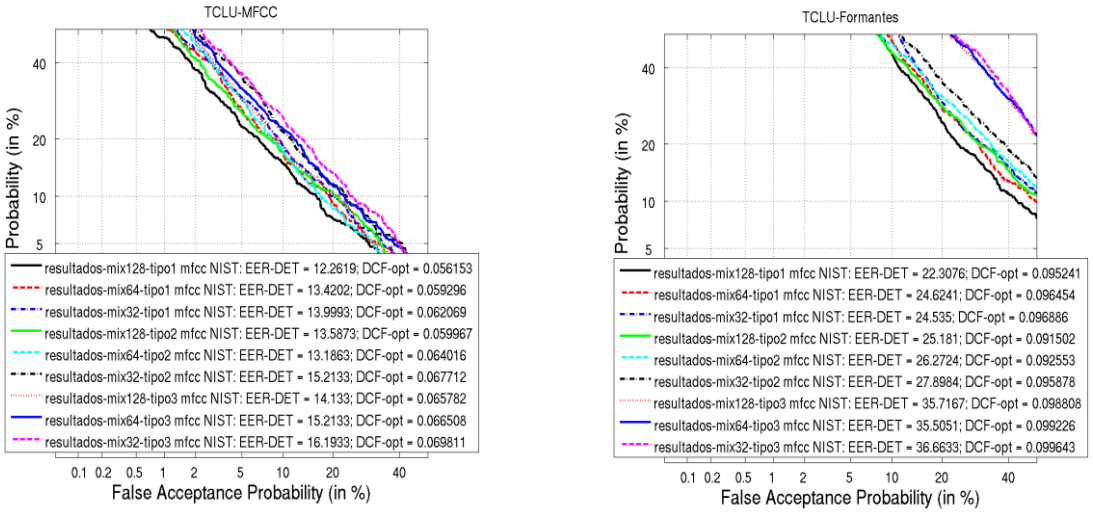

Figura 24: Curvas DET para parametrizaciones TCLU-MFCC y TCLU-Formantes sobre fonemas.

Para los contornos de coeficientes cepstrales del **tipo 1** se observa valores de FAR (Probabilidad de falsa aceptación) y FRR (probabilidad de falso rechazo) por debajo del resto. Una posible expliación a este comportamiento, en el caso de usar sólo la información dinámica (**tipo3**), es que el tipo de información no sea suciente para llegar a un nivel discriminación considerado comparado al uso de sólo información estática (**tipo1**). Por otra parte, en el caso de los coeficientes cepstrales, al usar información estática más información dinámica (**tipo2**) hace que la dimensión de los vectores de características sea igual a 190 (38 coef cepstrales x 5) y considerando que el UBM se entrena con 2219612 de vectores de características (NIST SRE 04 y 05), estaría produciendose lo que se conoce como “Maldición de la dimensionalidad”. Esto se debe a que los modelos tradicionales de estadísticas como los GMM no pueden manejar datos de alta dimensión. Los datos de entramiento necesarios para una correcta estimación de las funciones de densidad crecen exponencialmente con el número de dimensión de los vectores de características. Por tanto, se deduce que los vectores de características de dimensión igual a 190 requiere más datos para el correcto entrenamiento.En el caso de contornos de frecuencias de formantes, la diferencia es más notable entre el **tipo1** y el resto (**tipo 2** y **tipo 3**).

La Tabla 3 recoge los resultados en términos de EER de la Figura 24:

	TCLU-MFCC			TCLU-Formantes		
Número de gaussianas	**128**	**64**	**32**	**128**	**64**	**32**
Tipo 1	12,26	13,42	13,99	22,30	24,62	24,53
Tipo 2	13,58	13,18	15,21	25,18	26,27	27,89
Tipo 3	14,13	15,21	16,19	35,71	35,50	36,66

Tabla 3. Valores de EER en función del número de mezclas para TCLU-MFCC y TCLU-Formantes.

Estos resultados demuestran lo contario a lo que se creia al inicio. Al tener más información los vectores de características (información estatítica + información dinámica) se creia que el sistema GMM-UBM podría funcionar mejor, incluso, aplicando posteriormente la parametrización de los contornos temporales el vector final de caracterísitcas recogería mayor variabilidad de los coeficientes cepstrales o, en caso de formantes, frecuencias más su evolución temporal. Sin embargo, esta hipótesis sí se cumple para el uso de contornos de frecuencias y anchos de banda de formantes para el modelo MVK (Multivariate Kernel Densities) [González-Rodríguez, 2011]. Por tanto, la selección del tipo de parametrización estará relacionada al tipo de modelado que se usará en el sistema, siendo en este proyecto GMM-UBM.

Con los resultados obtenidos, los coeficientes cepstrales y los formantes, ambos del **tipo 1,** se usarán para la parametrización de contornos temporales en el sistema GMM-UBM global.

4.3.2 Influencia de compensación de variabilidad en vectores de características

En este proyecto, se ha aplicado compensación de sesión de vararibilidad a los vectores de características. De esta forma, se pretende estudiar la influencia de la compensación de varabilidad de sesión en el sistema.

4.3.2.1 Compensación de variabilidad en el dominio de características

La compensación de variabilidad de sesión se implementa mediante Joint Factor Analysis, JFA (2.6.4.1 Joint Factor Analysis). Esta técnica asume que existe una variabilidad no deseada dentro de un subespacio de baja dimensionalidad que modifica el supervector de locutor *s* para una locución *h*.

$$\mu_{sh} = \mu_s + Ux_h \tag{4.1}$$

donde el termino **U** representa el subespacio de variabilidad de sesión y $\mathbf{x_h}$ son los factores de canal o *channels factors*, ambos independientes del locutor **s**, por tanto la compensación de variabilidad de sesión en el dominio de características [González-Rodríguez et al., 2012] se basa en la substracción del componente aditivo de sesión en cada observación o dimensión del vector de características $\mathbf{o_h(t)}$:

$$\hat{o}_h(t) = o_h(t) - \sum_c y_c(t)\, U_c x_h \qquad (4.2)$$

donde $\hat{o}_h(t)$ es la observación compensada, $y_c(t)$ es la probabilidad de ocupacion de la gaussiana de trama *t* respecto a la componente **c** de la gaussiana del UBM y U_c es la submatriz del subespacio de variabilidad de sesion correspondiente a la Gaussiana c.

4.3.2.2 Resultados de compensación de variabilidad de sesión

La compensación de variabilidad se ha realizado sobre los coeficientes cepstrales de diferentes tipos (Tabla 2) de unidades de fonemas. Una vez obtenidos estos coeficientes compensados, se procede a la parametrización de los contornos temporales. Las matrices U de subespacio de variabiliadad de sesion han sido entrenadas mediante Principal Component Analysis (PCA) más 5 iteraciones EM y con los datos de Switchboards I&II, NIST SRE05 y SRE06. Para el entrenamiento del modelo UBM y modelos de locutor se ha empleado 64 mezclas de gaussianas.

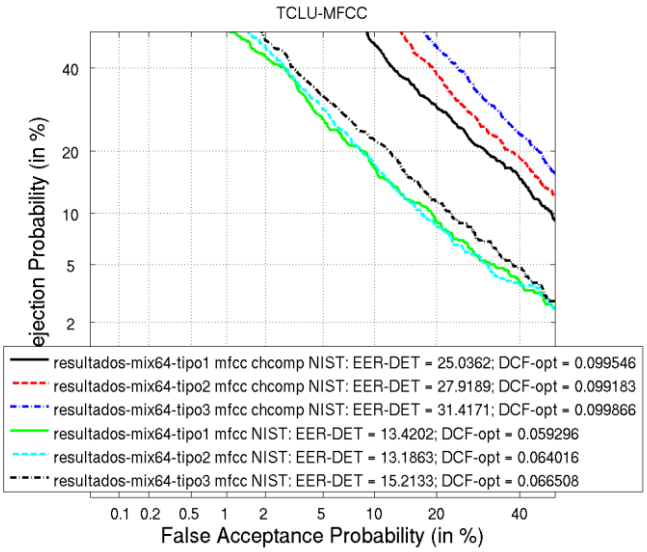

Figura 25. Coeficientes cepstrales compensados Vs. coeficientes cepstrales no compensados

En la Figura 25 se observa que los coeficientes cepstrales sin compensar funcionan mucho mejor en terminos de EER. Por tanto, la compensación de variabilidad de sesión no tiene efecto sobre los contornos temporales de vectores de características produciendo una degradación en éstos.

Incluso, los contornos temporales de coeficientes cepstrales compensados (MFCC compensados) en comparación con los contornos de las frecuencia de formantes (F123) presentan un peor rendimiento en términos de EER.

	EER (%), UBM 64 mezclas		
	F123	**TCLU-MFCC**	**TCLU-MFCC compensados**
Tipo 1	24,62	**13,42**	25,03
Tipo 2	26,27	**13,18**	27,91
Tipo 3	35,50	**15,21**	31,41

Tabla 4. Comparación de valores de EER en función del tipo de contornos de características: Frecuencias de formantes, MFCC y MFCC con compensación de variabilidad.

Con el fin de analizar mucho mejor la influencia de la compensación de variabilidad de sesión en el domino de características, se ha realizado pruebas sobre el sistema de referencia, pero con otras mezclas exactamente desde 8 hasta 1024 mezclas, y usando coeficientes cepstrales, tanto compensados como sin compensar. Cabe señalar, en base a trabajos anteriores, que cuando se trabaja sólo con coeficientes cepstrales, el mejor rendimiento del sistema se obtiene con el tipo 2 (19 MFCC + 19 deltas).

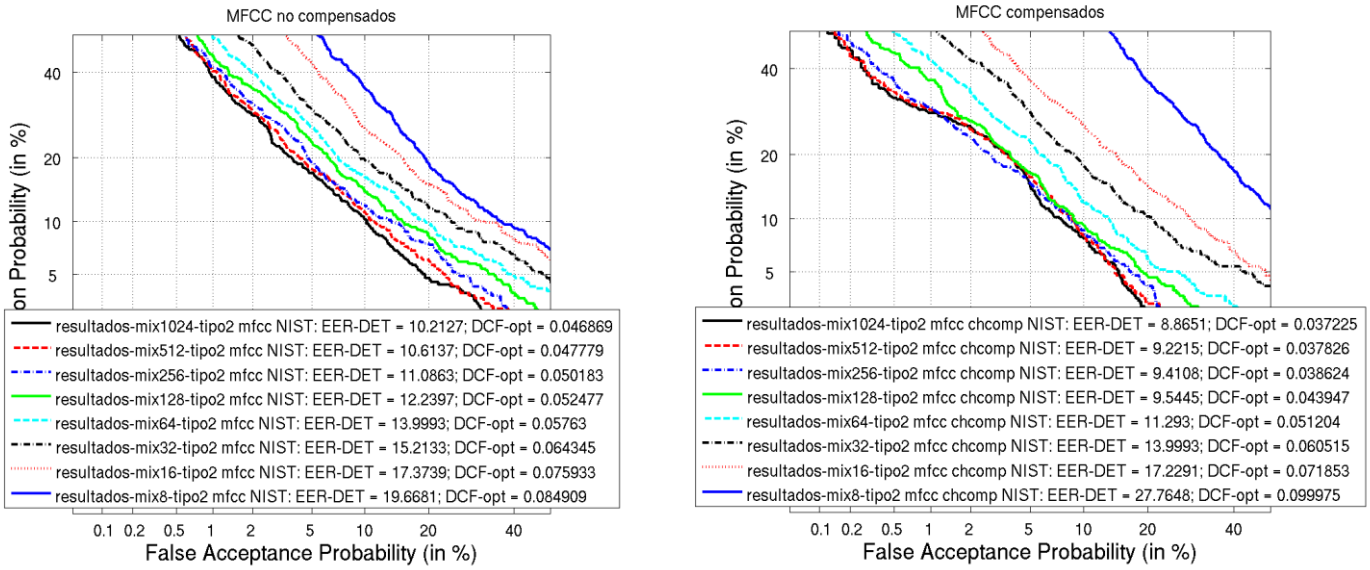

Figura 26. Coeficientes cepstrales compensados vs coeficientes cepstrales sin compensar.

En la Figura 26 se observa que usando los coeficientes cepstrales compensados el sistema obtiene un mejor rendimiento, por ejemplo, para 1024 mezclas de gaussianas se obtiene un valor de EER = 8,86 %, mientras que usando coeficientes no compensados se obtiene un valor de EER = 10,21 %. Por tanto, la compensación de variabilidad sí surge efecto sobre los coeficientes cepstrales, demostrando un mayor rendimiento en términos de EER.

De esta forma se confirma que la parametrización de los contornos temporales de estos coeficientes genera distorsión en ellos previa compensación de variabilidad de sesión en los coeficientes cepstrales.

Con los resultados obtenidos, los coeficientes cepstrales, correspondientes al **tipo 1** y sin compensar, se usarán para la parametrización de contornos temporales en el sistema GMM-UBM global.

4.3.3 Influencia en el número de mezclas de gaussianas en el sistema GMM-UBM Global

A continuación, se presenta los resultados de las pruebas en el sistema GMM-UBM Global con diferentes número de mezclas de gaussianas, desde 2 hasta 1024 con incrementos de múltiplos de 2; y con contornos temporales de coeficientes cepstrales por fonema, difonema y trifonema.

Se observa en la Figura 27, para el caso de unidades de fonemas y difonema, a medida que incrementamos el número de gaussianas el rendimiento del sistema mejora en términos del EER. Este resultado nos da a entender que si el modelo UBM se entrena con máyor número de mezclas gaussianas, los modelos de locutor, adaptados del UBM, representarán mucho mejor las características del propio locutor, obteniéndose así una mejor puntuación o *score* de verificación en el sistema, y por tanto, un mayor valor de EER.

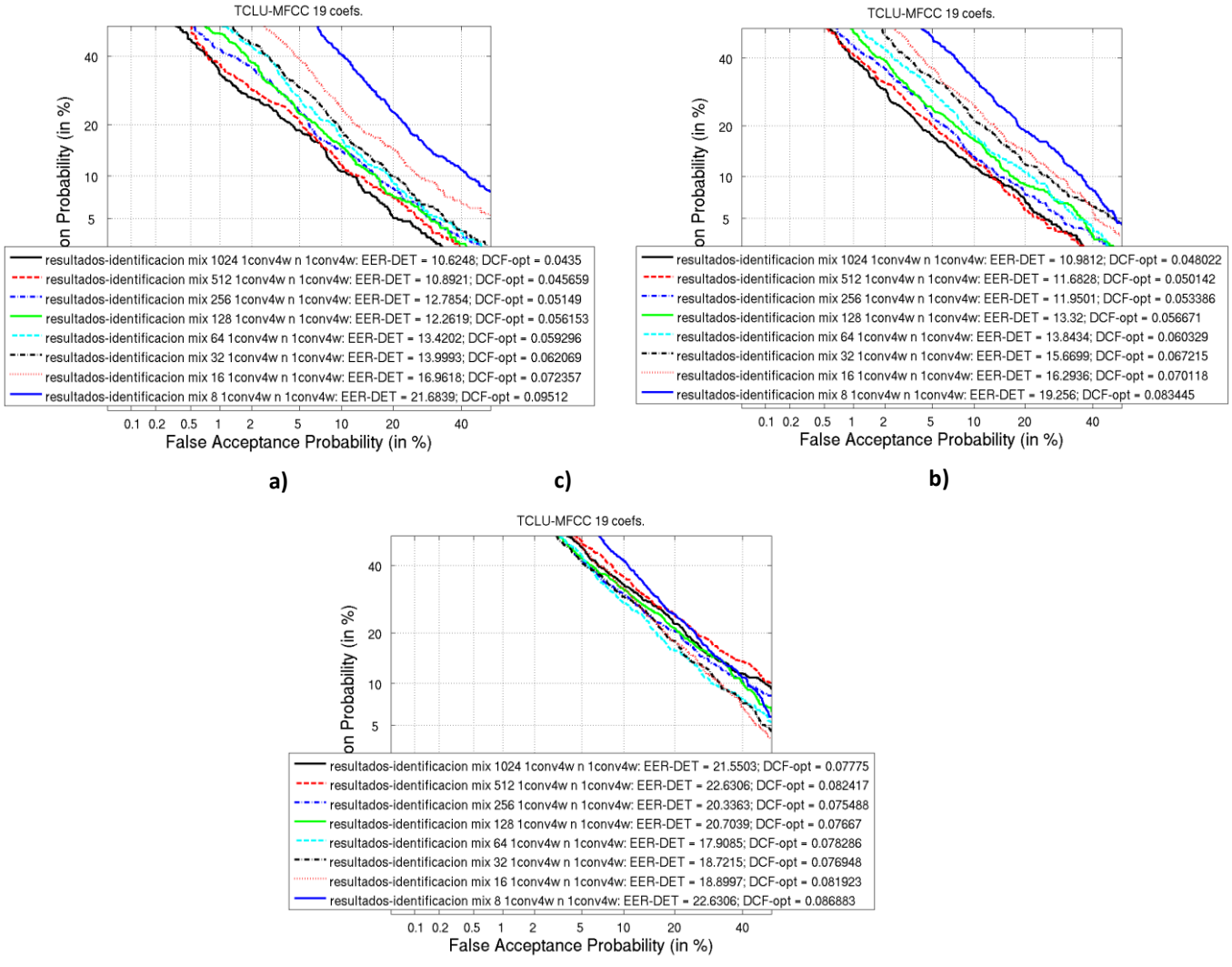

Figura 27. Curvas DET para distintos número de mezclas y unidades: **a)** fonemas, **b)** difonemas y **c)** trifonemas.

Sin embargo, el incrementar el número de mezclas para entrenar el UBM no siempre se traduce en un mejor rendimiento en el sistema, dado que existe un punto óptimo o número de mezclas óptimo que permite obtener un mejor rendimiento. Pasado ese punto, el rendimiento empieza a disminuir, es decir, los valores de EER, FAR (Probabilidad de Falsa Aceptación) y FRR (Probabilidad de Falso Rechazo) empiezan a aumentar. Este comportamiento se observa en el caso de los trifonemas; el rendimiento mejora a medida que incrementamos el número de gaussianas hasta llegar al número óptimo de mezclas igual a 64 con un valor de EER = 17,90 %, pasado ese punto el rendimiento comienza a degradarse. Para verificar si presenta el mismo comportamiento en el caso de los fonemas, se ha entrenado un UBM con 2048 mezclas.

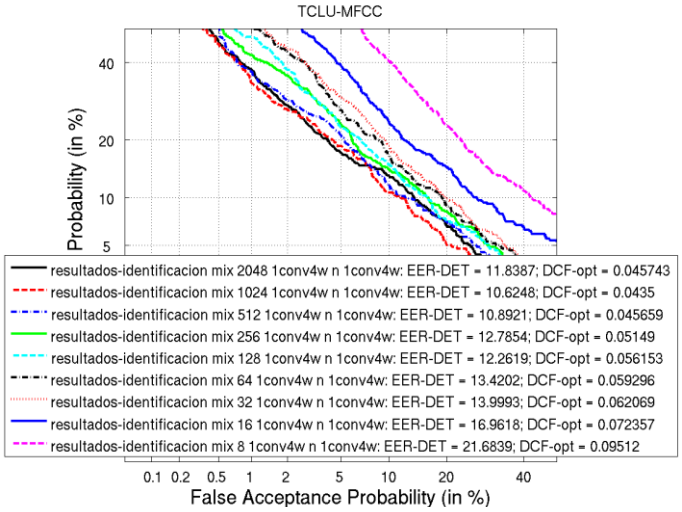

Figura 28. Curvas DET para UBM con 2 mezclas hasta 2048 mezclas para fonemas.

En la Figura 28 observamos el mismo comportamiento descrito anteriormente. Por tanto, el nivel óptimo de mezclas es igual a 1024 mezclas de gaussianas, pasado ese punto el rendimiento empieza a degradarse.

Otro factor que se muestra en la Figura 27, es que el mejor rendimiento del sistema se obtiene usando unidades de fonemas. Una de las posibles razones es que las unidades de difonemas y trifonemas cubren intervalos de tiempos mayores a los de los fonemas, por tanto, sus trayectorias temporales son mucho más complejas. Por tanto, existe una mejor discriminación entre locuciones segmentadas en estas unidades. Cabe recordar que para el entrenamiento del UBM Global los datos de entrenamiento son locuciones segmentadas en unidades.

	EER (%)		
Mezclas	**Fonemas**	**Difonemas**	**Trifonemas**
1024	**10,62**	**10,98**	**21,55**
512	10,89	11,68	22,63
256	12,78	11,95	20,33

Tabla 5. Valores de EER para UBM con distinto número de mezclas.

Por otra parte, se desea obtener valores de LR a partir de las puntuaciones y así tener la probabilidad de una hipótesis de ser verdadera frente a la hipótesis opuesta (2.5.2.1).

Nº Gaussianas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
1024	**10,62**	**0,0435**	0,4397	**0,3573**
512	10,89	0,0457	1,3090	0,3821
256	12,78	0,0515	0,8753	0,4168
128	12,26	0,0562	0,4811	0,4176
64	13,42	0,0593	0,5425	0,4506
32	13,99	0,0621	2,1120	0,4682
16	16,96	0,0724	0,6083	0,5464
8	21,68	0,0951	0,7212	0,6817

Tabla 6. Valores de EER, Cllr y minCllr para locuciones segmentadas en fonemas.

En la Tabla 6, se observa que para la mayoría de las mezclas la pérdida de calibración es menor ($Cllr^{cal}$ = C_{llr} - $minC_{llr}$), por tanto, esto nos permite obtener valores de LR bien calibrados. Por otra parte, el mejor rendimiento, en términos de EER, se obtiene con 1024 mezclas, con un valor de EER = 10,62 %, además presenta el menor coste de C_{llr} y el menor valor de $minC_{llr}$, esto se traduce en un buen nivel de discriminación respecto al resto de mezclas.

Usando los resultados de las pruebas con el sistema de referencia con distintas mezclas, se genera la siguiente tabla comparativa entre los coeficientes cepstrales y sus contornos temporales, ambos sin compensación de variabilidad de sesión.

	MFCC		TCLU-MFCC	
Nº Gaussianas	**EER (%)**	**minDCF**	**EER (%)**	**minDCF**
1024	**10,21**	**0,0468**	**10,62**	**0,0435**
512	10,61	0,0477	10,89	0,0457
256	11,08	0,0501	12,78	0,0515
128	12,23	0,0524	12,26	0,0562
64	13,99	0,0576	13,42	0,0593
32	15,21	0,0643	13,99	0,0621
16	17,37	0,0759	16,96	0,0724
8	19,66	0,0849	21,68	0,0951

Tabla 7. Comparación de valores de EER y minDCF en función de MFCC y TCLU-MFCC.

De acuerdo con la Tabla 7, los contornos temporales de los coeficientes cepstrales (TCLU-MFCC) obtiene un rendimiento similar al de los coeficientes cepstrales (MFCC), por tanto, el uso de contornos temporales presenta una buena perspectiva en el uso de este tipo de parametrización en pruebas futuras, permitiéndonos acercarnos a un enfoque forense; uno de los motivos principales de este proyecto.

4.3.4 Normalización de puntuaciones

En este apartado se presentará los resultados de aplicar normalización a las puntuaciones obtenidas del sistema. Tal como se explicó en el apartado 2.5.3, existe un desalineamiento de las puntuaciones debido a que se usa un umbral común a todos los usuarios. Por tanto, mediante la normalización se pretende que la distribución de las puntuaciones siga una media igual a cero y varianza unidad. Siguiendo con los resultados anteriores, la normalización se aplicó a las puntuaciones del sistema GMM-UBM Global para 512 mezclas. Se ha aplicado tres tipos de normalización: Z-NORM, T-NORM y ZT-NORM.

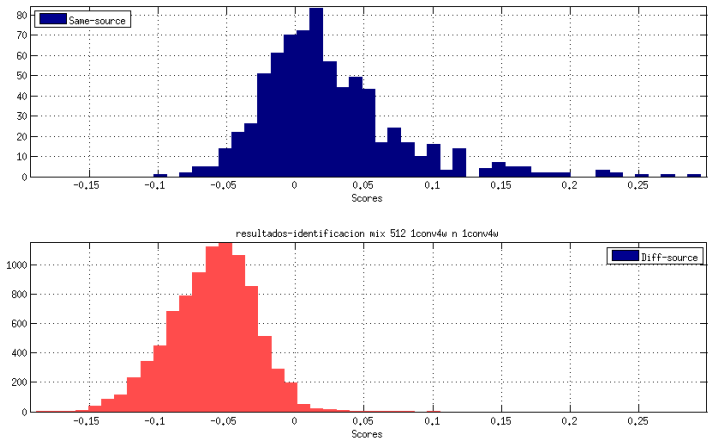

Figura 29. Histograma de puntuaciones para UBM con 512 mezclas.

En la Figura 29 se muestra el histograma de puntuaciones sin normalizar del sistema GMM-UBM Global con 512 mezclas y con un valor de EER = 10,89 %. En la figura se observa que tanto la distribución de puntuaciones de la hipótesis “misma fuente” (en la figura same-source) e hipótesis “distinta fuente” (en la figura diff-source) no están relativamente separadas, es decir, una parte de las puntuaciones “distinta fuente” puntúan alto, dando más apoyo a la hipótesis “misma fuente”, de la misma forma, parte de las puntuaciones “misma fuente” llegan a valores bajos, dando apoyo a las hipótesis contraria. Por ello, se pretende, mediante la normalización, aumentar la distancia entre las puntuaciones y así obtener mejor discriminación entre usuarios. Se debe tener en cuenta que se está hablando de puntuaciones, por tanto, no se sabe el grado de apoyo a una hipótesis

A la hora de aplicar la normalización T-Norm la selección de cohortes de modelos es un elemento importante. Estos modelos han de ser lo más parecidos posible a los modelos de usuario, y su número ha de ser elevado, dado que se debe estimar una gaussiana a partir de las puntuaciones obtenidas.

Para la parte experimental se ha empleado los modelos de la evaluación NIST SRE 2004. De la misma forma, para Z-Norm se ha empleado los datos de test de la evaluación NIST SRE 2004.

Normalización	EER (%)	minDCF	Valor medio EER (%) por modelo	Valor medio EER (%) por test
Raw	10,89	0,0456	4,46	4,26
TNorm	10,34	0,0446	**3,86**	4,26
ZNorm	8,74	0,0423	4,46	**2,42**
ZTNorm	**8,50**	**0,0372**	4,17	**2,42**

Tabla 8. Comparación de puntuaciones normalizadas por tipos de normalización

En la Figura 30 se observa que el rendimiento del sistema mejora luego de aplicar normalización a las puntuaciones, obteniéndose el mejor rendimiento con la normalización del tipo ZTnorm con un valor de EER = 8.5087 %. Por tanto, se ha logrado bajar de un EER de 10,8921 % a 8,5087 %, de la misma forma, este rendimiento supera el obtenido por el sistema de referencia.

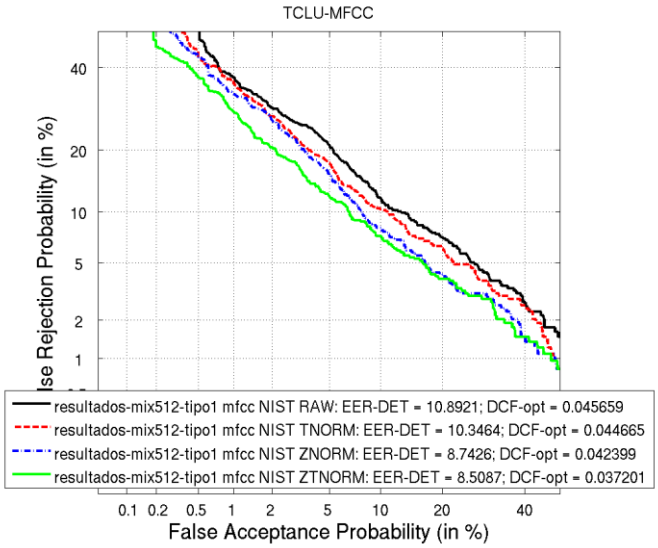

Figura 30. Representación de curvas DET con distinto tipos de normalización.

Por otra parte, esta mejora también se observa en los histograma de las puntuaciones normalizadas (Figura 31), en donde las nuevas puntuaciones de la "misma fuente" (en la figura same-source) puntúan mucho más alto, haciendo que las puntuaciones den mayor apoyo a la hipótesis "misma fuente" (en la figura same-source).

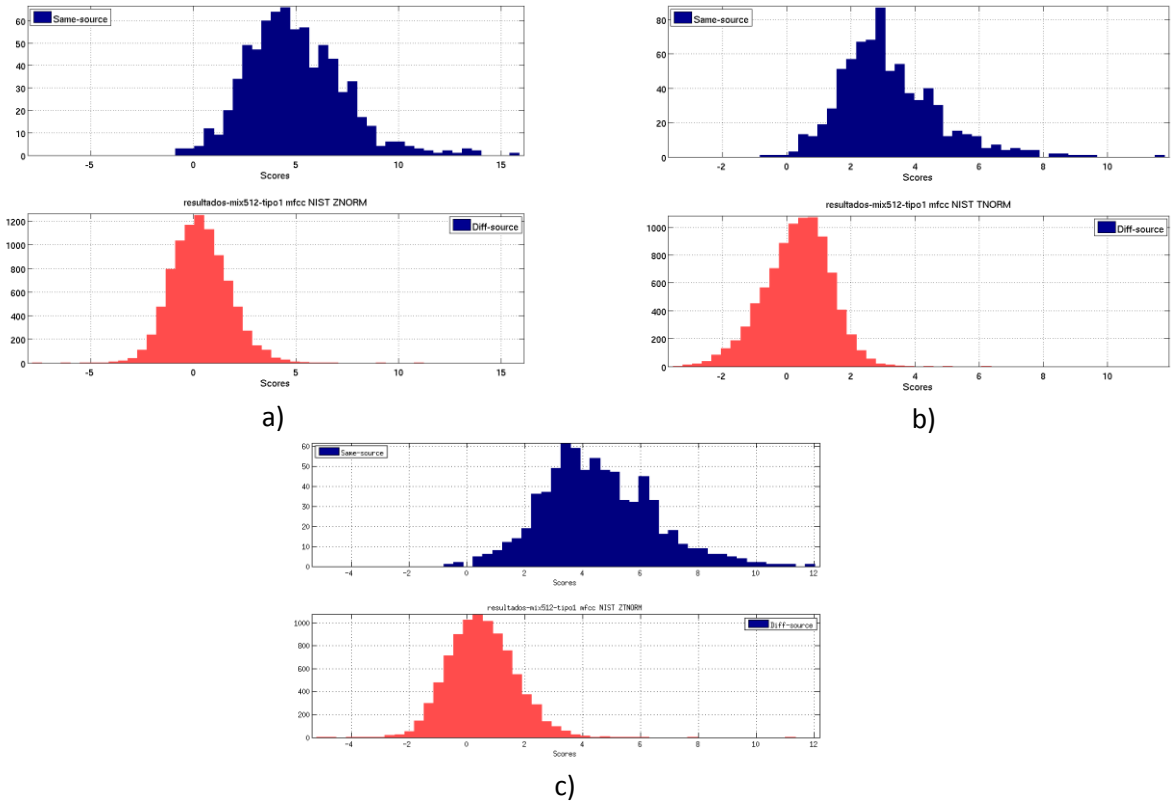

Figura 31. Histograma de puntuaciones normalizadas: a) ZNorm, b) TNorm y c) ZTNorm

4.3.5 Fusión inter-unidad

Luego de aplicar normalización a las puntuaciones, se procedió a la fusión de sistemas con puntuaciones sin normalizar. Para ello, mediante fusión suma promedio y fusión por regresión logística lineal se pretende combinar los sistemas GMM-UBM Global segmentados por unidades con mejor rendimiento: GMM-UBM Global 1024 mezclas-fonemas, GMM-UBM Global 1024 mezclas - difonemas y GMM-UBM Global 64 mezclas - trifonemas.

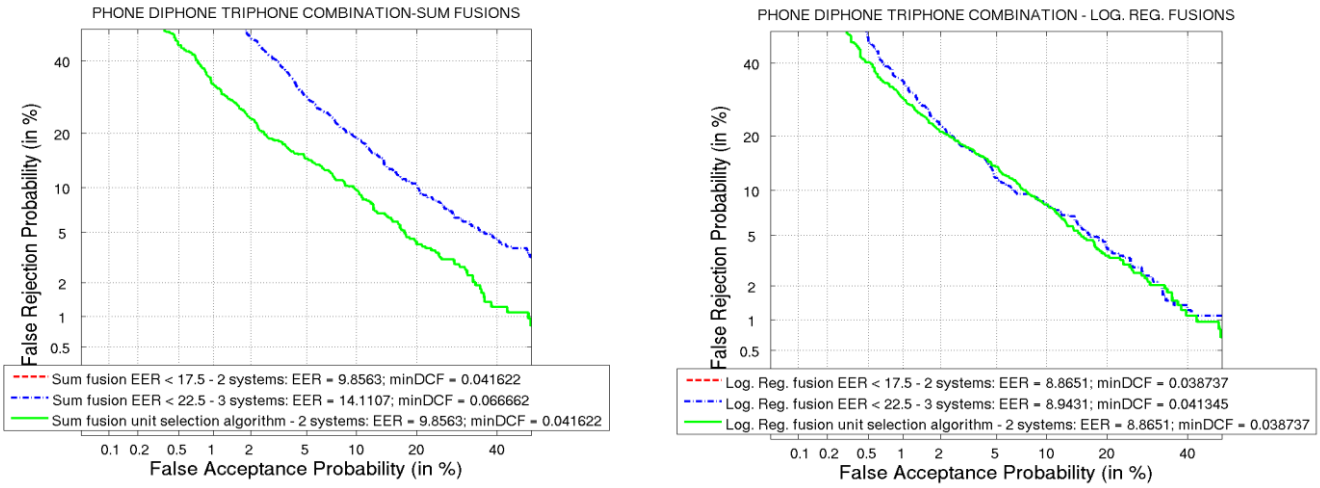

Figura 32. Curvas DETS con MFCC y MFCC compensados.

Se observa en la Figura 32 que el rendimiento mejora al aplicar los dos tipos de fusión, obteniéndose el mejor resultado, en caso de fusión con regresión logística, un EER = 8,8651 %, y, en caso de fusión suma promedio, un EER = 9,8563 %. Comparando con el sistema de referencia, se logra superar el rendimiento de éste en casi un 3%. Respecto a las pérdidas de calibración, éstas son menores para la fusión por regresión logística, el cual nos permite obtener mejores valores de LR calibrados, tal como se muestra en la Tabla 9.

Tipo de fusión	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Fusión suma promedio	9,8563	0,0416	0,64	0,48
Fusión por regresión logística lineal	**8,8651**	**0,0387**	**0,42**	**0,31**

Tabla 9. Valores de EER, minDCF, Cllr y minCllr por técnica de fusión.

Con la fusión por regresión logística lineal se obtiene mejor rendimiento en términos de EER, minDCF, incluso el coste por pérdida de discriminación ($minC_{llr}$)es menor, tal como muestra la curva Tippet en la Figura 33, donde se observa que existe una menor tasa de error de valores de LR que apoyan las hipótesis incorrectas, aproximadamente un 3,37 % de valores de LR relacionados a la hipótesis "misma fuente" (Same-source en la figura) que apoyan la hipótesis contraria y un 22,83 % para el caso de "distinta fuente" (Diff-source en la figura).

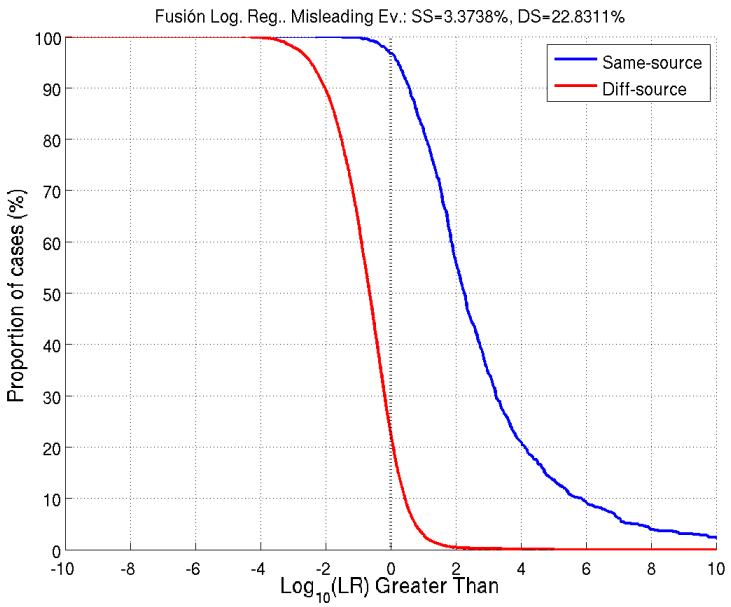

Figura 33. Curva tippet del sistema fusionado mediante regresión logística lineal.

Comparando los resultados de fusión por regresión logística con los de normalización de puntaciones mediante ZTnorm (Tabla 8), con estos últimos se obtiene mejor rendimiento, tanto en valor de EER como en coste mínimo de detección (minDCF).

4.4 Experimentos del sistema GMM-UBM dependiente de unidad

En este apartado, se presenta los resultados obtenidos mediante el modelado Constrained GMM-UBM, usando la parametrización de contornos temporales en unidades linguisticas de coeficientes cepstrales y formantes.

4.4.1 Sistema dependiente de unidad

Primeramente, se ha probado tres configuraciones de GMM-UBM dependiente de unidad, usando contornos temporales de coeficientes cepstrales: i**)** modelo UBM y modelo de locutor entrenados con datos independientes de unidad; datos test dependientes de unidad; **ii)** modelo UBM entrenado con datos independientes de unidad; modelos de locutor y datos test dependientes de unidad; **iii)** modelo UBM, modelo de locutor y datos test dependiente de unidad. Cada configuración ha sido probada con diferentes números de mezclas, desde 2 hasta 1024 con incrementos en potencia de 2. La Tabla 10 recoge la configuración óptima en función del número de mezclas.

	Tipo de configuración	Número de mezclas
Fonema	(**iii**)	8
Difonema	(**iii**)	4
Trifonema	(**iii**)	4

Tabla 10. Mejor configuración del sistema GMM-UBM dependiente de unidad, TCLU-MFCC.

Por tanto, las pruebas que se presentan en los apartados posteriores se basarán en la configuración del tipo (iii) y, dependiendo del tipo de unidad, se corresponderá con el número de mezclas óptimo (Tabla 10).

4.4.2 Tipos de coeficientes cepstrales y formantes

Al igual que en el sistema Global GMM-UBM, se ha realizado pruebas para determinar el tipo de coeficientes y tipo de formante (Tabla 2), cuyas trayectorias temporales otorgue mejor rendimiento al sistema Constrained GMM-UBM. Para esta prueba se ha empleado 128 mezclas gaussianas para entrenamiento de los modelos de locutor y UBM. La siguiente tabla se muestra los resultados de estas pruebas hechas a nivel de fonema.

TCLU-MFCC				
Tipos de coeficientes cepstrales	**Nº fonemas con EER (%) < 30 %**	**media EER (%)**	**min EER (%)**	**max EER (%)**
19 coef. estáticos (Tipo1)	**19**	**24,95**	**15,92**	**29,39**
19 coef. estáticos + 19 coef. dinámicos (Tipo2)	20	25,65	17,9	29,16
19 coef. dinámicos (Tipo3)	17	26,82	19,93	29,8
TCLU-Formantes				
Tipos de formantes	**Nº fonemas con EER (%) < 30 %**	**media EER (%)**	**min EER (%)**	**max EER (%)**
F1, F2, F3 (Tipo1)	**13**	**27,20**	**24,92**	**29,11**
F1, F2, F3 + FBW1, FBW2, FBW3 (Tipo2)	4	26,92	24,46	29,69
FBW1, FBW2, FBW3 (Tipo3)	0	0	0	0

Tabla 11. Rendimiento del sistema en función del tipo de coeficientes-cepstrales/formantes.

De acuerdo con la Tabla 11, el mejor rendimiento se obtiene con los coeficientes/formantes del tipo 1 (19 coef. estáticos para coeficientes cepstrales y las primeras tres frecuencias para formantes). Este rendimiento se relaciona con: el número de fonemas con valores de EER por debajo del 30 %, el valor medio de EER, valor mínimo de EER y valor máximo de EER. Este tipo de coeficiente/formante será empleado en las pruebas posteriores dependiente de unidad.

4.4.3 Influencia de compensación de variabilidad en vectores de características

Siguiendo el procedimiento para la compensación de variabilidad de sesión del apartado 4.3.2.1, se presentará los resultados de aplicar compensación sobre el dominio de características. La compensación de variabilidad se ha realizado sobre los coeficientes cepstrales del tipo 1 (Tabla 11) de unidades de fonemas, ya que son los que otorgan mejor rendimiento al sistema. Para esta prueba se ha empleado 8 mezclas de gaussianas para el entrenamiento de los modelos de locutor y el modelo UBM.

Fonema	TCLU-MFCC EER (%)	TCLU-MFCC compensados EER (%)
AE	18,98	48,41
AH	29,39	49,96
AX	27,08	**46,67**
AY	21,68	46,76
DH	28,57	49,32
EY	26,35	49,08
IH	26,98	48,02
IY	23,33	49,28
K	27,71	48,33
L	26,44	52,39
M	22,26	48,82
NG	29,39	50,43
N	**15,92**	47,18
OW	24,63	49,15
PUH	24,15	48,94
R	24,65	50,81
T	27,89	47,86
UW	24,69	49,62
Y	24,01	49,32

Tabla 12. Valores de EER por fonema debajo del 30 % usando coeficientes cepstrales sin compensar y valores de EER para los mismos fonemas usando coeficientes cepstrales compensados.

De acuerdo con los resultados de la Tabla 12, usando los coeficientes cepstrales sin compensación de variabilidad se obtiene un rendimiento muy superior, mientras que los coeficientes compensados obtienen valores de EER por encima del 30 %. Esto confirma los resultados del sistema Global GMM-UBM (4.3.2.2) donde la influencia de la compensación de variabilidad produce distorsión en la parametrización de los contornos temporales. Para las pruebas posteriores se usará coeficientes cepstrales sin compensación de variabilidad.

4.4.4 Pruebas a nivel de fonema

Estas pruebas corresponden a la evaluación individual por fonema correspondiente a la evaluación NIST SRE 2006 con la tarea **1conv4wire-1conv4wire** en idioma inglés (3.4.1.1).

4.4.4.1 TCLU-MFCC

Para estas pruebas se han empleado los contornos temporales de los coeficientes cepstrales sin compensación de variabilidad y del tipo 1 (19 coeficientes estáticos). De acuerdo con la Tabla 10, se ha empleado la configuración dependiente total de unidad: modelos de locutor, modelo UBM y datos de test dependientes de fonema.

Tal como muestra la Tabla 13, el número de mezclas óptimo para las unidades de fonemas es igual a 8 mezclas. Para llegar a este resultado, se ha probado el sistema desde 2 hasta 1024 mezclas de gaussianas con un incremento igual a potencia de 2. La siguiente tabla recoge estos resultados:

Nº mezclas	Nº unidades EER < 30 %	Media EER (%)	Min EER (%)	Max EER (%)
1024	3	25,38	22,49	27,26
512	3	25,41	22,9	28,87
256	6	25,59	18,42	29,85
128	12	26,42	17,93	29,42
64	16	26,19	17,58	29,98
32	17	25,44	16,14	29,73
16	19	25,08	16,33	29,25
8	**19**	**24,95**	**15,92**	**29,39**
4	19	25,16	17,29	28,97
2	18	26,19	18,35	29,78

Tabla 13. Valores medios, mínimos y máximos de EER (%) del rendimiento del sistema TCLU-MFCC por fonema, con valores de EER < 30 %.

De acuerdo con los resultados de la Tabla 13 el número de mezclas óptimo es igual a 8 mezclas gaussianas debido a que presenta más unidades de fonemas cuyo rendimiento está por debajo del 30 %, y además, presenta el valor más pequeño respecto a la media de EER. Cabe recordar que esto valores de EER (%) han sido obtenidos en función de las puntuaciones por trial. Una vez obtenido el número de mezclas óptimo para el sistema se procede a la evaluación por fonema.

En la Tabla 14 se presentan el rendimiento por fonema en base a valores de LR, obtenidos mediante calibración logística lineal de las puntuaciones.

Este grupo de fonemas corresponde a aquellos que presentan un valor de EER por debajo de 30 %, los cuales presentan valores de EER y minDCF altos, sin embargo, la mayoría presenta poca pérdida de calibración (C_{llr}^{cal} = C_{llr} - $minC_{llr}$). Esto nos permite valores de LR calibrados por locución segmentada en fonema. En la Figura 34 se representa la curva tippet del fonema "N", cuyo rendimiento en el sistema es el más alto.

Fonema	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AE	18,99	0,0813	0,6087	0,5832
AH	29,39	0,0969	0,8235	0,7967
AX	27,09	0,0947	0,7882	0,7512
AY	21,68	0,0869	0,6822	0,6428
DH	28,43	0,0934	0,8132	0,7867
EY	26,41	0,0925	0,7713	0,7515
IH	26,95	0,0948	0,7964	0,7495
IY	23,32	0,0923	0,7453	0,7002
K	27,76	0,0961	0,8219	0,7832
L	26,52	0,0935	0,7789	0,7451
M	22,29	0,0857	0,6824	0,6583
NG	29,38	0,0934	0,8178	0,7959
N	**15,93**	**0,0713**	**0,5415**	**0,5082**
OW	24,66	0,0987	0,7917	0,7396
PUH	24,19	0,0908	0,7359	0,7149
R	24,66	0,0887	0,7295	0,7116
T	27,90	0,0921	0,8005	0,7647
UW	24,79	0,0898	0,7391	0,7198
Y	24,00	0,0906	0,7313	0,7062

Tabla 14. EER (%), minDCF, C_{llr}, $minC_{llr}$ por fonemas con valores de EER < 30 % en TCLU-MFCC.

Estos valores de EER por fonema tienen un rendimiento peor en comparación con el sistema de referencia. No obstante, mediante las técnicas de fusión a nivel intra-unidad, se logra superar el rendimiento del sistema de referencia.

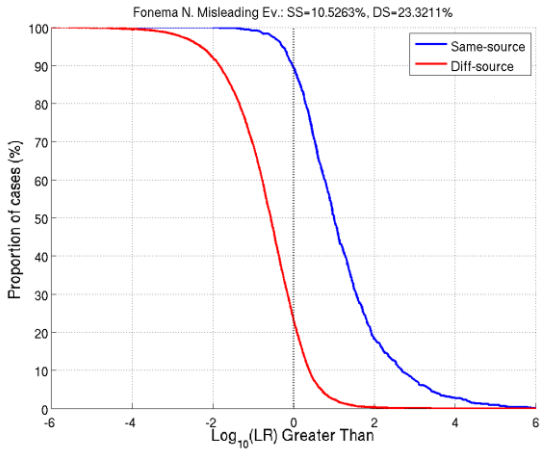

Figura 34. Curva tippet para el fonema con mejor rendimiento "N".

En la Figura 35 se presenta la fusión del sistema a nivel de unidades. En ella se observa que usando el algoritmo de selección de unidad se obtienen mejores resultados en ambas técnicas de fusión. El mejor rendimiento se alcanza con la fusión suma promedio, el cual presenta mayor número de unidades fusionada, con un valor de EER = 7,1166 % y minDCF = 0,042. Este rendimiento es un 30 % superior al sistema de referencia en términos de EER. Por otra parte, la menor pérdida de calibración se presenta en la fusión regresión logística (Tabla 15).

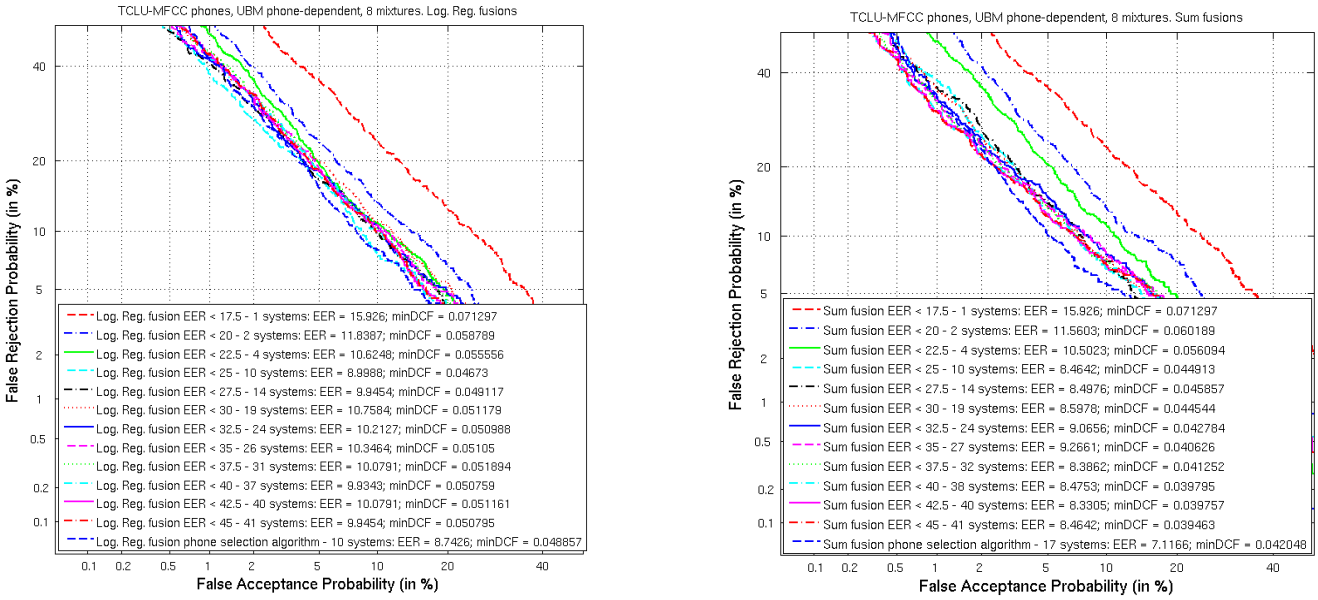

Figura 35. Curvas DET de los sistemas fusionados mediante regresión logística lineal y suma con TCLU-MFCC por fonema.

Técnica de Fusión (selección de unidad)	Nº de unidades fusionadas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Suma promedio	**17**	**7,11**	**0,0420**	0,61	0,27
Regresión logística lineal	10	8,74	0,0488	0,38	0,32

Tabla 15. Valores de EER (%), minDCF, C_{llr} y $minC_{llr}$ para cada técnica de fusión, TCLU-MFCC por fonema.

En la Figura 36 se muestra la curva tippet para la fusión suma promedio con técnica de selección de unidad. En ella se observa la buena discriminación presente en el sistema gracias al menor valor de $minC_{llr}$. Esta buena discriminación se hace visible en los valores de LR no muy altos que apoyan la hipótesis incorrecta “same-source” y valores de LR no tan pequeños que apoyan la hipótesis contraria “diff-source”. Alrededor del 4,183% de los valores de LR correspondientes a la hipótesis “misma fuente” (Same source) apoyan a la hipótesis contraria, mientras que en el caso contrario, el 14,82 % de los valores de LR de la hipótesis “diferente fuente (Diff-source) apoya la hipótesis contraria, por tanto, la discriminación del sistema por unidad mejora al fusionar las unidades.

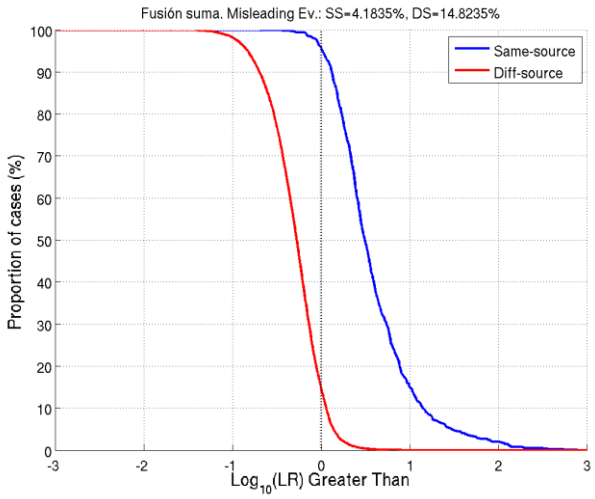

Figura 36. Curva tippet de la fusión suma con TCLU-MFCC por fonema y 8 mezclas.

Adicionalmente, con el fin de comparar el rendimiento con el uso de sólo coeficientes cepstrales, se ha realizado pruebas para distintos número de mezclas en el sistema a través de las puntuaciones por trial. Cabe recordar que el mejor rendimiento del sistema se obtiene empleando coeficientes cepstrales del tipo 2 (Tabla 2), es decir, 19 coeficientes estáticos más sus respectivas derivadas temporales. En la siguiente tabla se presenta el número de fonemas cuyo rendimiento individual está por debajo del 30 %, tanto para coeficientes cepstrales y sus contornos temporales.

MFCC				
Nº de mezclas	**Nº unidades EER < 30 %**	**Media EER (%)**	**Min EER (%)**	**Max EER (%)**
1024	30	21,29	12,66	29,35
512	31	21,06	12,56	29,91
256	**31**	**20,63**	**12,91**	**29,83**
128	30	20,48	12,64	28,29
64	30	21,12	13,85	28,23
32	29	21,37	14,33	29,18
16	29	22,21	14,8	29,91
8	31	23,61	15,9	29,91
TCLU-MFCC				
Nº de mezclas	**Nº unidades EER < 30 %**	**Media EER (%)**	**Min EER (%)**	**Max EER (%)**
1024	3	25,38	22,49	27,26
512	4	25,41	22,9	28,87
256	6	25,69	18,42	29,85
128	12	26,42	17,93	29,42
64	16	26,19	17,58	29,98
32	17	25,44	16,14	29,73
16	19	25,08	16,33	29,25
8	**19**	**24,95**	**15,92**	**29,39**

Tabla 16. Comparación de rendimiento entre coeficientes cepstrales (MFCC) y sus contornos (TCLU-MFCC) para unidades de fonemas.

De acuerdo con la Tabla 16, desde una perspectiva general, el rendimiento es mucho mayor para el caso de coeficientes cepstrales, dado que existen muchos fonemas que permiten un rendimiento por debajo del 30 % a comparación de los contornos temporales. Sin embargo, mediante la fusión de valores de LR, en ambos casos, los rendimientos son muy parecidos, tal como se muestra en la siguiente tabla:

Tipo	Nº de unidades fusionadas	Tipo de fusión	EER (%)	minDCF	C_{llr}	$minC_{llr}$
MFCC	12	Suma promedio	**6,22**	0,0310	**0,37**	**0,23**
TCLU-MFCC	17	Suma promedio	**7,11**	0,0420	0,61	0,27

Tabla 17. Valores de EER (%), minDCF, C_{llr} y $minC_{llr}$ de la fusión de sistemas con MFCC y TCLU-MFCC para unidades de fonemas.

De acuerdo con la Tabla 17 la menor pérdida de calibración ($\mathbf{C_{llr}}$ - $\mathbf{minC_{llr}}$) se presenta en los coeficientes cepstrales (MFCC), pero el nivel de discriminación sigue siendo buena para ambos casos, tal como se muestra en la siguientes curvas tippet de la Figura 37. En ella se observa que la proporción de valores de LR que apoyan la hipótesis contraria, en ambos casos, es muy pequeña.

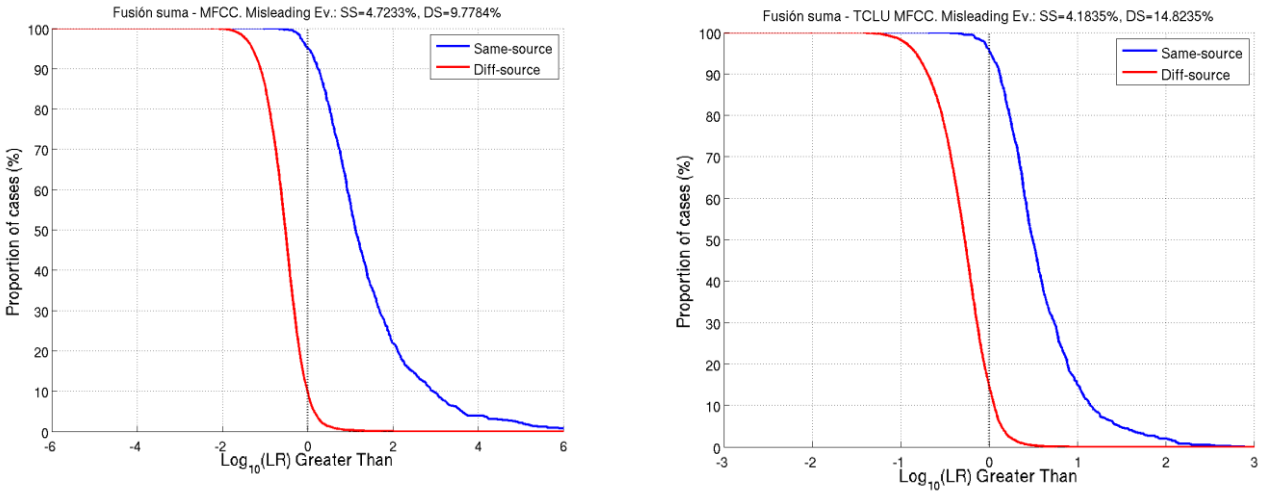

Figura 37. Curvas tippet de la fusiones suma de MFCC (256 mezclas) y TCLU-MFCC (8 mezclas).

4.4.4.2 TCLU-Formantes

Para estas pruebas se han empleado los contornos temporales de las tres primeras frecuencias de formantes. Para el caso de formantes, se ha usado la configuración dependiente total de unidad: modelos de locutor, modelo UBM y datos de test dependientes de fonema.

Primeramente, usando las puntuaciones por trial, se ha realizado pruebas del sistema para determinar el número de mezclas óptimo. Se ha probado con 8 hasta 1024 mezclas de gaussianas.

Nº mezclas	Nº unidades EER < 30 %	Media EER (%)	Min EER (%)	Max EER (%)
1024	5	27,63	24,92	29,11
512	5	26,24	23,58	29,35
256	10	27,14	23,11	29,82
128	**13**	**27,20**	**22,46**	**29,96**
64	11	26,41	22,55	29,58
32	12	26,42	21,67	29,18
16	12	27,42	23,44	30,58
8	8	27,70	24,65	29,98

Tabla 18. Valores medios, mínimos y máximos de EER (%) por número de mezcla, TCLU-Formantes.

De acuerdo con los resultados de la Tabla 18 el número de mezclas óptimo es igual a 128 mezclas gaussianas. Esta elección se basa en que para 128 mezclas existen mayor número de unidades de fonemas con rendimientos por debajo del 30 %.

Una vez obtenido el número de mezclas óptimo para el sistema se procede a la evaluación por fonema mediante valores LR. En la Tabla 19 se presentan los fonemas con valores de EER por debajo de 30 %. Se observa que el rendimiento es similar al caso de contornos de coeficientes cepstrales (Tabla 14). Además, presentan valores pequeños de pérdida de calibración (C_{llr} - $minC_{llr}$), el cual permite obtener información útil de los valores de LR calibrados por locución.

Fonema	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AE	23,3	0,0882	0,7355	0,6848
AH	28,29	0,0977	0,8483	0,8142
AX	28,08	0,0964	0,8187	0,7877
AY	**22,46**	**0,0898**	**0,7634**	**0,6797**
EH	29,95	0,0951	0,8388	0,8083
IH	28,06	0,0959	0,8176	0,7881
L	24,27	0,0934	0,7727	0,7235
M	29,96	0,0955	0,8784	0,8241
N	26,13	0,0906	0,7868	0,7504
OW	29,51	0,0934	0,8684	0,8046
PUH	28,01	0,0907	0,9032	0,7920
R	26,95	0,0940	0,8748	0,7713
Y	28,74	0,0943	0,8370	0,7956

Tabla 19. EER (%), minDCF, C_{llr}, $minC_{llr}$ por fonemas con valores de EER < 30 %, TCLU-Formantes.

En la Figura 38 se presenta la curva tippet para el fonema con mejor rendimiento, "AY". En ella se observa una mayor proporción de valores de LR que apoyan a las hipótesis contrarias. Para el caso de LR de "misma fuente" (Same-source) se tiene un 34,50 % de valores de LR que apoyan la hipótesis contraria, mientras que para "distinta fuente" (Diff-source), se tiene un 13,5 % de valores de LR que poyan la hipótesis contraria.

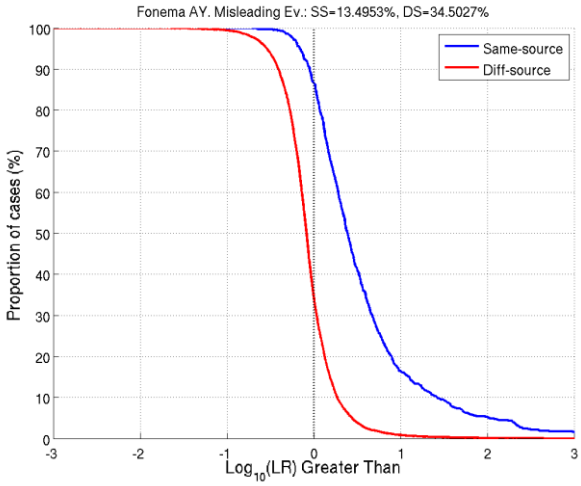

Figura 38. Curva tippet para el fonema con mejor rendimiento "AY".

Por otra parte, mediante las técnicas de fusión suma y regresión logística, en ambos casos usando el algoritmo de selección de unidad, se obtienen rendimientos similares al sistema de referencia (Figura 39). El mejor rendimiento, en términos de EER y coste de detección, se obtienen mediante la fusión por regresión logística y usando la técnica de selección de unidad, EER = 12,27 % y minDCF = 0,0663. Este rendimiento es aproximadamente un 2 % inferior respecto al sistema de referencia en términos de EER.

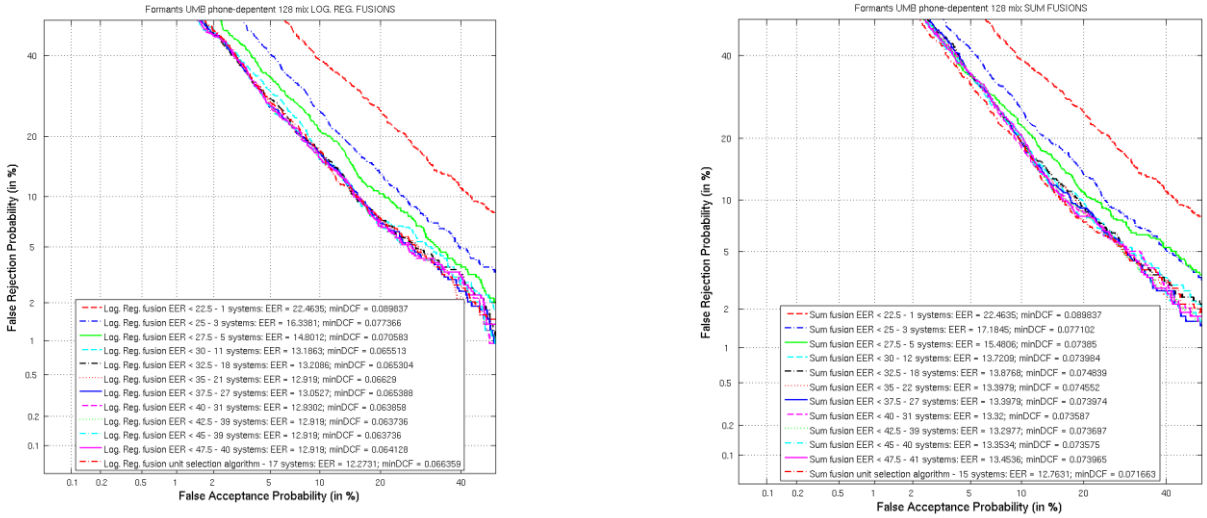

Figura 39. Curvas DET de los sistemas fusionados mediante regresión logística lineal y suma promedio con TCLU-Formantes por fonema.

Respecto a las pérdidas por calibración (C_{llr} - $minC_{llr}$), la fusión por regresión logística permite obtener un valor pequeño de pérdida por calibración (Tabla 20). En la Figura 40 se presenta la curva tippet para la fusión por regresión logística. En ella se ve claramente la mejora de la discriminación del sistema dado al valor pequeño de $minC_{llr}$. A comparación del rendimiento por unidad (Figura 38) la proporción de valores de LR que apoyan hipótesis se reduce considerablemente.

Técnica de Fusión	Nº unidades fusionadas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Suma	15	12,76	0,0716	0,75	0,46
Regresión logística lineal	**17**	**12,27**	**0,0663**	**0,5**	**0,43**

Tabla 20. Valores de EER (%), minDCF, Cllr y minCllr para cada técnica de fusión, TCLU-Formantes para unidades de fonemas.

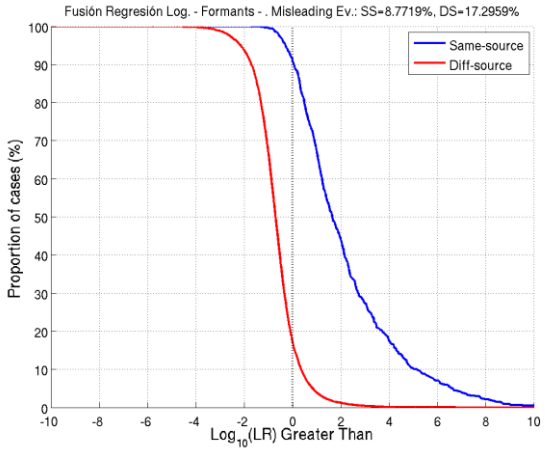

Figura 40. Curva tippet para la fusión por reg. log. con TCLU – Formantes para unidades de fonemas.

Para obtener una mejor medida del rendimiento del sistema, es necesario disponer de pruebas usando solo información de formantes (frecuencias y anchos de banda), por tanto, se ha realizado pruebas en el sistema sólo usando, como vectores de características, las tres primeras frecuencias de formantes más sus respectivos anchos de banda. Los valores de EER han sido obtenidos a partir de las puntuaciones por trial.

En la siguiente tabla se muestra una comparativa en función del rendimiento en términos de EER de cada caso.

F1, F2, F3 – FBW1, FBW2, FBW3				
Nº de mezclas	**Nº unidades EER < 30 %**	**Media EER (%)**	**Min EER (%)**	**Max EER (%)**
1024	18	25,76	20,73	29,76
512	21	25,32	20,69	29,5
256	21	24,88	19,24	29,78
128	**23**	**25,18**	**19,69**	**29,37**
64	22	24,96	20,47	29,42
32	23	25,29	20,2	29,91
16	20	26,01	22,19	29,78
8	17	26,74	23,47	29,78
TCLU - F1, F2, F3				
Nº de mezclas	**Nº unidades EER < 30 %**	**Media EER (%)**	**Min EER (%)**	**Max EER (%)**
1024	5	27,63	24,92	29,11
512	5	26,24	23,58	29,35
256	10	27,14	23,11	29,82
128	**13**	**27,2**	**22,46**	**29,96**
64	11	26,41	22,55	29,58
32	12	26,42	21,67	29,18
16	12	27,42	23,44	30,58
8	9	27,73	24,65	29,98

Tabla 21. Comparación de rendimiento entre formantes tipo 2 (F1, F2, F3 – FBW1, FBW2, FBW3) y contornos temporales de las tres primeras frecuencia de formantes (TCLU-F1, F2, F3)

En la Tabla 21 se observa que los rendimientos para ambos casos están muy cercanos, siendo el número de mezclas óptimo igual a 128. Sin embargo, el rendimiento de ambos no supera el del sistema de referencia, solo mediante la fusión, en caso de contornos temporales, se llega a un rendimiento cercano al sistema de referencia.

4.4.4.3 Influencia del número de orden de la DCT para la codificación de contornos temporales

En la parametrización de los contornos temporales se busca aproximar dichas contornos mediante el ajuste polinomial o la DCT, siendo este último el mejor candidato debido a la compactación de la información en pocos coeficientes y a las propiedades pseudo-ortogonales presentes en los coeficientes DCT.

Con el objetivo de analizar la influencia del tipo de parametrización DCT de los contornos temporales de los coeficientes/formantes en el rendimiento del sistema, se ha realizado pruebas para la configuración óptima del sistema (128 mezclas para TCLU-Formantes y 8 mezclas para TCLU-MFCC) y con parametrizaciones DCT de orden 5, 7, y 9 de los contornos temporales. En la siguiente tabla se muestra los resultados de estas pruebas:

	TCLU-Formantes EER (%)			TCLU-MFCC EER (%)		
Fonema	**DCT - Orden 5**	**DCT - Orden 7**	**DCT - Orden 9**	**DCT - Orden 5**	**DCT - Orden 7**	**DCT - Orden 9**
AE	23,3	27,73	30,17	18,98	22,09	22,09
AH	28,29	32,88	36,26	29,39	30,05	32,48
AX	28,08	31,41	34,68	27,08	28,35	29,11
AY	22,46	24,59	28,7	21,68	24,92	24,98
IH	28,06	30,9	33,83	26,98	27,65	29,96
L	24,27	26,31	31,04	26,44	29,18	30,77
M	29,96	32,77	38,19	22,26	23,84	26,04
N	26,13	28,7	33,28	15,92	17,3	18,72
OW	29,51	31,94	34,36	24,63	27,64	28,03
PUH	28,01	31,17	33,79	24,15	25,36	27,11

Tabla 22. Valores de EER en función de distintos órdenes de parametrización DCT para unidades de fonemas.

En la Tabla 22 se observa que el orden del DCT influye de manera negativa, degradando el rendimiento del sistema. Una de las posibles razones es que la aplicación de este tipo de DCT (orden 7 y 9) hace que muchas peculiaridades de los contornos se pierdan al momento de parametrizarlos. Este comportamiento también se presenta para las unidades de difonemas y trifonemas.

4.4.5 Pruebas a nivel de difonema

Estas pruebas corresponden a la evaluación individual por difonema correspondiente a la evaluación NIST SRE 2006 con la tarea **1conv4wire-1conv4wire** en idioma inglés.

4.4.5.1 TCLU-MFCC

Para estas pruebas a nivel de difonema, al igual que para el caso de fonemas, se han empleado los contornos temporales de los coeficientes cepstrales sin compensación de variabilidad y del tipo 1 (19 coeficientes estáticos). De acuerdo con la Tabla 10, se ha empleado la configuración dependiente total de unidad: modelos de locutor, modelo UBM y datos de test dependientes de difonema.

Tal como muestra la Tabla 10, el número de mezclas óptimo para las unidades de fonemas es igual a 4 mezclas. Para llegar a este resultado, se ha probado el sistema desde 2 hasta 1024 mezclas de gaussianas con un incremento igual a potencia de 2.

En la siguiente tabla se presenta el rendimiento del sistema por unidad cuyo rendimiento, en términos de EER, esté por debajo del 30 %. Estos valores de EER han sido obtenidos a partir de las puntuaciones por trial.

Nº mezclas	Nº unidades. EER < 30 %	Media EER (%)	Min EER (%)	Max EER (%)
16	3	26,62	25,99	27,62
8	4	26,94	24,59	29,75
4	**5**	**26,94**	**23,84**	**29,79**
2	3	28,02	25,6	29,37

Tabla 23. Valores medios, mínimos y máximos de EER (%) del rendimiento del sistema TCLU-MFCC por difonema cuyos valores de EER están por debajo de 30 %.

De acuerdo con la Tabla 23, el mejor rendimiento se obtiene para 4 mezclas ya que presenta más unidades de difonemas con rendimiento por debajo del 30 %, y además, presenta valores bajos de media, mínimo y máximo de EER. Con el número de mezclas óptimo para el sistema se realiza la evaluación por unidad de difonema en función de valores de LR (puntuaciones calibradas). En la siguiente tabla (Tabla 24) se presenta los diez mejores difonemas con mayor rendimiento en el sistema.

Difonema	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AEN	30,72	0,0993	0,8479	0,823
AET	31,89	0,0969	0,872	0,8526
AXN	**23,84**	**0,0899**	0,7583	0,7097
AYK	32,45	0,0970	0,8494	0,8356
LAY	29,11	0,0972	0,8156	0,7955
ND	24,92	0,0876	0,7563	0,7037
NOW	30,86	0,0995	0,8455	0,8185
UWN	32,20	0,0953	0,8417	0,8188
YAE	29,78	0,0976	0,8383	0,8094
YUW	27,18	0,0960	0,8223	0,7812

Tabla 24. EER (%), minDCF, C_{llr} y $minC_{llr}$ de los diez mejores difonemas que presentan mayor rendimiento en el sistema, TCLU-MFCC.

De acuerdo con la Tabla 24, se observa que el rendimiento del sistema por difonema es muy bajo a comparación de los fonemas. Una posible razón es que las unidades de difonemas comprenden intervalos de tiempos más largos que los fonemas, por tanto, éstas presentan trayectorias mucho más complejas que la de los fonemas. Por otra parte, sí presentan buena calibración por unidad debido a la poca pérdida de calibración (C_{llr} - $minC_{llr}$). No obstante, este rendimiento se mejora mediante la fusión de unidades de difonemas, el cual permite llegar a un rendimiento similar al conseguido por unidades de fonemas (Figura 41).

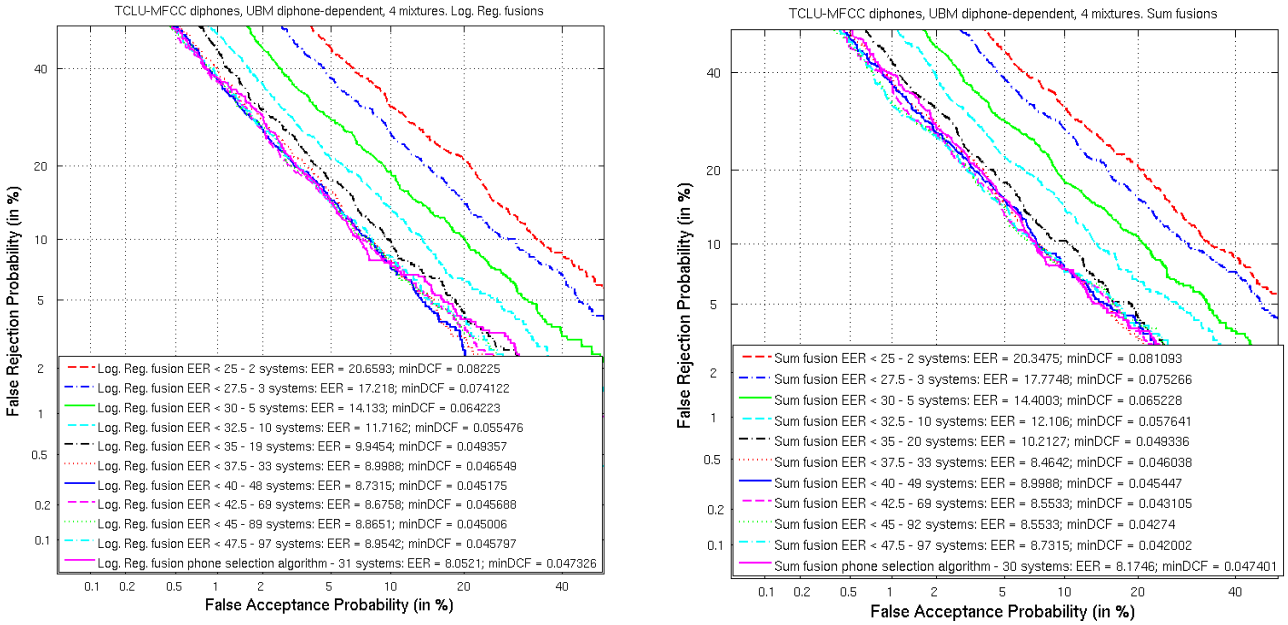

Figura 41. Curvas DET de los sistemas fusionados a nivel de difonema mediante suma y regresión logística con TCLU-MFCC.

Se observa en la Figura 41 que el mejor rendimiento se obtiene mediante la fusión regresión logística con el algoritmo de selección de unidad, EER = 8,05 % y minDCF = 0,0473. Por otra parte, este rendimiento es un 20 % superior respecto al sistema de referencia en términos de EER. En la siguiente tabla se muestra las pérdidas de calibración por técnica de fusión:

Técnica de Fusión (selección de unidad)	Nº de unidades fusionadas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Suma promedio	30	8,17	0,0474	0,79	0,31
Regresión logística lineal	**31**	**8,05**	**0,0473**	**0,39**	**0,31**

Tabla 25. Valores de EER (%), minDCF, C_{llr} y $minC_{llr}$ para cada técnica de fusión, TCLU-MFCC por difonema.

En la Tabla 25 se observa que la menor pérdida de calibración se obtiene con la fusión por regresión logística, el cual permite obtener valores de LR bien calibrados.

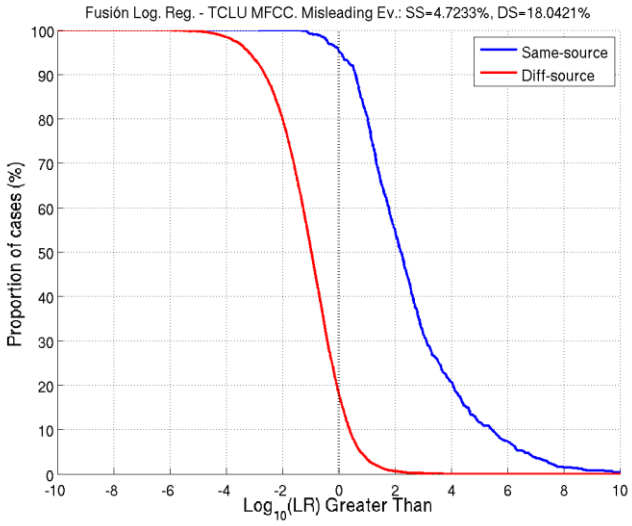

Figura 42. Curva tippet de la fusión reg. log. con TCLU-MFCC por difonema y 4 mezclas .

En la Figura 42 se observa la buena discriminación de los valores de LR, esto se manifiesta en la poca proporción de valores de LR que apoyan a las hipótesis contrarias. Alrededor de un 4,72 % de LR correspondiente a la hipótesis de "misma fuente" (Same-source) que apoyan la hipótesis contraria, y un 18,04 % de LR correspondiente a "diferente fuente" (Diff-source) que apoyan la otra hipótesis.

En la siguiente se presenta una comparación del rendimiento del sistema usando fonemas y difonemas.

Unidad	Tipo de fusión	Nº unidades fusionadas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Fonema	Suma promedio	**17**	**7,11**	0,0420	0,61	**0,27**
Difonema	Reg. Logística	**31**	**8,05**	0,0473	0,39	**0,31**

Tabla 26. EER (%), minDCF, C_{llr} y min C_{llr} de fusión óptima de sistemas para fonema y difonema, TCLU-MFCC.

El rendimiento obtenido de la fusión a nivel de unidades de difonemas, tal como se muestra en la Tabla 26, es similar al obtenido por los fonemas. Sin embargo, el nivel de discriminación es menor respecto al de los fonemas. Esto se aprecia en los valores de EER y $minC_{llr}$, los cuales son mayores que el para el caso de los fonemas.

Adicionalmente, con el fin de comparar el rendimiento del sistema usando sólo coeficientes cepstrales, en la siguiente tabla se muestra el rendimiento óptimo del sistema usando coeficientes cepstrales (MFCC) y sus contornos temporales (TCLU-MFCC). Estos valores de EER se han obtenido a partir de las puntuaciones por trial.

MFCC				
Nº de mezclas	**Nº unid. EER < 30 %**	**Media EER (%)**	**Min EER (%)**	**Max EER (%)**
128	35	26,33	17,55	29,99
TCLU-MFCC				
Nº de mezclas	**Nº unid. EER < 30 %**	**Media EER (%)**	**Min EER (%)**	**Max EER (%)**
4	5	26,94	23,84	29,79

Tabla 27. Comparación de rendimiento entre coeficientes cepstrales (MFCC) y sus contornos (TCLU-MFCC) para unidades de difonemas.

Se observa en la Tabla 27 que para el caso de MFCC se tiene más unidades cuyo rendimiento están por debajo del 30 %. Por otra parte, en la fusión a nivel de unidades de difonema (Tabla 28) el rendimiento en los coeficientes cepstrales (MFCC) es aproximadamente un 27 % superior al de los obtenidos por los contornos temporales (TCLU-MFCC) en términos de EER.

Tipo	Nº unid. fusionadas	Tipo de fusión	EER (%)	minDCF	C_{llr}	$minC_{llr}$
MFCC	23	Suma promedio	**5,82**	0,0364	0,63	0,25
TCLU-MFCC	30	Reg. logística	**8,05**	0,0420	0,39	0,31

Tabla 28. Valores de EER (%), minDCF, C_{llr} y $minC_{llr}$ de la fusión de sistemas con MFCC y TCLU-MFCC para unidades de difonemas.

4.4.5.2 TCLU-Formantes

Para estas pruebas se han empleado los contornos temporales de las tres primeras frecuencias de formantes. Para el caso de formantes, se ha usado la configuración dependiente total de unidad: modelos de locutor, modelo UBM y datos de test dependientes de difonema.

Primeramente, se ha realizado pruebas del sistema para determinar el número de mezclas óptimo. Se ha probado con 2 hasta 1024 mezclas de gaussianas. En la siguiente tabla se presenta el rendimiento de los sistemas en términos de EER por debajo del 30 %. Estos valores de EER han sido obtenidos a partir de las puntuaciones por trial.

Nº mezclas	Nº sist. EER < 30 %	Media EER (%)	Min EER (%)	Max EER (%)
128	1	29,91	29,91	29,91
64	1	29,62	29,62	29,62
32	**3**	**29,23**	**28,93**	**29,76**
16	2	27,97	27,73	28,22

Tabla 29 . Valores medios, mínimos y máximos de EER (%) del rendimiento del sistema TCLU-Formantes por difonema cuyos valores de EER están por debajo de 30 %.

Obtenido el número de mezclas óptimo, se procede al cálculo de los valores de LR mediante regresión logística. En la siguiente tabla se muestra el rendimiento de los cinco mejores difonemas.

Difonema	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AXN	30,76	0,0980	0,8612	0,8379
LAY	31,27	0,0969	0,8439	0,8129
UWN	30,99	0,0942	0,8609	0,8196
YAE	**30,01**	**0,0953**	**0,8355**	**0,8114**
YUW	30,09	0,0979	0,8375	0,8089

Tabla 30. EER (%), minDCF, C_{llr} y $minC_{llr}$ de los cinco mejores difonemas que presentan mayor rendimiento en el sistema, TCLU-Formantes.

De acuerdo con la Tabla 30 el rendimiento de los contornos de las frecuencias de formantes es peor respecto al caso de los fonemas (Tabla 19). Esto se debe, al igual que sucede en el caso de los contornos temporales de coeficientes cepstrales, las trayectorias temporales de las frecuencias abarcan más tiempo, por tanto, son más complejas. Respecto al sistema de referencia, presentan un rendimiento por debajo del 90 %, sin embargo, mediante la fusión a nivel intra (unidades de difonemas) se llega a un rendimiento cercano al del sistema de referencia. En la siguiente figura se presenta las fusiones a nivel de difonema:

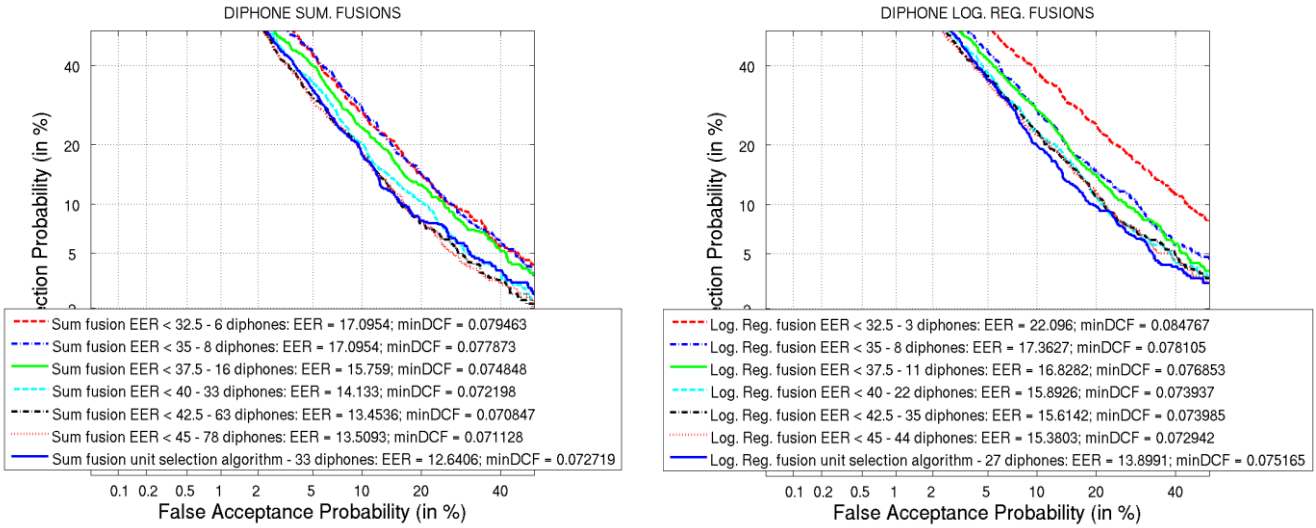

Figura 43. Curvas DET de los sistemas fusionados a nivel de difonema mediante suma y regresión logística con TCLU-Formantes.

En la Figura 43 se observa que el mejor rendimiento se obtiene con la fusión suma promedio y usando la técnica de selección de unidad, EER = 12,64 % y minDCF = 0,0727. En comparación con el sistema de referencia, el rendimiento del sistema fusionado es un 20 % peor en términos de EER.

Respecto a las pérdidas de calibración (C_{llr} - $minC_{llr}$), la fusión por regresión logística presenta una menor pérdida de calibración respecto a la fusión suma (Tabla 31). En la Figura 44 se presentan las curva tippet de cada tipo de fusión.

Técnica de Fusión	Nº unidades fusionadas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Suma promedio	33	**12,64**	**0,0727**	**0,86**	**0,47**
Regresión logística lineal	27	13,89	0,0751	0,55	0,49

Tabla 31. Valores de EER (%), minDCF, C_{llr} y $minC_{llr}$ para cada fusión suma y regresión logística con TCLU-Formantes por difonemas.

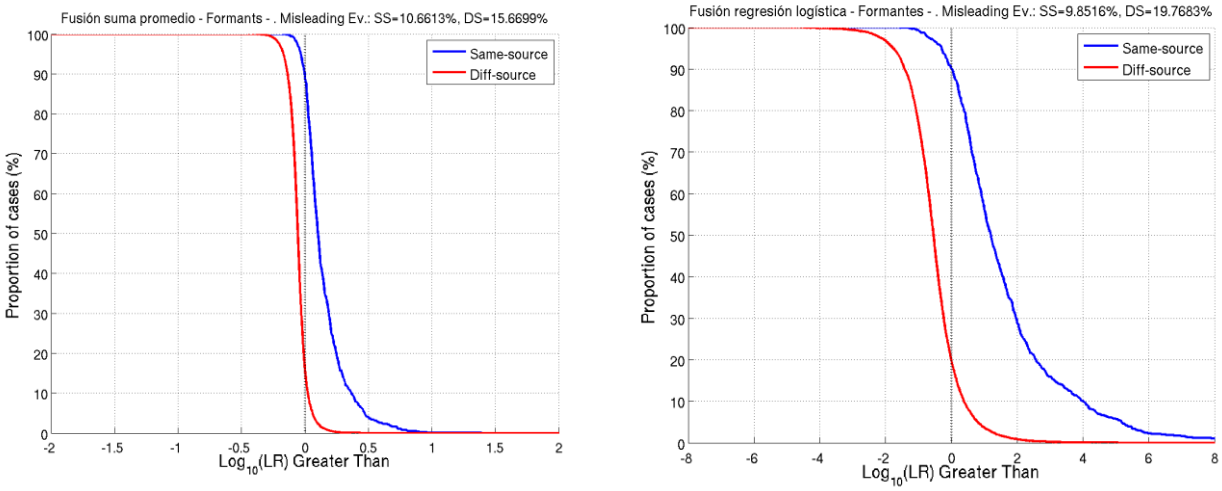

Figura 44. Curvas tippet de fusión suma y regresión logística con TCLU-Formantes por difonema.

La diferencia entre las pérdidas de discriminación ($minC_{llr}$) de ambas fusiones se aprecia mejor en las curvas tippet de la Figura 44, donde para el caso de la fusión por regresión logística existe una mayor proporción de valores de LR que apoyan sus respectivas hipótesis, es decir, el grado de apoyo es mucho mayor a comparación de la técnica de fusión suma promedio.

Por otra parte, comparando los resultados obtenidos con los fonemas, el rendimiento es similar, con la excepción que en los difonemas se requiere de más unidades a fusionar para llegar al rendimiento óptimo. En la Tabla 32 se presenta la comparación de los rendimientos óptimos por unidad.

Unidad	Tipo de fusión	N º unidades fusionadas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Fonema	Reg. Logística	**17**	**12,27**	0,0663	0,5	0,43
Difonema	Suma promedio	**33**	**12,64**	0,0727	0,86	0,47

Tabla 32. EER (%), minDCF, Cllr y min Cllr de fusión óptima de sistemas para fonema y difonema, TCLU-Formantes.

Adicionalmente, se ha realizado pruebas del sistema usando, como vectores de características, las tres primeras frecuencias y sus respectivos anchos de banda (F1, F2, F3 – FBW1, FBW2, FBW3). Se debe tener en cuenta que los valores de EER (Equal Error Rate) se han obtenido a partir de las puntuaciones por trial.

Como resultado de esta prueba, se obtiene como número de mezclas óptimo igual a 32 mezclas. En la siguiente tabla se presenta una comparación entre los dos sistema: Formantes (F1, F2, F3 – FBW1, FBW2, FBW3) y TCLU-Formantes (TCLU - F1,F2,F3).

Tipo de característica	Nº de mezclas	Nº unid. EER < 30 %	Media EER (%)	Min EER (%)	Max EER (%)
Formantes	32	23	27,93	22,70	29,98
TCLU – Formantes	32	3	29,23	28,93	29,76

Tabla 33. Comparación de rendimientos en función de EER para Formantes y TCLU-Formantes a nivel de trifonema.

4.4.6 Pruebas a nivel de trifonema

4.4.6.1 TCLU-MFCC

Para estas pruebas a nivel de trifonema, al igual que para el caso de fonemas, se han empleado los contornos temporales de los coeficientes cepstrales sin compensación de variabilidad y del tipo 1 (19 coeficientes estáticos). De acuerdo con la Tabla 10, se ha empleado la configuración dependiente total de unidad: modelos de locutor, modelo UBM y datos de test dependientes de trifonema.

Tal como muestra la Tabla 10, el número de mezclas óptimo para las unidades de trifonemas es igual a 4 mezclas. Para llegar a este resultado, se ha probado el sistema desde 2 hasta 1024 mezclas de gaussianas con un incremento igual a potencia de 2. En la siguiente tabla se presenta el rendimiento del sistema por unidad cuyo rendimiento, en términos de EER, esté por debajo del 30 %. Estos valores de EER han sido obtenidos a partir de las puntuaciones por trial.

Nº mezclas	Nº unidades. EER < 30 %	Media EER (%)	Min EER (%)	Max EER (%)
8	1	29,86	29,86	29,86
4	**1**	**28,44**	**28,44**	**28,44**
2	1	29,58	29,58	29,58
1	1	28,96	28,96	28,96

Tabla 34. Valores medios, mínimos y máximos de EER (%) del rendimiento del sistema TCLU-MFCC por trifonema cuyos valores de EER están por debajo de 30 %.

Se observa en la Tabla 34 que el número de mezclas óptimo es igual a cuatro ya que presenta el menor valor medio de EER como también los valores máximos y mínimos. Por otra parte, el rendimiento es muy bajo donde sólo una unidad por mezclas tiene un valor de EER por debajo del 30 %. Con este resultado y los presentados en la parte de fonemas y difonemas, se confirma que a medida que la longitud de la unidad es mayor, la trayectoria de dicha unidad incrementa en complejidad, resultando difícil paramétrizar dichos coeficientes relacionados a los contornos.

Con el número mezclas óptimo se procede a evaluar el sistema por trifonema mediante los valores de LR (puntuaciones calibradas). En la siguiente tabla se presenta el rendimiento del sistema con cuatro mezclas gaussianas correspondiente a los cinco mejores unidades de trifonemas.

Trifonema	EER (%)	minDCF	C_{llr}	$minC_{llr}$
DHAET	**33,96**	**0,0992**	**0,8872**	**0,8715**
THIHNG	36,26	0,0975	0,9104	0,8752
AXND	37,48	0,0994	0,9263	0,9102
LAYK	38,16	0,0990	0,9056	0,8443
YAE-	39,09	0,1	1,0101	0,9593

Tabla 35. EER (%), minDCF, C_{llr} y $minC_{llr}$ de los cinco mejores trifonemas que presentan mayor rendimiento en el sistema, TCLU-MFCC.

De acuerdo con la Tabla 35, el rendimiento está muy por encima del 30 %, sin embargo, los valores de pérdida de calibración se mantienen en valores pequeños. Por otra parte, aplicando fusión a nivel de trifonema el rendimiento mejora pero es aproximadamente un 50 % peor respecto al sistema de referencia en términos de EER. En la siguiente figura se presenta las curvas DET correspondientes a cada técnica de fusión.

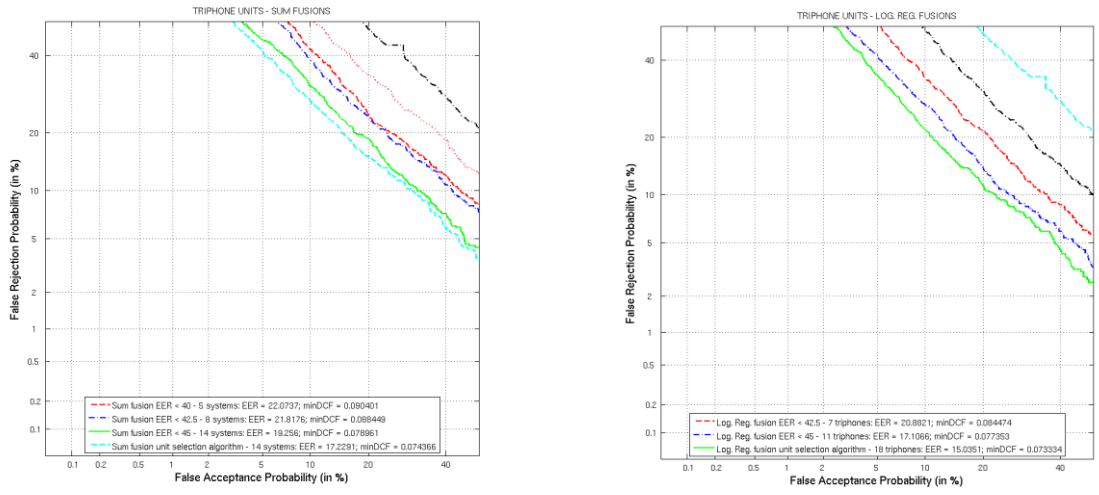

Figura 45. Curvas DET de los sistemas fusionados a nivel de trifonema mediante suma y regresión logística con TCLU-MFCC.

Se observa en la Figura 45 que el mejor rendimiento se obtiene usando el algoritmo de selección por unidad en la fusión por regresión logística, con un valor de EER = 15,03 % y minDCF = 0,0733. Respecto a la pérdida de calibración, en la Tabla 36 se observa que la menor pérdida de calibración (C_{llr} - $minC_{llr}$) se presenta en la fusión por regresión logística.

Técnica de Fusión	Nº unidades fusionadas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Suma promedio	14	17,22	0,0743	0,85	0,56
Regresión logística lineal	**14**	**15,03**	**0,0733**	**0,54**	**0,5**

Tabla 36. Valores de EER (%), minDCF, C_{llr} , $minC_{llr}$ y C_{llr}^{cal} para cada fusión suma y regresión logística con TCLU-Formantes por trifonemas.

Respecto al nivel de discriminación entre cada técnica de fusión se visualiza mucho mejor en la curvas tippet de la Figura 46. En ella se observa que para la fusión por regresión logística la proporción de valores de LR que apoyan a la hipótesis incorrecta es menor que en el caso de la fusión suma promedio.

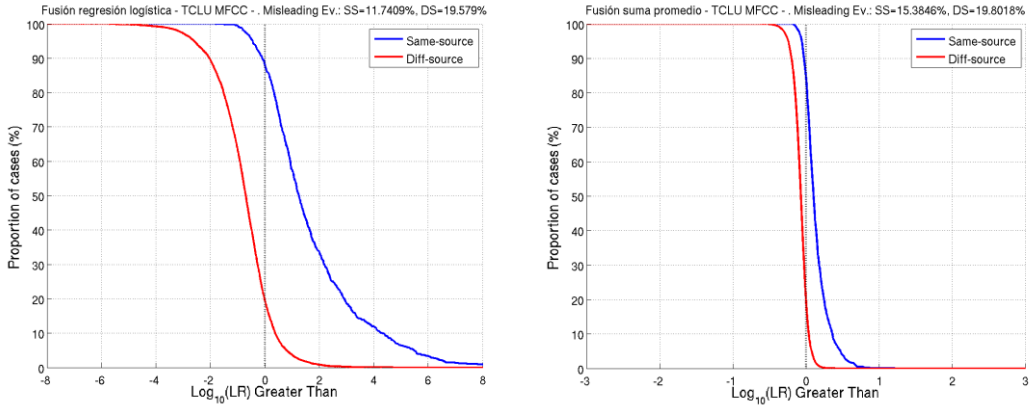

Figura 46. Curvas tippet de fusión regresión logística y suma con TCLU-MFCC por trifonema.

Comparando el rendimiento del sistema fusionado a distintos niveles de unidad, tal como se muestra en la Tabla 37, se observa que el rendimiento empeora a medida que usamos unidades de mayor longitud. Tal como se explicó en resultados anteriores, una de las posibles razones es la complejidad de paramétrizar los contornos temporales de la unidades. Otra posible razón se relaciona con la frecuencia de aparición de las unidades en las locuciones, por tanto, el número de veces que se repite un difonema o trifonema en una locución de 20 ms de duración (2.4.1 Coeficientes MFCC (Mel-Frecuency Cepstral Coefficients)) es mucho menor respecto a las unidades de fonemas. Se debe tener en cuenta que las unidades que mayor frecuencia de aparición presentan son los fonemas dado que son las unidades más cortas, de ahí el mejor rendimiento que poseen en todas las pruebas.

Unidad	Tipo de fusión	Nº unidades fusionadas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Fonema	Suma promedio	**17**	**7,11**	0,0420	0,61	0,27
Difonema	Reg. Logística	**31**	**8,05**	0,0473	0,39	0,31
Trifonema	Reg. Logística	**14**	**15,03**	0,0733	0,54	0,5

Tabla 37. Valores de EER (%), minDCF, C_{llr} y $minC_{llr}$ de sistemas fusionados a nivel de unidad: fonema, difonema y trifonema. TCLU-MFCC.

En la siguiente tabla se presenta una comparación entre los rendimientos óptimos del sistema usando sólo coeficientes cepstrales y contornos temporales. Los valores de EER se han obtenido a partir de las puntuaciones por trial.

MFCC				
Nº de mezclas	**Nº unid. EER < 30 %**	**Media EER (%)**	**Min EER (%)**	**Max EER (%)**
128	**12**	26,52	23,15	29,6
TCLU-MFCC				
Nº de mezclas	**Nº unid. EER < 30 %**	**Media EER (%)**	**Min EER (%)**	**Max EER (%)**
4	**1**	28,44	28,44	28,44

Tabla 38. Comparación de rendimiento entre coeficientes cepstrales (MFCC) y sus contornos (TCLU-MFCC) para unidades de trifonemas.

En la Tabla 38 se observa que para coeficientes cepstrales (MFCC) existe un mayor número de unidades cuyo rendimiento, en términos de EER, es menor que el 30 %. Esta prevalencia se mantiene en la fusión de a nivel de unidades de trifonemas, donde se alcanza un mejor rendimiento respecto al sistema de referencia, caso contrario con los contornos temporales. La siguiente tabla recoge estos resultados.

Tipo	Nº unid. fusionadas	Tipo de fusión	EER (%)	minDCF	C_{llr}	$minC_{llr}$
MFCC	17	Reg. Logística	**9,88**	0,0595	0,4	0,35
TCLU-MFCC	14	Reg. Logística	**15,03**	0,0733	0,54	0,5

Tabla 39. Valores de EER (%), minDCF, C_{llr} y $minC_{llr}$ de la fusión de sistemas con MFCC y TCLU-MFCC para unidades de trifonemas.

4.4.6.2 TCLU-Formantes

Para estas pruebas se han empleado los contornos temporales de las tres primeras frecuencias de formantes. Para el caso de formantes, se ha usado la configuración dependiente total de unidad: modelos de locutor, modelo UBM y datos de test dependientes de trifonema.

Primeramente, se ha realizado pruebas del sistema para determinar el número de mezclas óptimo. Se ha probado con 16 hasta 256 mezclas de gaussianas.

El resultado de esta prueba indica que el rendimiento para cada trifonema está por encima del 30%. Por tanto, la elección del número de mezclas óptimo se basará en la menor media que presente. En la siguiente tabla se presenta los valores medios de EER para diferentes números de mezclas.

Nº de mezclas	Valor medio EER (%)
256	42,64
128	40,08
64	39,46
32	**38,37**
16	38,44

Tabla 40. Valores medios de EER para diferentes números de mezclas. TCLU-Formantes-trifonemas.

De acuerdo con la Tabla 40, el número de mezclas óptimo es igual a 32 con un valor medio de EER igual a 38,37 %. A continuación, se procede a calcular los valores de LR a partir de las puntuaciones obtenidas para dicho número de mezclas. En la siguiente tabla se muestra los cinco trifonemas con mejor rendimiento en términos de EER.

Trifonema	EER (%)	minDCF	C_{llr}	$minC_{llr}$
DHAET	**33,29**	**0,0989**	**0,8973**	**0,8739**
YUWN	33,40	0,0988	0,9109	0,8708
LAYK	34,19	0,0979	0,8869	0,8618
UWNOW	35,68	0,0996	0,9010	0,8756
YAE-	36,39	0,0999	0,9298	0,9076

Tabla 41. EER (%), minDCF, C_{llr} y $minC_{llr}$ de los cinco mejores trifonemas que presentan mayor rendimiento en el sistema, TCLU-Formantes.

Se observa en la Tabla 41 que el mejor rendimiento se obtiene con el trifonema DHAET, con un valor de EER igual a 33,29 % y minDCF = 0,0989. Por otra parte, las pérdidas de calibración son bajas, esto nos permite obtener valores de LR bien calibrados. Por otra parte, mediante la fusión a nivel de unidades (intra) de trifonemas se pretende mejorar este rendimiento. En la siguiente figura se muestra el resultado de fusionar el sistema a nivel intra mediante suma promedio y regresión logística.

En la Figura 47 se observa que el mejor rendimiento se alcanza con la fusión suma promedio y el algoritmo de selección de unidad, con un valor de EER = 20,47 % y minDCF = 0,0830. El rendimiento se ha mejorado respecto al rendimiento por unidad, sin embargo, es casi un 90 % peor al sistema de referencia en términos de EER. Por otra parte, la menor pérdidad de calibración se presenta en la fusión por regresión logística (Tabla 42). En la siguiente tabla se recoge los valores de EER (%), minDCF, C_{llr} y $minC_{llr}$ para cada típo de fusión.

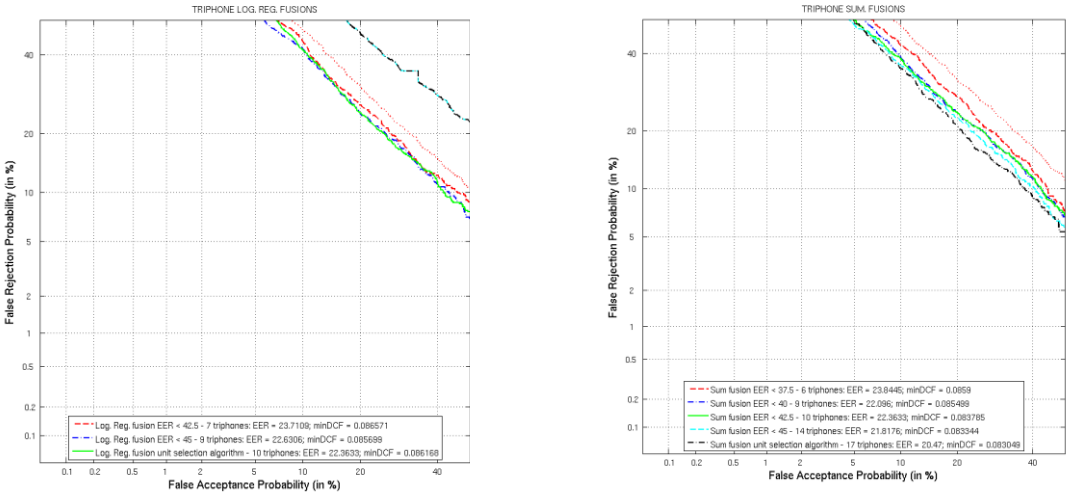

Figura 47. Curvas DET de los sistemas fusionados a nivel de trifonema mediante suma y regresión logística con TCLU-Formantes.

Técnica de Fusión	Nº unidades fusionadas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Suma promedio	17	**20,47**	**0,0830**	**0,74**	**0,63**
Regresión logística lineal	10	22,36	0,0861	0,71	0,68

Tabla 42. Valores de EER (%), minDCF, C_{llr} y $minC_{llr}$ para cada fusión suma y regresión logística con TCLU-Formantes por trifonemas.

Respecto a la pérdida de discriminación ($minC_{llr}$), ésta es relativamente alta. Esto se traduce en una mayor tasa de error de los valores de LR que apoyan las hipótesis contrarias, siendo aproximadamente un 20 % de tasa de error en ambos casos (Figura 48).

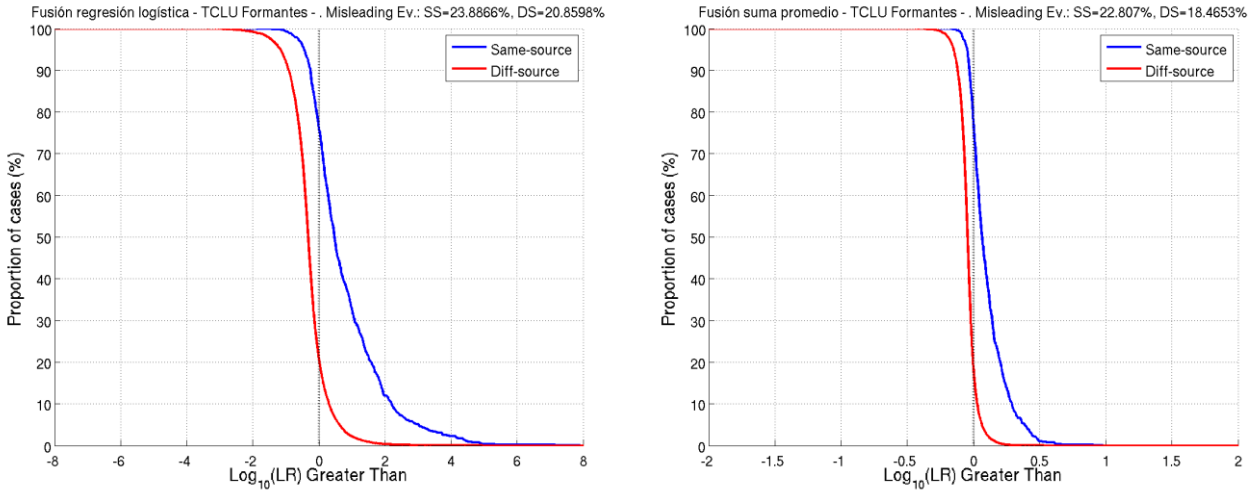

Figura 48. Curvas tippet de fusión regresión logística y suma con TCLU-Formantes por trifonema.

Por otra parte, el rendimiento óptimo de fusión suma comparado al rendimiento óptimo de fusión de otras unidades es muy bajo, aproximadamente un 60 % peor respecto a los otros. En la Tabla 43 se presenta la comparación de los rendimientos óptimos por unidad.

Unidad	Tipo de fusión	N º unidades fusionadas	EER (%)	minDCF	C_{llr}	$minC_{llr}$
Fonema	Reg. Logística	**17**	**12,27**	0,0663	0,5	0,43
Difonema	Suma promedio	**33**	**12,64**	0,0727	0,86	0,47
Trifonema	Suma promedio	**17**	**20,47**	0,0830	0,74	0,63

Tabla 43. EER (%), minDCF, Cllr y min Cllr de fusión óptima de sistemas para fonema, difonema y trifonema , TCLU-Formantes.

Adicionalmente, se ha realizado pruebas del sistema usando, como vectores de características, las tres primeras frecuencias y sus respectivos anchos de banda (F1, F2, F3 – FBW1, FBW2, FBW3). Los valores de EER se calculan a partir de las puntuaciones por trial.

La Tabla 44 presenta los resultados de esta prueba en función del número de mezclas. En ella se observa un rendimiento superior respecto a los contornos temporales. Una de las posibles razones se debe, tal como se comentó en apartados anteriores, a la complejidad de las trayectorias de cada unidad, por ende, la dificultad de paramétrizar dichas trayectorias; y considerando la poca frecuencia de ocurrencia de trifonemas en las locuciones, hace que el sistema presenta un rendimiento muy bajo usando los contornos o trayectorias temporales.

Nº de mezclas	Nº unidades EER < 30 %	Media EER (%)	Min EER (%)	Max EER (%)
1024	4	27,65	23,35	29,98
512	7	28,05	23,65	29,89
256	7	27,63	23,96	29,98
128	8	27,22	24,26	29,85
64	7	26,88	24,71	29,72
32	10	27,69	25,33	29,62
16	6	27,80	25,49	29,79
8	5	28,76	27,07	29,76

Tabla 44. Rendimiento del sistema usando tres primeras frecuencias de formantes y sus respectivos anchos de banda para unidades de trifonemas.

4.4.7 Fusión inter-unidad

En los apartados anteriores se ha presentado los resultados de las fusiones realizadas a nivel de unidad, también llamado intra-unidad. En base a los resultados se ha visto cuan bien funcionan la fusión unidades, logrando una gran mejora en el rendimiento del sistema.

En este apartado se presentará los resultados de la fusión inter-unidad, es decir, se fusionará todos los sistemas con la configuración óptima que se obtuvo para cada unidad. Adicionalmente, se fusionará los sistemas combinando los diferentes tipos de características, como por ejemplo la fusión entre los sistemas de contornos temporales de coeficientes cepstrales (TCLU-MFCC) y contornos de frecuencias de formantes (TCLU-Formantes).

En la siguiente tabla (Tabla 45) se recoge las configuraciones óptimas para cada tipo de parametrización de contorno.

	TCLU-MFCC	TCLU-Formantes
Unidad	**Nº mezclas óptimo**	**Nº mezclas óptimo**
Fonema	8	128
Difonema	4	32
Trifonema	4	32

Tabla 45. Configuración óptima para cada sistema por unidad usando TCLU-MFCC y TCLU-Formantes.

En las siguientes figuras se presenta las fusiones inter-unidad por tipo de característica.

TCLU-MFCC

- Fonemas + Difonemas

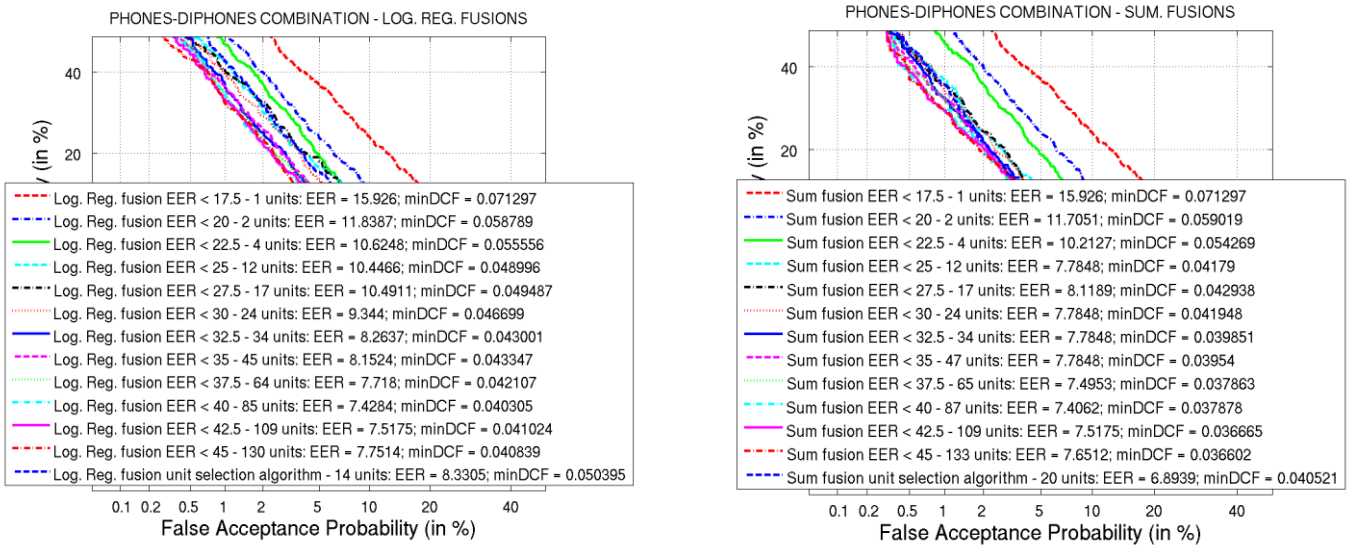

Figura 49. Curvas DET de fusiones regresión logística y suma promedio a nivel inter-unidad (fonemas y difonemas), TCLU –MFCC.

- Fonemas + Difonemas + Trifonemas

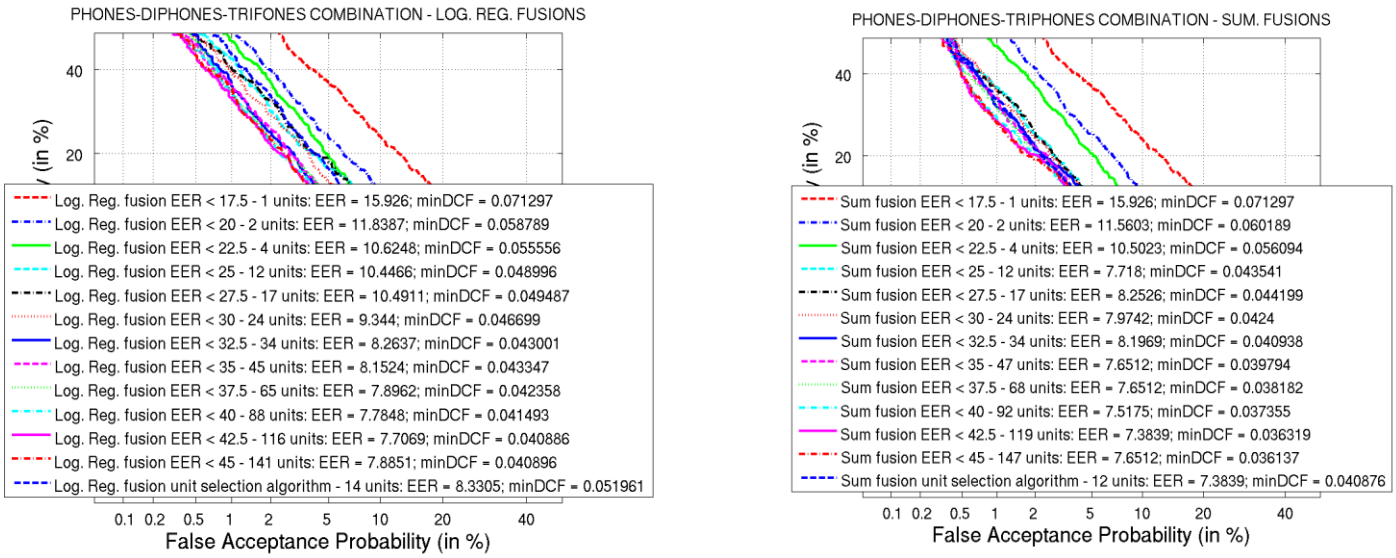

Figura 50. Curvas DET de fusiones regresión logística y suma promedio a nivel inter-unidad (fonemas, difonemas y trifonemas), TCLU-MFCC.

TCLU-Formantes

- Fonemas + Difonemas

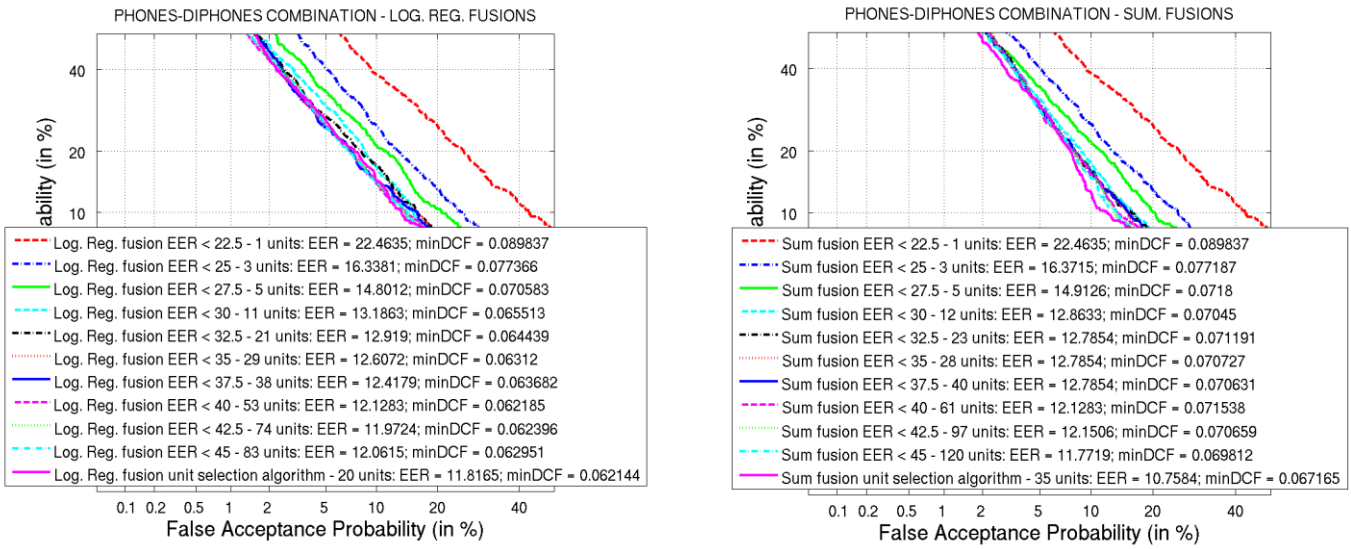

Figura 51. Curvas DET de fusiones regresión logística y suma promedio a nivel inter-unidad (fonemas, y difonemas), TCLU-Formantes.

- Fonemas + Difonemas + Trifonemas

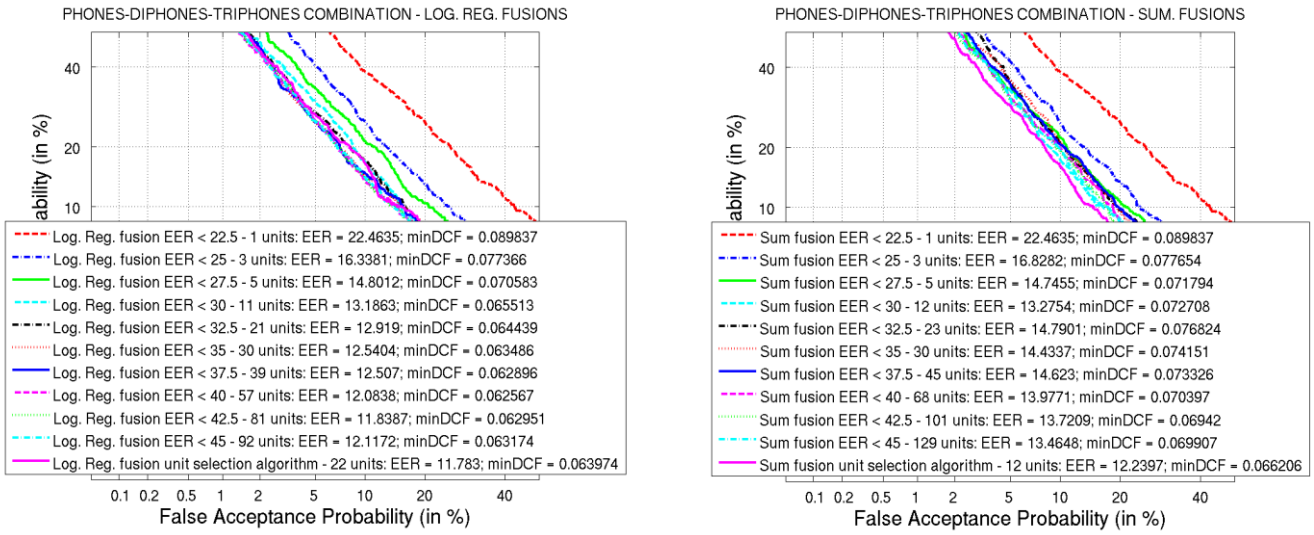

Figura 52. Curvas DET de fusiones regresión logística y suma promedio a nivel inter-unidad (fonemas, difonemas y trifonemas), TCLU-Formantes.

En la siguiente tabla se presenta la fusión óptima inter-unidad para contornos de coeficientes cepstrales (TCLU-MFCC) y formantes (TCLU-Formantes).

Tipo de característica	Tipo de fusión	Unidades	ERR (%)	minDCF	C_{llr}	$minC_{llr}$
TCLU MFCC	**Suma promedio**	**fonemas, difonemas**	**6,89**	**0,0405**	0,64	0,26
	Suma promedio	**fonemas, difonemas y trifonemas**	**7,38**	**0,0363**	0,77	0,27
	Regresión logística	fonemas, difonemas	7,42	0,0403	0,33	0,27
	Regresión Logística	fonemas, difonemas y trifonemas	7,70	0,0408	0,33	0,27
TCLU Formantes	**Suma promedio**	**fonemas, difonemas**	**10,75**	**0,0671**	0,9	0,41
	Suma promedio	fonemas, difonemas y trifonemas	12,23	0,0662	0,55	0,46
	Regresión logística	fonemas, difonemas	11,80	0,0616	0,47	0,41
	Regresión logística	**fonemas, difonemas y trifonemas**	**11,42**	**0,0620**	0,47	0,41

Tabla 46. EER (%), minDCF, Cllr y minCllr para diferentes fusiones a nivel inter unidad y tipos de características: TCLU-MFCC y TCLU-Formantes.

De acuerdo con la Tabla 46, para el caso de los TCLU-MFCC, el mejor rendimiento, en términos de EER, se obtiene mediante la fusión suma promedio de las unidades de fonemas y difonemas. De la misma forma, para el caso de los TCLU-Formantes, el mejor rendimiento se obtiene con la fusión suma promedio mediante la fusión de unidades de fonemas y difonemas. En ambos casos, si incorporamos al grupo de fusión los trifonemas, el rendimiento empieza a degradarse. Una posible razón de esta degradación se relaciona con mal rendimiento que se obtiene con las unidades de trifonemas, tanto para TCLU-MFCC y TCLU-Formantes. Cabe recordar, que los mejores rendimientos para fonemas y difonemas son similares. (Ver Tabla 37 y Tabla 43).

Por otra parte, se observa que el rendimiento óptimo alcanzado mediante la fusión inter-unidad supera al rendimiento obtenido por fusión intra-unidad. Una posible razón se debe a la distinta naturaleza de cada tipo unidad que se fusiona, permitiendo alcanzar valores óptimos de EER, sin embargo, a medida que mejora el valor de EER, la pérdida de calibración empeora.

Respecto al sistema de referencia, en términos de EER, el rendimiento de TCLU-MFCC (EER = 6,89 %) es aproximadamente un 40 % superior a éste (EER = 10.21 %). Para el caso de TCLU-Formantes (EER = 10,75 %), su rendimiento es menor respecto al sistema de referencia, aproximadamente un 5 %.

Adicionalmente, se ha realizado fusiones inter-unidad pero al mismo nivel de unidad, es decir, fusión de distintos tipos de características pero sobre la misma unidad. En las siguientes figuras se presentan las curvas DET de las fusiones realizadas.

FONEMA

- TCLU-MFCC + MFCC

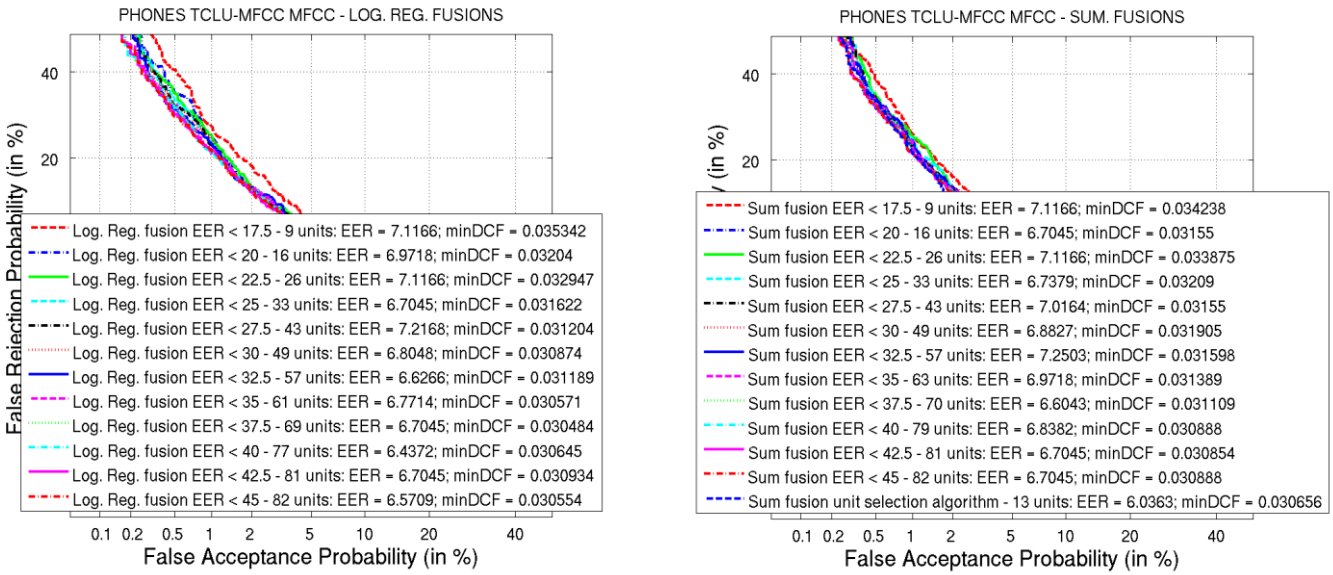

Figura 53. Curvas DET de fusiones regresión logística y suma promedio a nivel de fonema (TCLU-Formantes y TCLU-MFCC).

- TCLU-MFCC + TCLU-Formantes

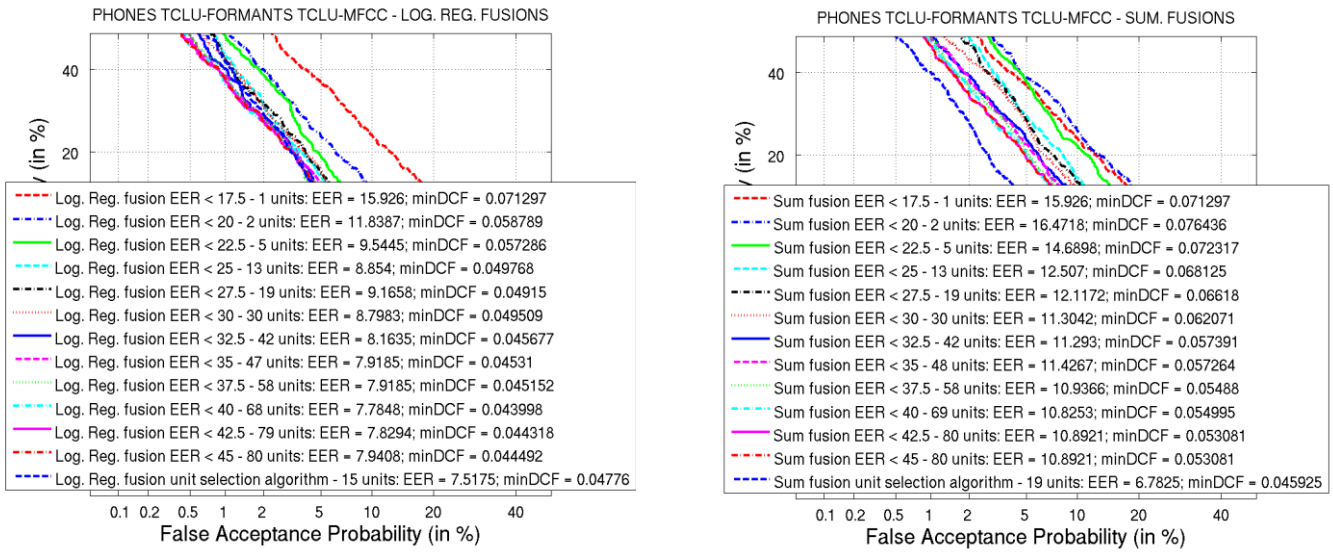

Figura 54. Curvas DET de fusiones regresión logística y suma promedio a nivel de fonema (TCLU-Formantes y TCLU-MFCC).

DIFONEMA

- TCLU-MFCC + MFCC

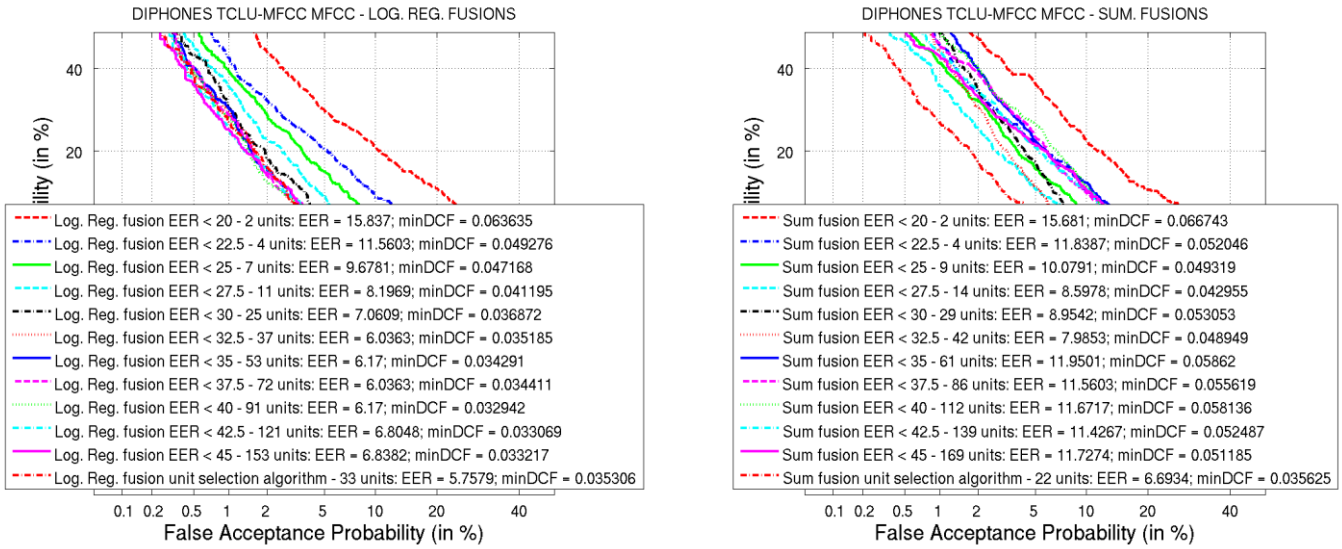

Figura 55. Curvas DET de fusiones regresión logística y suma promedio a nivel de difonema (TCLU-MFCC y MFCC).

- TCLU-MFCC + TCLU-Formantes

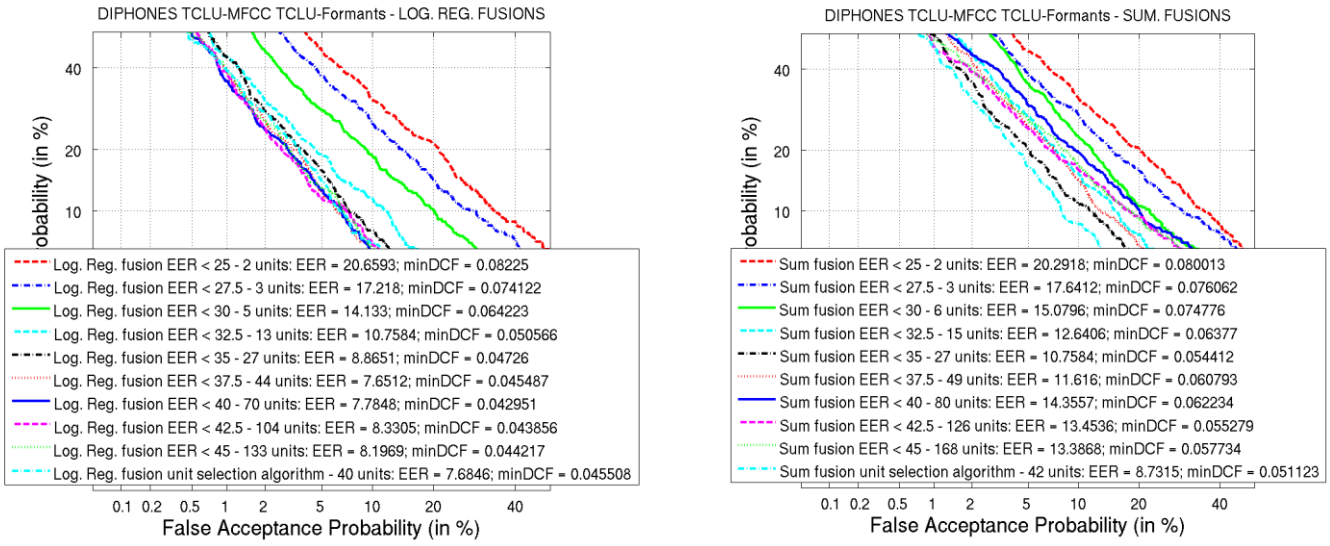

Figura 56. Curvas DET de fusiones regresión logística y suma promedio a nivel de difonema (TCLU-Formantes y TCLU-MFCC).

TRIFONEMA

- TCLU-MFCC + MFCC

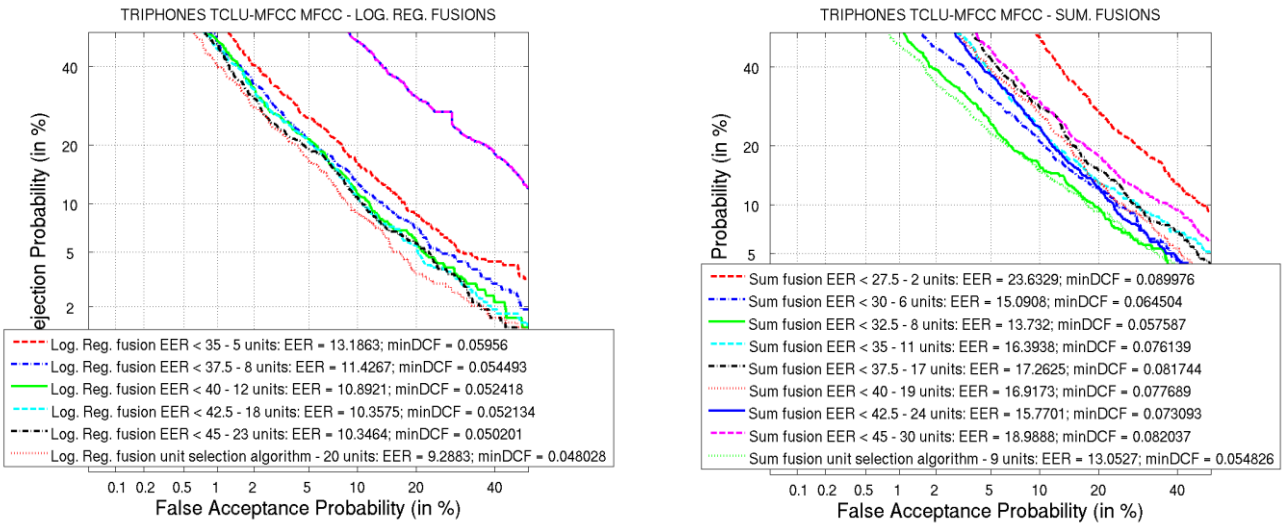

Figura 57. Curvas DET de fusiones regresión logística y suma promedio a nivel de trifonema (TCLU-MFCC y MFCC).

- TCLU-MFCC + TCLU-Formantes

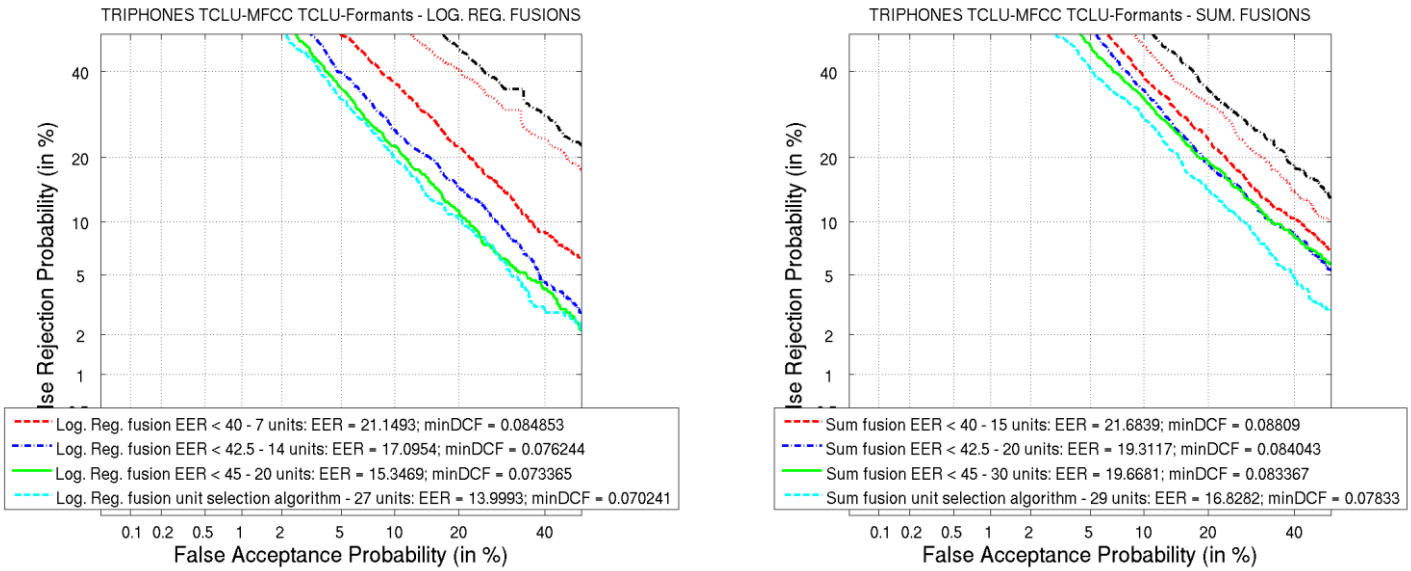

Figura 58. Curvas DET de fusiones regresión logística y suma promedio a nivel de trifonema (TCLU-Formantes y TCLU-MFCC).

En la siguiente tabla se presenta la fusión óptima para contornos de coeficientes cepstrales (TCLU-MFCC) y formantes (TCLU-Formantes) por unidad lingüística.

Unidad	Tipo de fusión	Características fusionadas	ERR (%)	minDCF	C_{llr}	$minC_{llr}$
Fonema	**Suma promedio**	**TCLU-MFCC y MFCC**	**6,03**	**0,0306**	0,37	0,23
	Regresión logística	TCLU-MFCC y MFCC	6,43	0,0306	0,30	0,23
	Suma promedio	**TCLU-MFCC y TCLU-Formantes**	**6,78**	**0,0459**	0,76	0,27
	Regresión logística	TCLU-MFCC y TCLU-Formantes	7,51	0,0477	0,36	0,28
Difonema	Suma promedio	TCLU-MFCC y MFCC	6,69	0,0356	0,94	0,25
	Regresión logística	**TCLU-MFCC y MFCC**	**5,75**	**0,0353**	0,31	0,24
	Suma promedio	TCLU-MFCC y TCLU-Formantes	8,73	0,0511	0,88	0,32
	Regresión logística	**TCLU-MFCC y TCLU-Formantes**	**7,65**	**0,0454**	0,37	0,28
Trifonema	Suma promedio	TCLU-MFCC y MFCC	13,05	0,0548	0,94	0,43
	Regresión logística	**TCLU-MFCC y MFCC**	**9,28**	**0,0480**	0,39	0,34
	Suma promedio	TCLU-MFCC y TCLU-Formantes	16,82	0,0783	0,67	0,55
	Regresión logística	**TCLU-MFCC y TCLU-Formantes**	**14**	**0,0702**	0,51	0,47

Tabla 47. EER (%), minDCF, Cllr y minCllr para diferentes fusiones a nivel inter unidad y tipos de características: TCLU-MFCC y TCLU-Formantes.

Respecto a la Tabla 47, se puede observar la contribución de cada tipo de característica. En los tres casos el rendimiento conseguido mediante la fusión supera al sistema de referencia. En el caso de los fonemas, los mejores rendimientos que se obtienen para cada tipo de combinación de característica son similares. Comparado con el sistema de referencia, es aproximadamente, en términos de EER, un 60 % mejor. Sin embargo, las pérdidas de calibración que se presentan son muy altas.

En el caso de los difonemas, los mejores rendimientos obtenidos para ambos casos son similares y son 30 % mejores que el sistema de referencia en términos de EER, pero, al igual que en el caso de los difonemas, la pérdida de calibración sigue siendo relativamente alta.

Finalmente, para los trifonema, los mejores rendimientos conseguidos son buenos, pero mucho menores que el resto de unidades, sin embargo, solo para la fusión TCLU-MFCC y MFCC supera en un 10 % al sistema de referencia; para el caso de TCLU-MFCC y TCLU-Formantes, el valor de EER es casi un 4 % inferior al de referencia.

Comparando este último respecto al rendimiento de unidades trifonema TCLU-Formantes (Tabla 42) es aproximadamente un 30 % superior. Respecto a las pérdidas de calibración, éstas son mejores respecto al resto de unidades. Esto nos permite obtener menos tasas de error relacionadas al apoyo de hipótesis incorrecta de los LR.

En conclusión, mediante esta prueba de fusiones a nivel inter-unidad se puede apreciar la contribución de las unidades que se fusionan, en muchos casos, la aportación de cada tipo de unidad o característica es más que otra, como por ejemplo, la fusión inter-unidad (fonemas, difonemas y trifonemas) para TCLU-Formantes (Tabla 46) , donde el rendimiento conseguido es inferior respecto al conseguido por fonemas o difonema pero es superior respecto al conseguido usando sólo trifonemas.

5. Conclusiones y trabajo futuro

En el presente proyecto se ha centrado en el desarrollo y análisis de los sistemas de reconocimiento automático de locutor basados en el modelado GMM-UBM, usando como parámetros de características los contornos temporales de coeficientes cepstrales y formantes sobre unidades lingüísticas. En base a los resultados presentados, y siendo uno de los principales objetivos del proyecto, se puede obtener valores de verosimilitud (LR) calibrados por unidad lingüística, lo cual se acerca mucho al contexto forense, pero con la ventaja que se realiza de forma automática usando los mismos parámetros que usan los lingüistas y fonetistas. Esto nos permite trabajar en entornos no controlados, donde existen sólo segmentos cortos de habla para ser analizados, llegando extraer una posible identidad a partir de dicho segmento.

Para llegar al desarrollo completo del sistema automático de reconocimiento de locutor, se ha evaluado el sistema usando dos tipos de modelado GMM-UBM (3.3), como también, diferentes tipos de características y tipos de unidades lingüísticas. En función de estos se ha procedido a mejorar el sistema mediante técnicas de fusión y normalización de puntuaciones. Las pruebas realizadas se dividen en dos bloques relacionadas a los tipos de modelado de GMM-UBM: GMM-UBM Global y GMM-UBM Constrained.

Dos puntos concluyentes en común para ambas pruebas se relacionan con el tipo de parametrización del contorno temporal y la compensación de variabilidad:

- Respecto al tipo de parametrización de contornos temporales, el mejor rendimiento en el sistema se obtiene mediante la parametrización de los contornos temporales tipo 1, es decir, 19 primeros coeficientes estáticos para el caso de los coeficientes cepstrales (MFCC) y 3 primeras frecuencias de los formantes (F1, F2, F3). En base a las pruebas realizadas, se comprueba que la adición de más información a los coeficientes-cepstrales/frecuencias-formantes, como por ejemplo, 19 coeficientes estáticos más sus respectivas derivadas o 3 primeras frecuencias de formantes más sus respectivos anchos de banda, influye de manera negativa en rendimiento del sistema mediante la parametrización de sus contornos temporales. Caso contrario a la parametrización MFCC y formantes, donde los coeficientes o formantes del tipo 2 (19 coeficientes estáticos más sus respectivas derivadas o 3 primeras frecuencias de formantes más sus respectivos anchos de banda), sí ofrecen un mejor rendimiento al sistema.

- Respecto a la compensación de variabilidad de sesión, la aplicación sobre el dominio de coeficientes cepstrales produce una distorsión en la parametrización de sus contornos temporales, el cual provoca un peor rendimiento en el sistema comparado a las características sin compensación de variabilidad.

5.1 GMM-UBM Global

El fin de estas pruebas fue analizar la influencia de las trayectorias temporales en el sistema global, donde, tanto el modelo de UBM como el de locutor no son dependientes de las unidades. Cabe recordar que las locuciones sí están segmentadas en unidades. De esta forma, se pretende medir el rendimiento del sistema global usando este tipo de parametrización. Además, mediante normalización y fusión inter-unidad se ha logrado obtener mejores rendimientos.

Para este tipo de modelado, el mejor rendimiento se obtiene con las locuciones segmentadas en unidades de fonemas, y el número de mezclas óptimo es igual a 1024, obteniéndose un valor de EER = 10,62 y un minDCF = 0,0430.

Posteriormente, se aplicó tres tipos normalización a las puntuaciones: T-Norm, Z-Norm y ZT-Norm. En base a los resultados obtenidos se confirma la mejora en el rendimiento del sistema tras la normalización de puntuaciones, con un valor óptimo de EER = 8,50 % y minDCF = 0,0372, y obtenido mediante ZT-Norm. Este resultado es un 16,71 % superior respecto al sistema de referencia en términos de EER.

Adicionalmente, se realizó las pruebas de fusión a nivel inter-unidad, mediante las técnicas de suma promedio y regresión logística. Se fusionaron las mejores configuraciones por tipo unidad del sistema GMM-UBM Global: fonema, difonema y trifonema. El resultado obtenido muestra un valor de EER = 8,86 % y minDCF = 0,0387. Este valor de EER es un 13,22 % superior respecto al sistema de referencia.

5.2 GMM-UBM Constrained

En esta parte se ha analizado la influencia de los contornos temporales en unidades lingüísticas sobre el rendimiento del sistema, donde los modelos de UBM y de locutor son dependientes de estas unidades. A partir de estas pruebas se ha obtenido resultados que confirman el buen rendimiento de este nuevo enfoque, el cual nos permite aproximarnos mucho más al contexto forense.

Comparando los tres tipos de unidades, el mejor rendimiento se obtiene con los contornos temporales de unidades de fonemas, cuyo rendimiento por unidad son inferiores respecto al sistema de referencia, siendo en mucho de los casos un 60 % inferior, pero permiten obtener valores de LR calibrados dado a la poca pérdida de calibración que presentan. Esto confirma el buen grado de discriminación que presenta los contornos temporales de los fonemas respecto al resto de las unidades.

Adicionalmente, se ha realizado pruebas para determinar la influencia del tipo de parametrización de los contornos temporales. Para ello, se realizó la parametrización de los contornos mediante la transformada discreta del coseno (DCT) de orden 5, 7 y 9. El mejor rendimiento, en ambas parametrizaciones (TCLU-MFCC y TCLU-Formantes), se obtuvo con la DCT de grado 5.

Posteriormente se realizó las fusiones a nivel intra-unidad mediante dos técnica: suma promedio y regresión logística. En base a los resultados, el mejor rendimiento óptimo se obtiene para las unidades de fonemas, con un valor EER = 7,11 y minDCF = 0,0420 para la parametrización TCLU-MFCC, y un valor de EER = 12,27 % y minDCF = 0,0663 para la parametrización TCLU-Formantes. Esto confirma la buena aportación al rendimiento del sistema la fusión intra-unidad, el cual permite superar o acercarse al rendimiento del sistema de referencia. Además, dada a la poca pérdida de calibración, se puede obtener valores de LR calibrados por unidad evaluada, siendo una de las grandes ventajas de este nuevo enfoque.

Finalmente, se realizó pruebas del sistema con la fusión a nivel inter-unidad, es decir fusión entre distintos tipos de unidades (fonema, difonemas y trifonema) y diferentes tipos de parametrizaciones (TCLU-MFCC, MFCC y TCLU-Formantes). En base a los resultados obtenidos, se confirma la gran aportación que hace al rendimiento del sistema este tipo de fusión, llegando, en el mejor de los casos, a un rendimiento 50 % superior respecto al sistema de referencia. Esto demuestra que mediante la fusión de distintos tipos de niveles de información se obtiene un sistema mucho más robusto frente al sistema de referencia o frente al sistema por unidad.

5.3 Trabajo futuro

El nuevo enfoque de la parametrización TCLU abre una nueva línea de investigación en el reconocimiento automático de locutor. Recientes estudios y publicaciones científicas confirman su efectividad en el reconocimiento, sin embargo, dado al gran espacio de condiciones y diferentes entornos que se presentan en el reconocimiento de locutor, se presentan nueva líneas futuras que continúen con el trabajo propuesto:

- Estudiar nuevos métodos de parametrización de los contornos temporales. En base a los resultados y estudios anteriores, la parametrización mediante la DCT de orden 5 es la que mejor resultado otorga al sistema, sin embargo, cuando se parametriza las características con más información (información estática más dinámica en caso de MFCC o frecuencias de formantes más sus respectivos anchos de banda) el rendimiento empeora. Por tanto, se plantea como trabajo futuro la búsqueda de nuevos métodos de parametrización que aprovechen mejor mucho mejor este tipo de información.

- Respecto a los formantes, se plantea realizar las mismas pruebas pero con una base de datos manual; es decir, donde las frecuencias de formantes y anchos de banda han sido extraídas manualmente. Para ello, se propone la base de datos VTR (Vocal Tract Resonance) [Li Deng et al., 2006]. Esta base de datos contiene las trayectorias de las tres primeras frecuencias de formantes, las cuales han sido extraídas manualmente. Esta base es un subconjunto representativo de la base de datos TIMIT [Garafolo et al., 1990]. Mediante el análisis de estas pruebas se pretende evaluar y comparar los rendimientos de los sistemas, obtenidos a partir de las distintas bases de datos: automática (NIST SRE 2006) y manual (VTR-TIMIT).

- Extender las pruebas realizadas a unidades lingüísticas de mayor longitud: silabas, palabras y centros de fonemas en trifonemas (extrae el contorno temporal sólo del fonema central en un trifonema).

- Realizar pruebas usando como técnica de reconocimiento las máquinas de soporte (SVM) y los contornos temporales de unidades lingüísticas.

- Mediante el modelado GMM-UBM, crear supervectores para implementar sistemas de reconocimiento basado en Factor Analysis. De esta forma se pretende llegar a sistema mucho más robustos frente a la variabilidad de inter-sesión.

6. Referencias bibliográficas

[Aitken *et al*, 2004]
Aitken, C.G.G., Taroni, F.: "*Statistics and the Evaluation of Evidencie for Forensic Scientist*". John Wiley & Sons, Chichester, 2004.

[Arpabet]
Código de transcripción fonética desarrollado por ARPA y descrito en Wikipedia.
http://en.wikipedia.org/wiki/Arpabet

[Atal, 1974]
B. Atal. "*Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification*." *J. Acoust. Soc. Amer.* 55 (6), 1304-1312, 1974.

[Auckenthaler *et al*., 2000]
R. Auckenthaler, M. Carey and H. Lloyd-Thomas. "*Score normalization for text-independent speaker verification systems.*" *Digital Signal Processing*, V10, pp. 42-54, 2000.

[Bimbot *et al*., 2004]
F.Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Margin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds. A tutorial on text-independent speaker verification. *Journal on Applied Signal Processing*, 2004(4):430-451, 2004.

[Brümmer and Preez, 2006]
N. Brümmer and J. du Preez. "*Application-Independent Evaluation of Speaker Detection*". Computer Speech & Language, 20(2-3):230–275, 2006.

[Brümmer *et al.*, 2007]
N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Scwartz, and A. Strasheim. Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. IEEE Transactions on Audio, Speech and Signal Processing, 15(7):2072–2084, 2007.

[Burget *et al*., 2007]
L. Burget, P. Matejka, O. Glembek and J. Cernocky. "*Analysis of feature extraction and channel compensation in a GMM speaker recognition system.*" *IEEE Trans. Audio, Speech Language Process*. 15 (7), 1979-1986, 2007.

[Burton, 1987]
D. Burton. "*Text-independent speaker verificaction using vector quantization source coding.*" *IEEE Trans. Acoustics, Speech, Signal Process.* 35 (2), 133-143, 1987.

[Campbell *et al*. 2004]
J. P. Campbell, H. Nkasone, C. Cieri, D. Miller, K. Walker, A. F. Martin and M. A. Przybocki. "*The MMSR bilingual and crosschannel corporal for speaker recognition research and evaluation.*" *Proc. of Oddyssey*, pp. 29.32, 2004.

[Campbell *et al*, 2005]
Campbell, W.M., Reynolds, D.A., Campbell, J.P., Brady, K.J.: "*Estimating and evaluating confidence for forensic speaker recognition*". In: Proc. Of ICASSP, pp. 717-720, 2005.

[Campbell *et al*., 2006a]
W. Campbell, J. Campbell, D. Reynolds, E. Singer, P. Torres-Carrasquillo. "*Support vector machines for speaker and language recognition.*" Comput. Speech Lang. 20 (2-3), 210-229, 2006.

[Campbell *et al*., 2006b]
W. Campbell, D. Sturim and D. Reynolds. "*Support vector machines using GMM supervectors for speaker verification.*" *IEEE Signal Process. Lett.* 13 (5), 3008-311, 2006.

[Castro *et al*, 2009]
Castro, A., Ramos, D. and González-Rodriguéz, J., "*Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking*", Proc Interspeech 2009, Bringhton, UK, 2009, pp. 2343-2346.

[Carr, 1999]
P. Carr. English Phonetics and Phonology: An Introduction. *Blackwell Publishing*, 1999.

[Cortes and Vapnik, 1995]
C. Cortes and V. Vapnik. Support-vector networks. Machine Learning, 20(3):273–297, 1995.

[Dehak *et al*, 2011]
Dehak, N., et al., "*Front-End Factor Analysis for Speaker Verification*", IEEE Trans. on Audio, Speech, and Lang. Proc., 19(4), 788-798, May 2011.

[Davis and Mermelstein, 1980]
S. Davis and P. Mermelstein. "*Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*." *IEEE Trans. Acoustics, Speech, Signal Process.* 28 (4), 357-366, 1980.

[Fierrez-Aguilar *et al*., 2003]
J. Fierrez-Aguilar, J. Ortega-García, D. García-Romero y J. González-Rodríguez. "*A comparative evaluation of fusion strategies for multimodal biometric verification.*" *Proc. 4th IAPR Intl. Conf. on Audio and Video Based Person Authentication AVBPA*, pp. 830-837, Junio 2003.

[Fierrez-Aguilar *et al*., 2005]
J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. "*Target dependent score normalization techniques and their application to signature verification.*" *IEEE Trans. on Systems, Man and Cybernetics, part C*, 35(3):418:425, 2005.

[Franco-Pedroso *et al*, 2012]
Franco-Pedroso, J., Espinoza-Cuadros, F., González-Rodríguez, J.: "*Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition*", IberSpeech, 2012.

[Furui, 1981]
S. Furui. "*Cepstral analysis technique for automatic speaker verification.*" *IEEE Trans. Acoustics, Speech Signal Process*. 29 (2), 254-272, 1981.

[Garafolo *et al.*, 1990]
J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, NTIS Order No. PB91-505065: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM (National Institute of Standards and Technology, 1990).

[Gersho and Gray, 1991]
A. Gersho and R. Gray. "*Vector Quantization and Signal Compression.*" *Kluwer Academic Publishers*, Boston, 1991.

[González-Rodríguez, 2011]
González-Rodríguez, J., "*Speaker recognition using temporal contours in linguistic units: the case of formant and formant-bandwidth*", Interspeech 2011, pp. 133-136, Florence, Italy, 2011.

[González-Rodríguez *et al.*, 2012]
González-Rodríguez, J., González-Dominguez, J., Franco-Pedroso, J., Ramos, D., "*A linguistically-motivated speaker recognition front-end through session variability compensated cepstral trajectories in pone units*", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[Hermanski and Morgan, 1994]
H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Trans. Speech Audio Process.* 2 (4), 578-589.

[Huang *et al*., 2001]
X. Huang, A. Acero and H.-W. Hon. "*Spoken Language Processing: a Guide to Theory, Algorithm, and System Development.*" *Prentice-Hall*, New Jersey, 2001.

[Kajarekar *et al,* 2009]
Kajarekar; S. S. et al., "*The SRI NIST 2008 Speaker Recognition Evaluation System.*" Proc. IEEE ICASSP' 09, pp. 4205-4209, Taipei, 2009.

[Kenny, 2006]
P. Kenny. "*Joint factor analysis of speaker and session variability: theory and algorithms*". *Technical Report* CRIM-06/08-14, 2006.

[Kenny *et al.*, 2008]
P. Kenny, P. Ouellet, N. Dehak, V. Grupta and P. Dumouchel. "*A study of inter-speaker variability in speaker verification*". *IEEE Trans. Audio, Speech Language Process.* 16 (5), 980-988, 2008.

[Kinnunen *et al.,* 2006]
Kinnunen, T., Karpov, E., Fränti, P., 2006. "*Real-time speaker identification and verification.*" IEEE Trans. Audio, Speech Language Process. 14 (1), 277–288.

[Kinnunen and Li, 2010]
Kinnunen, T., and Li, H., "*An overview of text-independent speaker recognition: from features to supervectors*", Speech Communication, vol. 52, pp. 12-40, 2010.

[Krause and Gazit, 2006]
N. Krause y R. Gazit. "*SVM-based speaker classification in the GMM model space.*" *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, pp. 1-5, 2006.

[Leeuwen and Brümmer, 2007]
D. A. Leeuwen and N. Brümmer. "*An introduction to application-independent evaluation of speaker recognition systems.*" In C. Müller, editor, Speaker Classification I, pages 330–353. Springer-Verlag, Berlin, Heidelberg, 2007.

[Linde *et al*., 1980]
Y. Linde, A. Buzo, R. Gray. "*An algorithm for vector quantizer design.*" *IEEE Trans. Comm.* 28 (1), 84-95.

[Li Deng *et al*., 2006]
Li Deng, C. Xiaodong, R. Pruvenok, J. Huang, S. Momem, Y. Chen and A. Alwan. "*A database of vocal tract resonance trajectories for research in speech processing*". *Proc. ICASSP*, 2006.

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

[López-Moreno *et al*., 2008]
I. Lopez-Moreno, D. Ramos, J. Gonzalez-Rodriguez, and D. T. Toledano. "*Anchor Model Fusion for LanguageRecognition.*" In Proceedings of Interspeech 2008, September 2008.

[López *et al*., 2003]
E. López, G. Sosa y M. Rocamora: "*Tratamiento de Voz*". Disponible en: http://iie.fing.edu.uy/investigacion/grupos/gmm/audio/seminario/seminariosviejos/2003/charlas/charla1/voz8.htm

[Louradour and Daoudi, 2005]
Louradour, J., Daoudi, K., 2005. "*SVM speaker verification using a new sequence kernel.*" In: Proc. 13th European Conf. on Signal Processing (EUSIPCO 2005), Antalya, Turkey, September2005.

[Malayath *et al*., 2000]
N. Malayath, H. Hermansky, S. Kajarekar and B. Yegnanarayana. "*Data-driven temporal filters and alternatives to GMM in speaker verification.*" *Digital Signal Process*. 10 (1-3), 55-74, 2000.

[Maltoni *et al.*, 2003]
D. Maltoni, D. Maio, A. K. Jain and S. Prabhakar. Handbook of Fingerprint Recognition, *Springer* 2003.

[Morrinson, 2009]
Morrinson, G. S., "*Likelihood-ratio-based forensic speaker comparison using parametric representation of vowel formant trajectories*", J. of the Ac. Soc. Of Am., 125, 2387-2397, 2009.

[NIST SRE]
Página web de las evaluaciones NIST de reconocimiento de locutor: http://www.nist.gov/itl/iad/mig/sre.cfm

[Oppenheim *et al*., 1999]
A. Oppenheim, R. Schafer and J. Buck. Discrete-Time Signal Processing, segunda edición. *Prentice-Hall*, 1999.

[Ortega-García, 2010]
J. Ortega-García: Ampliación de señales aleatorias. Disponible en: http://arantxa.ii.uam.es/~jortega/VQ_AlgoritmiaRecVoz_ASAL.pdf

[Pelecanos y Sridharan, 2001]
J. Pelecanos y S. Sridharan. "*Feature warping for robust speaker verification.*" *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001)*, Creta, Grecia, Junio de 2001, pp. 213-218.

[Ramos-Castro, 2007]
D. Ramos Castro. "*Forensic evaluation of the evidence using automatic speaker recognition systems*". Tesis doctoral. Escuela Politécnica Superior, Universidad Autónoma de Madrid. Noviembre de 2007.

[Reynolds and Rose, 1995]
D. Reynolds and R. Rose. "*Robust text-independent speaker identification using Gaussian mixture speaker models.*" *IEEE Trans. Speech Audio Process.* 3, 72-83, 1995.

[Reynolds *et al*., 2000]
D. A. Reynolds, T. F. Queatieri and R. B. Dunn: "*Speaker Verification Using Adapted Gaussian Mixture Models"*. Digital Signal Processing 10, 19-41 (2000).

[Reynolds *et al*., 2003]
Douglas Reynolds, Walter Andrews, Joseph Campell, Jiri Navratil, Barbara Peskin, Andre Adam, Qin Jin, David Klusacek, Joy Abramson, Radu Mihaescu, Jack Godfrey, Doug Jones, Bing Xiang. SuperSID Project: "*Exploiting high-level information for high accuracy speaker recognition*". *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 4, pp 784-787, Abril 2003.

[Reynolds, 2003]
D. Reynolds. "*Channel robust speaker verification via feature mapping*." *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, Vol. 2, Hong Kong, China, Abril de 2003, pp. 53-56.

[Roch, 2006]
Roch, M., 2006. "*Gaussian-selection-based non-optimal search for speaker identification.*" Speech Commu. 48, 85–95.

[Rose, 2002]
P. Rose. "*Forensic Speaker Identification.*" *Taylor & Francis*, London, 2002.

[Saastamoinen *et al.*, 2005]
Saastamoinen, J., Karpov, E., Hautamäki, V., Fränti, P., 2005. "*Accuracy of MFCC based speaker recognition in series 60 device.*" EURASIP J.Appl. Signal Process. 17, 2816–2827.

[Shireberg, 2007]
Shireberg, E., "*Higher-level features in speaker recognition*", in Speaker Classification I: Fundamentals, Feature and Methods, C. Müller, Ed., number 4343 in Lecture notes in Artificial Intelligence, pp. 241-259, Springer, 2007.

[Sjolander and J. Beskow, 2000]
K. Sjolander and J. Beskow, "*Wavesurfer – an open source speech tool*". Proc. ICSLP 2000, Beijing, China, 2000.

[Soong *et al.*, 1987]
F. K. Soong, A. E. Rosenberg, B. -H. Juang and L. R. Rabiner. "*A vector quantization approach to speaker recognition*." *AT&T Technical J.* 66, 14-26. 1997.

[Soong and Rosenberg, 1988]
F. Soong and A. Rosenberg. "*On the use of instantaneous and transitional spectral information in speaker recognition.*" *IEEE Trans. Acoustics, Speech Signal Process*. 36 (6), 871-879, 1988.

[SRE 2006]
Página web del plan de evaluación de reconocimiento locutor del NIST de 2006: http://www.itl.nist.gov/iad/mig/tests/sre/2006/sre-06_evalplan-v9.pdf

[Vogt and Sridharan, 2008]
R. Vogt and S. Sridharan. "*Explicit modeling of session variability in text-independent speaker verification.*" *Comput. Speech Lang.* 22 (1), 17-38, 2008.

[Wolf, 1972]
J. Wolf. "*Efficient acoustic parameters for speaker recognition.*" *J. Acoust. Soc. Amer.* 51 (6), 2044-20556 (Part 2), 1972.

A. Presupuesto

1) Ejecución Material

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

2) Gastos generales

- 16 % sobre Ejecución Material..... 180 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material..... 70 €

4) Honorarios Proyecto

- 1500 horas a 15 € / hora..... 22500 €

5) Material fungible

- Gastos de impresión ..... 100 €
- Encuadernación ..... 200 €
- Total de material fungible.....300 €

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, Septiembre 2012

El Ingeniero Jefe de Proyecto

Fdo.: Fernando Manuel Espinoza Cuadros
Ingeniero Superior de Telecomunicación

B. Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un *SISTEMA DE IDENTIFICACIÓN DE HABLANTES A PARTIR DE TRAYECTORIAS TEMPORALES EN UNIDADES LINGÜÍSTICAS SOBRE GRANDES BASES DE DATOS*.

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% del presupuesto y la provisional del 2%.

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 Contrata” y 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. Apéndice

A continuación se incluye el artículo que se mencionó en el resumen del proyecto. Este artículo recoge parte de los resultados obtenidos en el presente proyecto.

IberSPEECH2012
IberSPEECH2012 "VII Jornadas en Tecnología del Habla" and III Iberian SLTech Workshop
21-23 November 2012, Madrid, Spain

Submission Summary

Track:	**REGULAR PAPERS**
Paper ID:	48
Title:	Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition
Abstract:	In this paper, the contributions of different linguistic units to the speaker recognition task are explored by means of temporal trajectories of their MFCC features. Inspired by successful work in forensic speaker identification, we extend the approach based on temporal contours of formant frequencies in linguistic units to design a fully automatic system that puts together both foren-sic and automatic speaker recognition worlds. The combination of MFCC fea-tures and unit-dependent trajectories provides a powerful tool to extract indi-vidualizing information. At a fine-grained level, we provide a calibrated likeli-hood ratio per linguistic unit under analysis (extremely useful in applications such as forensics), and at a coarse-grained level, we combine the individual contributions of the different units to obtain a highly discriminative single sys-tem. This approach has been tested with NIST SRE 2006 datasets and proto-cols, consisting of 9,720 trials from 219 male speakers for the 1side-1side Eng-lish-only task, and development data being extracted from 367 male speakers from 1,808 conversations from NIST SRE 2004 and 2005 datasets.
Created On:	6/25/2012 1:08:03 PM
Modified On:	7/2/2012 3:33:25 PM
Authors:	Javier Franco Pedroso , *javier.franco@uam.es* Fernando Espinoza , *fernando.espinoza@estudiante.uam.es* Joaquin González-Rodríguez , *joaquin.gonzalez@uam.es*
Primary Contact:	Javier Franco Pedroso , *javier.franco@uam.es*
Primary Subject Area:	1.b. Speech technology and applications: Speech and speaker recognition
Secondary Subject Areas:	
Files:	IberSpeech_2012_-_Javier_Franco-Pedroso.pdf (1,177,951 bytes uploaded on 7/2/2012 3:33:25 PM)

Cepstral Trajectories in Linguistic Units for Text-Independent Speaker Recognition

Javier Franco-Pedroso, Fernando Espinoza-Cuadros and Joaquin Gonzalez-Rodriguez

ATVS – Biometric Recognition Group
Universidad Autonoma de Madrid, Spain
javier.franco@uam.es

Abstract. In this paper, the contributions of different linguistic units to the speaker recognition task are explored by means of temporal trajectories of their MFCC features. Inspired by successful work in forensic speaker identification, we extend the approach based on temporal contours of formant frequencies in linguistic units to design a fully automatic system that puts together both forensic and automatic speaker recognition worlds. The combination of MFCC features and unit-dependent trajectories provides a powerful tool to extract individualizing information. At a fine-grained level, we provide a calibrated likelihood ratio per linguistic unit under analysis (extremely useful in applications such as forensics), and at a coarse-grained level, we combine the individual contributions of the different units to obtain a highly discriminative single system. This approach has been tested with NIST SRE 2006 datasets and protocols, consisting of 9,720 trials from 219 male speakers for the 1side-1side English-only task, and development data being extracted from 367 male speakers from 1,808 conversations from NIST SRE 2004 and 2005 datasets.

Keywords: automatic speaker recognition, forensic speaker identification, temporal contours, linguistic units, cepstral trajectories.

1 Introduction[1]

Automatic speaker recognition has focused in the last decade on two concurrent problems: the compensation of session variability effects, mainly through high-dimensional supervectors and latent variable analysis [2] [7] [8], and the production of an application-independent calibrated likelihood ratio per speaker recognition trial [1], able to elicit useful speaker identity information to the final user with any given application prior. The results are highly efficient text-independent systems in controlled conditions, as NIST SRE evaluations, where lots of data from hundreds of speakers in similar conditions are available. Thus, all the speech available in every trial is used to produce detection performances difficult to imagine a decade ago.

[1] Supported by MEC grant PR-2010-123, MICINN project TEC09-14179, ForBayes project CCG10-UAM/TIC-5792 and Catedra UAM-Telefonica. Thanks to ICSI (Berkeley, CA) for hosting the preliminary part of this work. Thanks to SRI for providing Decipher labels for SRE datasets.

However, in the presence of strong mismatch (as e.g. in forensic conditions, where acoustic and noise mismatch, apart from highly different emotional contexts, speaker roles or health/intoxication states can be present between the control and questioned speech), those acoustic/spectral systems could be unusable as all our knowledge about the two speech samples is deposited into a single likelihood ratio, obtained from all the available speech in the utterance, that could be strongly miscalibrated (being then highly misleading) as the system has been developed under severe database mismatch between training and testing data. Moreover, it is difficult (or even impossible) to collect enough data to develop a system robust to every combination of mismatch factors present in actual case data, an important problem in real applications.

A usual procedure in forensic laboratories is that a speech expert, typically a linguist/phonetician, can isolate or mark segments of compatible/comparable speech between both samples, segments being from seconds long to just some short phonetic events in given articulatory contexts. The number and types of comparable units for analysis is always a case-dependent subject, and therefore flexible strategies for analysis and combination are needed.

The proposed approach gives an answer to this application framework, providing informative calibrated likelihood ratios for every linguistic unit under analysis. Moreover, the combination of the different units yields good discrimination capabilities allowing to obtain speaker detection performance levels similar to equivalent acoustic/spectral systems when enough usable units are available.

The remainder of the paper is organized as follows. In Sections 2 and 3 we present, respectively, our proposed front-end for feature extraction over linguistic units and the system in use. Section 4 describes the databases and the experimental protocol used for testing the system. Section 5 shows results for the different linguistic units individually and for several combination methods, to finally conclude in Section 6 summarizing the main contributions and future extensions of this work.

2 Cepstral trajectories extraction from linguistic units

Many attempts have been made to incorporate the temporal dynamics of speech into features, from the simplest use of the velocity (delta) and acceleration (delta-delta) derivative coefficients to modulation spectrograms, frequency modulation features or even TDCT (temporal DCT) features (see [9] for a review). However, to the best of our knowledge none of the previous approaches, with the exception of SNERFs [4] and [12] for prosodic information, take advantage of the linguistic knowledge provided by an automatic speech recognizer (ASR) to extract non-uniform-length sequences of spectral vectors to be converted into constant-size feature vectors characterizing the spectro-temporal information in a given linguistic unit. In our proposed front-end, we obtain a constant-size feature vector from non-uniform-length MFCC features sequence within a phone unit.

2.1 ASR region conditioning

In this work, both phone and diphone units have been used for defining time intervals in order to extract the temporal contours over the MFCC features. For this purpose, the phonetic transcription labels produced by SRI's Decipher conversational telephone speech recognition system [6] were used first. For this system, trained on English data, the Word Error Rate (WER) of native and nonnative speakers on transcribed parts of the Mixer corpus, similar to NIST SRE databases used for this work, was 23.0% and 36.1% respectively. These labels define both phonetic content and time interval of speech regions containing the phone units to be segmented. For this work, 41 phone units from an English lexicon were used, represented by the Arpabet phonetic transcription code [13]. Diphone units are defined by the combination of any two consecutive phone units, although only a subset of 98 diphones of the possible combinations was used (those presenting higher frequency of occurrence).

2.2 Cepstral trajectories parameterization

By means of SRI's Decipher phone labels, trajectories (i.e., the temporal evolution of each MFCC vector dimension) of 19 static MFCC are extracted from phone and diphone units, yielding a MFCC matrix of 19 coefficients x #frames/unit for each linguistic unit. This variable-length segment is duration equalized to a number of frames equivalent to 250 ms. Finally, those trajectories are coded by means of a fifth order discrete cosine transform (DCT), yielding our final 19 x 5 fixed-dimension feature vector for each linguistic unit

3 System description

3.1 Unit-dependent acoustic systems

Proposed systems are based on the well known GMM-UBM framework [11], using duration-equalized DCT-coded MFCC trajectories per linguistic unit as feature vectors. The GMM-UBM systems have been the state-of-the-art in the text-independent speaker recognition field for many years until the emergence of JFA [7] and total variability [2] techniques, which have outperformed the former ones through accurately modeling the existing variability in the supervector feature space.

For this work, GMM-UBM systems have been chosen for two main reasons: i) as we are using a new type of features, we need first to find the optimal configuration for this GMM-UBM new-framework, which is the basis of supervector-based systems; and ii) because we aim to model speakers in a unit-dependent way, a much smaller amount of data is available for training purposes, so probably not enough data would be available to capture the existing variability in each unit domain (also having into account that we only have ASR labels from the SRE04, SRE05 and SRE06 datasets).
Three different unit-dependent GMM-UBM configurations were tested previously to perform experiments reported in this paper:

1. UBM and speaker models trained on unit-independent data; evaluation trials performed on unit-dependent test data (as we did in our first approach [5]).
2. UBM trained on unit-independent data; speaker models adapted from unit-dependent training data; evaluation trials performed on unit-dependent test data.
3. UBM and speaker models trained on unit-dependent data; evaluation trials performed on unit-dependent test data (fully unit-dependent).

For each configuration, different numbers of mixtures were tested, ranging from 2 up to 1024 mixtures increasing in powers of 2. It was found out that best results were obtained for the fully unit-dependent configuration, using 8 mixtures in the case of phone units and 4 mixtures in the case of diphone units. These configurations are those used to obtain the individual linguistic unit results reported in this paper.

3.2 Fusion schemes and linguistic units combinations

Both individual unit performance and different unit combinations have been analyzed in this paper. On the one hand, individual linguistic-unit systems allow us to report useful speaker verification LR's for very short speech samples where usual state-of-the-art systems are not directly applicable (as it is the case of forensic applications). On the other hand, when more data is available, individual units can be combined to achieve better discriminative capabilities.
In addiction to obtaining test results for each linguistic unit, these individual systems were combined in both intra- and inter-unit manners, i.e. fusing phone/diphone units between them and fusing phone and diphone units together. Two different fusion techniques were used: sum fusion and logistic regression fusion. The former one was performed after linear logistic regression calibration, while the latter one was performed in a single calibration/fusion step.
Another issue is what should be the selected units to be fused. Two strategies have been used in this work. The first of them is to select the n-best performing units by setting a threshold for the EER of the units to be fused, leaving out those performing worse. However, this procedure do not guaranty that the best fused system will be achieved because some units with lower performance by itself could contribute to the fused system if its LR's are sufficiently low correlated with those produced by the other units to be fused. On the other hand, testing all of the possible combinations would be a very complex task, so we used a unit selection algorithm (similar to that used in [3]) based on the following steps:

1. Take the best performing unit in terms of EER as the initial units set.
2. Take the next best performing unit and fuse with the previous set. If the fusion improves the performance of the previous set, this unit is added to the units set, otherwise rejected.
3. The previous step is repeated for all the units in increasing EER order.

This procedure allows us to find complementarities between units that otherwise would not have been revealed, but avoiding the complex task of testing each possible combination.

4 Datasets and experimental setup

NIST SRE datasets and protocols have been used to develop and test our proposed system, in particular those of years 2004, 2005 and 2006. As region conditioning for linguistic units definition and extraction rely on SRI's Decipher ASR system (trained on English data), English-only subsets of the NIST SRE datasets have been used. SRE 2004 and 2005 datasets were used as the background dataset for UBM training, consisting of 367 male speakers from 1,808 conversations (only male speakers were used for this work). English-only male 1side-1side task from SRE 2006 was used for testing purposes.
This dataset and evaluation protocol comprises both native and nonnative speakers across 9,720 same-sex different-telephone-number trials from 298 male speakers. SRE 2005 evaluation set was also used to obtain scores in order to train the calibration rule (linear logistic regression).
Performance evaluation metrics used are the Equal Error Rate (EER) and the Detection Cost Function (DCF) as defined in the NIST SRE 2006 evaluation plan [10]. Cllr and minCllr [1] (and its difference, calibration loss) are also used to evaluate the goodness of the different detectors after the calibration process.

5 Results

5.1 Reference system performance

As we are using the GMM-UBM framework to model unit-dependent systems, our baseline reference system is also a GMM-UBM system based on MFCC features. A classical configuration with 1024 mixtures and diagonal covariance matrices was used, and MFCC features include 19 static coefficients plus first order derivatives, cepstral mean normalization, RASTA filtering and feature warping. The performance of this system in the English-only male 1side-1side task from SRE 2006 is EER=10.26% and minDCF=0.0457. This system does not include any type of score normalization.

5.2 Phone units: individual and combined systems performances

Table 1 shows individual performance of phone units for the NIST SRE 2006 English-only male 1side-1side task. It can be seen that, although most of the phones have high EER and minDCF values, almost all of them are well calibrated (low difference between Cllr and minCllr). This allows us to obtain informative calibrated likelihood ratios from very short speech samples (as low as some phone units), as we can see in the tippet plot in Figure 1 for the best performing phone unit ('N'). Moreover, there are lots of units that can be combined, and despite their lower individual performance (around 60% worse than the reference system for the best performing phone), combined system can outperform reference system by means of sum or logistic regression fusion, as it can be seen in Figure 2. This is due to the highly complementarity of acoustic systems coming from different linguistic content.

Phone unit	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AA	32.20	0.0983	0.8633	0.8452
AE	18.98	0.0813	0.6087	0.5832
AH	29.39	0.0969	0.8235	0.7967
AO	34.36	0.0992	0.9065	0.8838
AW	36.99	0.0991	0.9241	0.9111
AX	27.08	0.0947	0.7882	0.7512
AY	21.68	0.0869	0.6822	0.6428
B	34.50	0.0986	0.8922	0.8778
CH	42.59	0.1000	0.9686	0.9538
D	32.07	0.0965	0.8661	0.8500
DH	28.43	0.0934	0.8403	0.7857
DX	40.44	0.0998	0.9670	0.9484
EH	31.69	0.0975	0.8574	0.8283
ER	35.18	0.0987	0.9107	0.8901
EY	26.40	0.0925	0.7713	0.7515
F	39.63	0.0993	0.9561	0.9397
G	35.71	0.1000	0.9291	0.9040
HH	39.80	0.0992	0.9527	0.9414
IH	26.95	0.0948	0.7964	0.7495
IY	23.32	0.0923	0.7453	0.7002
JH	39.69	0.0997	0.9487	0.9339
K	27.76	0.0961	0.8219	0.7832
L	26.51	0.0935	0.7789	0.7451
M	22.28	0.0857	0.6824	0.6583
N	15.92	0.0713	0.5520	0.5082
NG	29.37	0.0934	0.9977	0.7958
OW	24.65	0.0987	0.7917	0.7396
P	39.50	0.0988	0.9466	0.9335
PUH	24.18	0.0908	0.7359	0.7149
PUM	34.15	0.0953	0.8644	0.8419
R	24.65	0.0887	0.7295	0.7116
S	30.04	0.0973	0.8451	0.8059
SH	39.36	0.0996	1.0546	0.9294
T	27.89	0.0921	0.8256	0.7647
TH	38.37	0.1000	1.1207	0.9298
UH	41.53	0.1000	0.9717	0.9593
UW	24.79	0.0898	0.7391	0.7198
V	35.86	0.0990	0.9093	0.8932
W	35.82	0.0993	0.9167	0.8966
Y	24.00	0.0906	0.7313	0.7062
Z	32.07	0.0968	0.8487	0.8312

Table 1. EER (%), minDCF, Cllr and minCllr for phone units in the NIST SRE 2006 English-only male 1side-1side task.

It should be noted that results equivalent to that of the reference system can be achieved by combining only 4 phone units (‘AE’, ‘AY’, ‘M’, ‘N’). Also, it can be seen that the unit selection algorithm used can achieve better fusion results than simply setting a threshold for the EER of the units to be fused, both for sum and logistic regression fusions. Furthermore, it is worth noting that some of the phone units selected to be fused have very low performance (‘CH’ in the sum fusion, ‘AO’ in both sum and logistic regression fusions).

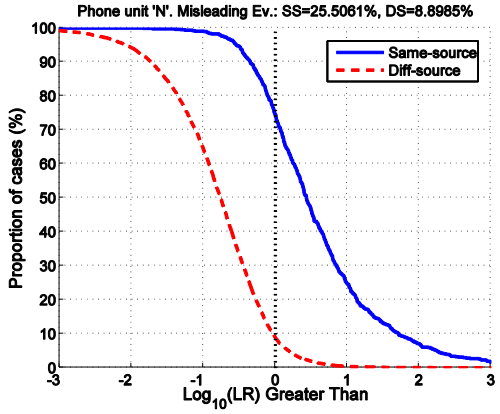

Fig. 1. Tippet plot for the best performing phone unit ('N') in the NIST SRE 2006 English-only male 1side-1side task.

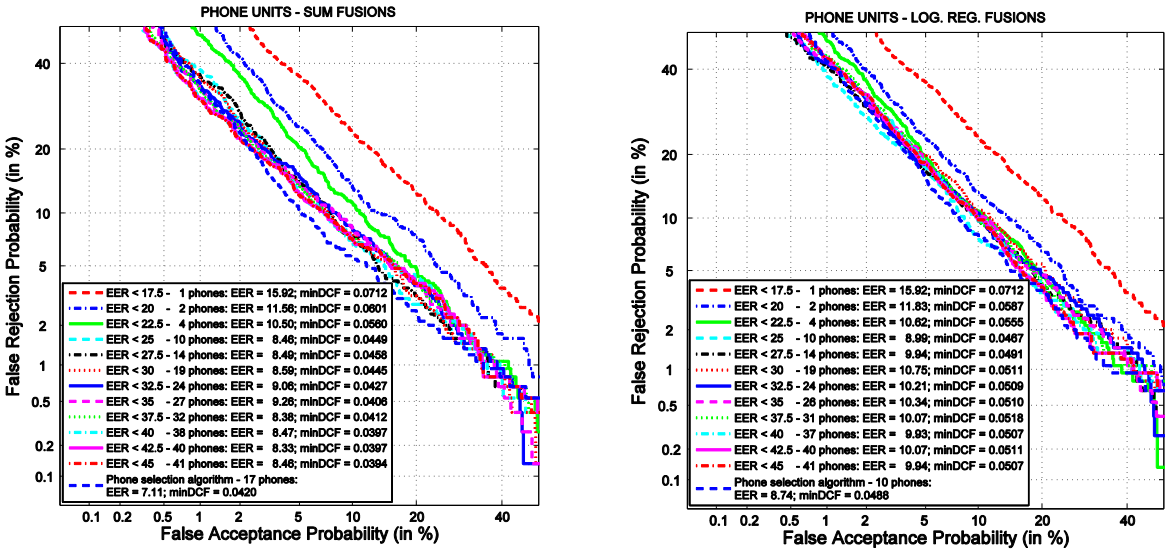

Fig. 2. DET curves for sum and log. reg. fused systems in the NIST SRE 2006 English-only male 1side-1side task for different phone selection schemes.

5.3 Diphone units: individual and combined systems performances

Table 2 shows individual performance for the ten best performing diphone units for the NIST SRE 2006 English-only male 1side-1side task. As it can be seen, diphone units have much lower performance than phone units. This may be due to the fact that, while diphones cover a longer time span that can present more complex trajectories, we are still using a 5 order DCT to code these trajectories. However, as it can be seen in Figures 3, diphone fusions can achieve as good performance as the phones unit fusions, although more units are needed to be fused.

Diphone unit	EER (%)	minDCF	C_{llr}	$minC_{llr}$
AEN	30.72	0.0993	0.8479	0.823
AET	31.89	0.0969	0.872	0.8526
AXN	23.84	0.0899	0.7583	0.7097
AYK	32.45	0.0970	0.8494	0.8356
LAY	29.11	0.0972	0.8156	0.7955
ND	24.92	0.0876	0.7563	0.7037
NOW	30.86	0.0995	0.8455	0.8185
UWN	32.20	0.0953	0.8417	0.8188
YAE	29.78	0.0976	0.8383	0.8094
YUW	27.18	0.0960	0.8223	0.7812

Table 2. . EER (%), minDCF, Cllr and minCllr for the 10 best performing diphone units in the NIST SRE 2006 English-only male 1side-1side task.

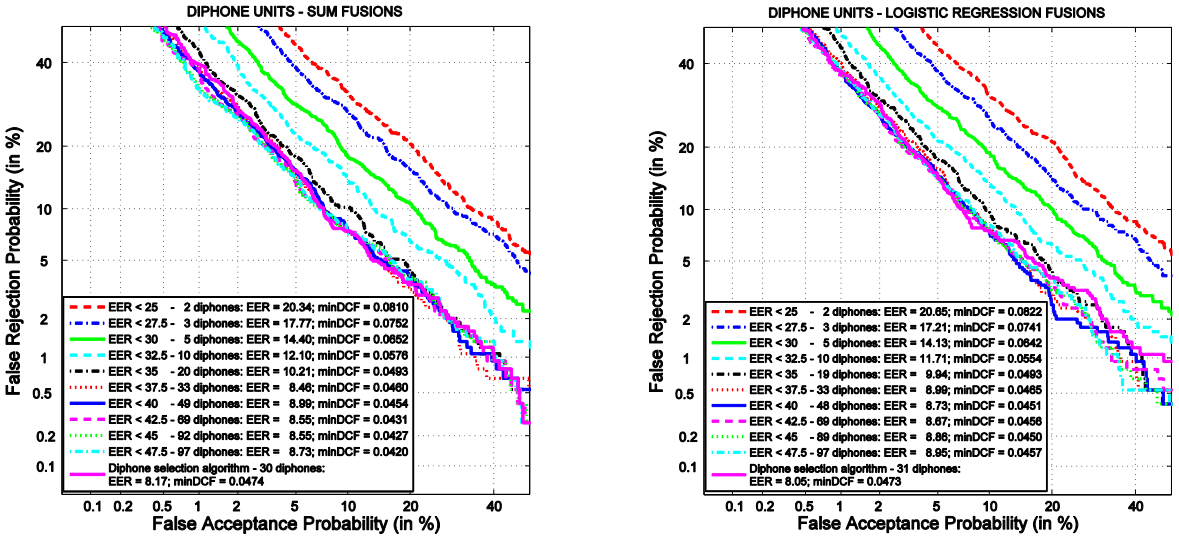

Fig. 3. DET curves for sum and log. reg. fused systems in the NIST SRE 2006 English-only male 1side-1side task for different diphone selection schemes.

5.4 Inter-unit combined system performance

In the previous paragraphs we have seen how well combine different units from each type (i.e., different phones between them and different diphones between them), but it is also interesting to see how can be combined units from different types between them. For this purpose, same fusion techniques and combination schemes have been used putting together both phones and diphones, yielding results show in Figure 4.

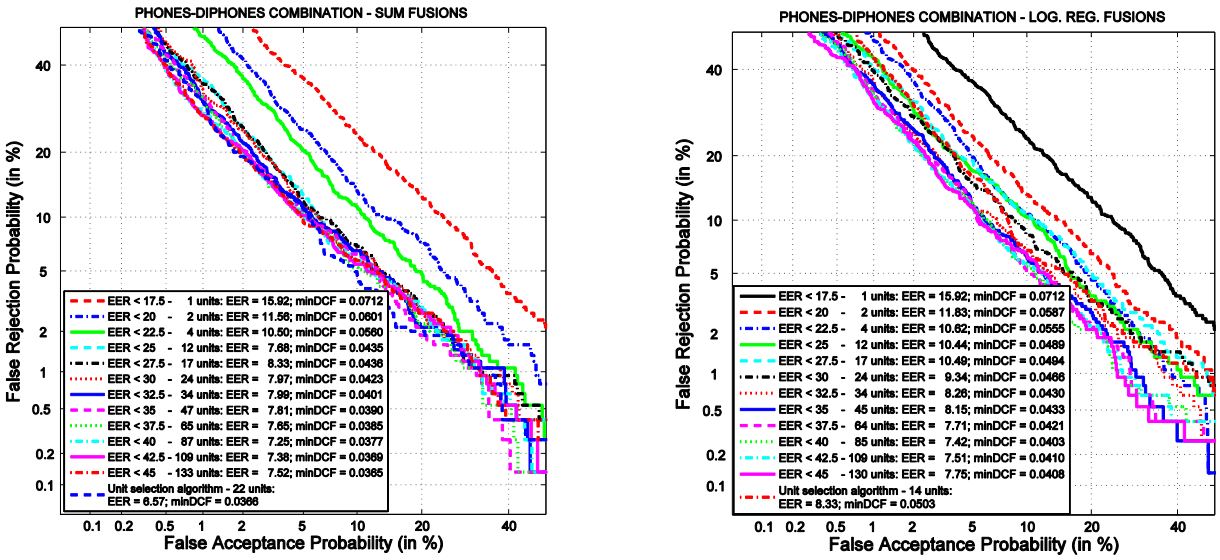

Fig. 4. DET curves for sum and log. reg. fused systems in the NIST SRE 2006 English-only male 1side-1side task for different phone-diphone selection schemes.

It can be seen that better results can be achieve by combining phones and diphones units than working in a intra-unit manner, taking advantage of different linguistic levels. This way, it is possible to achieve improvements around 35% in terms of EER over the reference system, as it can be seen in Table 3.

6 Summary and conclusions

In this paper we have presented an analysis of the contributions of individual linguistic units to automatic speaker recognition by means of their cepstral trajectories, showing that some of them can be used to obtain informative likelihood ratios very useful in forensic applications, with the advantage of being a completely automatic system and using parameters similar to those used by linguists or phoneticians. This way it is possible to deal with uncontrolled scenarios where only some short segments are available to be compared, making it possible to infer a conclusion about the speaker identity in the speech sample. This procedure cannot be done by the usual automatic speaker recognition systems because they use all available speech data as a whole, and usually they are tuned to work with fixed-length training and testing segments. Furthermore, when more testing data is available, individual units can be combined to improve the discrimination capabilities of the resulting system, having shown that these combinations, both at intra- and inter-unit levels, can outperform the results obtained with the same system framework based on MFCC features.

System	# fused units	EER (%)	minDCF
Reference	-	10.26	0.0457
Phones – best fused system (sum)	17	7.11	0.0420
Diphones – best fused system (log. reg.)	31	8.05	0.0473
Phones+diphones – best fused system (sum)	22	6.57	0.0366

Table 3. Performance comparison between the reference system and unit-based fused systems in the NIST SRE 2006 English-only male 1side-1side task

7 References

1. Brummer, N. et al., "Application-independent evaluation of speaker detection", Comp. Speech Lang., (20) 230-275, 2006.
2. Dehak, N., et al., "Front-End Factor Analysis for Speaker Verification", IEEE Trans. on Audio, Speech and Lang. Proc., 19(4), 788-798, May 2011.
3. Castro, A. d., Ramos, D., and Gonzalez-Rodriguez, J., "Forensic speaker recognition using traditional features comparing automatic and human-in-the-loop formant tracking", in Proceedings of Interspeech 2009, pp. 2343-2346, September 2009.
4. Ferrer, L., "Statistical modeling of heterogeneous features for speech processing tasks", Ph.D. dissertation, Stanford Univ., 2009 (http://www.speech.sri.com/people/lferrer/thesis.html)
5. Franco-Pedroso, J., Gonzalez-Rodriguez, J., Gonzalez-Dominguez, J., and Ramos, D. "Fine-grained automatic speaker recognition using cepstral trajectories in pone units". Proceedings of IAFPA 2012, Santander, Spain.
6. Kajarekar, S. et al., "The SRI NIST 2008 Speaker Recognition Evaluation System", Proc. IEEE ICASSP'09, pp. 4205-4209, Taipei, 2009.
7. Kenny, P. et al., "A Study of Inter-speaker Variability in Speaker Verification", IEEE Trans. on Audio, Speech and Lang. Proc., 16(5):980-988, 2008.
8. Kenny, P., "Bayesian speaker verification with heavy tailed priors", Keynote presentation at Odyssey 2010, Brno, 2010.
9. Kinnunen, T., and Li, H., "An overview of text-independent speaker recognition: from features to supervectors", Speech Communication, vol. 52, pp. 12-40, 2010.
10. NIST SRE 2006 Evaluation Plan: http://www.itl.nist.gov/iad/mig/tests/sre/2006/sre-06_evalplan-v9.pdf
11. Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., "Speaker verification using adapted gaussian mixture models", Digital Signal Processing 10, pp. 19-41, 2000.
12. Shriberg, E., "Modeling prosodic feature sequences for speaker recognition", Speech Communication, 46 (3-4), July 2005, pp. 455-472, Jan. 2005.
13. Wikipedia contributors. "Arpabet". *Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/wiki/Arpabet (19 July, 2012.)