UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

PROYECTO FIN DE CARRERA

Reconocimiento Automático de Locutor en Entornos Forenses basado en Técnicas de Factor Analysis aplicadas a Nivel Acústico

Eugenio Arévalo González

JULIO 2011

Reconocimiento Automático de Locutor en Entornos Forenses basado en Técnicas de Factor Analysis aplicadas a Nivel Acústico

AUTOR: Eugenio Arévalo González
TUTOR: Javier González Domínguez

Área de Tratamiento de Voz y Señales (ATVS)
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Julio de 2011

Agradecimientos

Dicen que “el orden de los factores no altera el producto” y es por eso que todas las personas que aparecen mencionadas aquí merecen de igual manera mi agradecimiento, porque todo lo que hemos compartido ha servido para que me convierta en la persona que soy. Son mis primeros agradecimientos, así que siento de antemano la parrafada.

Para empezar quiero agradecer a mis padres, **Eugenio** e **Inés**, que hayan confiado siempre en mí, que me hayan dado libertad para todo y que hayan dejado que siga mi ritmo en la carrera, sin presiones, y que, aun sabiendo que podría haber terminado mucho antes, siempre tienen una sonrisa para mí y un “no te preocupes hijo”. Os quiero.

Por supuesto tengo que agradecer también al resto de mi familia, que siempre me ha tenido en un pedestal inmerecido, y que me deja seguir siendo “el niño” tenga la edad que tenga. Gracias abuelos, tíos y primos.

Gracias a todos los compañeros de universidad que han compartido conmigo prácticas, exámenes, agobios, clases, salidas, fiestas, viajes… Gracias por aguantarme en los momentos de estrés cuando me pongo borde e insoportable. Para algunos tengo que añadir más:

Gracias **Alicia**, por no perder la confianza conmigo, por seguir en contacto, por lo amenas que me hacías las clases, por lo mucho que me río contigo y porque siempre consigues hacer que vea que las cosas no tienen tanta importancia como yo les quiero dar.

Gracias **Eva**, mi compi de risas en el laboratorio, por hacerme más llevaderas las clases y el pfc, por “convivir” conmigo en B-203 y C-109, por nuestros gabinetes de crisis en los que nos quejamos de todo, y porque no recordamos cómo ni cuándo nos conocimos, pero ha merecido la pena haberlo hecho. Y gracias **Pi**, por tus consejos, tu paciencia y tu amabilidad.

Gracias a mis amigos de Madrid de toda la vida, mis **Inclusos**: por seguir ahí tropecientos años después, animándome y diciendo que mi carrera es muy difícil y que es normal que tarde en acabarla. Cris, Belén y Gema, siempre seréis mis niñas.

No puedo olvidarme de mi gente de **Malpartida City**, y menos de los que habéis estado siempre ahí, poniendo cara de que mi proyecto os resultaba interesante: gracias **Bea** porque siempre nos entendemos, **Iván** porque eres un baboso al que quiero, y **Nacho** porque eres mi mejor amigo, mi hermano mayor y mi compañero de aventuras europeas.

Y dentro del universo extremeño he de agradecer muy especialmente a **Laura**, mi *surimi*, la mejor amiga que puede haber, porque lo sabe todo de mí, porque me anima, porque nos reímos juntos, porque me llama cuando sabe que estoy mal y porque la quiero más que a nada.

También quiero dar las gracias a 3 personas que me han acompañado durante la carrera, que han aguantado mis días de agobio y mal humor, y que han entendido que pasase más tiempo con una compañera de prácticas que con ellas. Gracias a las que habéis sido “mi pareja” en algún momento de la vida universitaria: **Iris**, **Elena** y **Patricia**.

No puedo olvidarme de mis compañeros y jefes de **Indra**, gracias por confiar en mí desde el principio, por brindarme una primera experiencia profesional, por aconsejarme y enseñarme, y por integrarme tan pronto en el grupo.

Gracias a todos mis compañeros de baile, porque nos los pasamos tan bien y me río tanto en clase, que me ha servido como terapia anti estrés de cara a la universidad durante estos años.

Para ese grupo de amigos que me han aceptado enseguida como uno más, que bromean con mi edad y el tiempo que he tardado en acabar la carrera y que no se enteran de nada cuando les cuento de qué va mi proyecto… Gracias por formar parte de mi gente, especialmente *Hurdianos* y *Power Rangers*. ¡Os quiero **Garcianos**!

Muchísimas gracias a todo mi equipo de trabajo, el ATVS, por hacerme sentir como uno más, por el buen rollo que hay siempre y por esas "reuniones de sillas" en el B203. Gracias a Almudena, Eva, Mario, Sergio, Virginia (¡¡esos mails bostonianos/florentinos!!), Álvaro, Jaime y todos los demás. Gracias Galbally por ser un apoyo en las prácticas de Patrones. Javier Franco, mil gracias por ayudar siempre con una sonrisa sin poner pegas, incluso estando ocupado. Agradecimiento especial para **Daniel Ramos** por buscarme proyecto y un tutor que encajase conmigo, y por esa alegría que me transmite cada vez que hablo con él.

Por supuesto quiero dar las gracias a mi tutor, **Javier González**, por su trato cercano, por el esfuerzo que ha llevado a cabo conmigo durante todos estos meses, por su paciencia, por todo lo que he aprendido con y de él, por su simpatía y porque sin él no habría pfc.

Hay alguien que he conocido al final de la carrera pero se ha convertido en una de las personas más importantes de mi vida. Me ha metido prisa para terminar la carrera, se ha interesado por mi proyecto, ha aguantado mis bajones, mis cambios de humor, mis quejas… Y siempre ha estado ahí, de forma incondicional. Tengo mucho, muchísimo, que agradecerle, pero vamos a dejarlo en un "gracias por hacerme sonreír". Como tú dijiste: TODO. Para ti **Dani**, porque eres un *ñu*, mil gracias. 私はあなたを愛しています。

Sólo me queda una persona por mencionar, y la he dejado para el final porque es lo que se hace cuando quieres decir muchas cosas en pocas líneas y no sabes cómo hacerlo. Nos conocimos en 1º, en prácticas de Física, poniendo la oreja en un tubo de cristal e intentando diferenciar "armónicos". Los 2 teníamos la misma cara de tonto. Ella es a la vez mi mujer y mi amiga, es mi ardilla aventurera (¡Chip!), mi más frecuente (y mejor) compañera de prácticas, hemos compartido alegrías y penas, *marujeos*, muchos viajes (y los que nos quedan) y mil cosas más. Nadie como ella sabe lo que siento y pienso en cada momento, y podemos estar a 50000 km de distancia y seguir contándonos todo como si fuésemos vecinos. **Lucía**, si algo hace que merezca la pena haber hecho Ingeniería de Telecomunicación es haberte conocido, así que sólo me queda decirte **GRACIAS**.

Eugenio Arévalo González

Resumen

El presente proyecto se dedica al estudio del impacto de la aplicación de técnicas basadas en *Factor Analysis* en un sistema de reconocimiento automático de locutor. Con este propósito se llevan a cabo diferentes experimentos sobre bases de datos pertenecientes a entornos controlados (evaluaciones NIST SRE) y sobre bases de datos forenses (Ahumada III) para realizar comparaciones entre los resultados obtenidos en los mismos.

El objetivo final es conseguir un decremento de las tasas de error o, al menos, realizar un estudio que compruebe el impacto de las diferentes técnicas utilizadas en los resultados finales.

Los métodos utilizados en la elaboración de los experimentos están formados por compensación de variabilidad intersesión e interlocutor, normalización de puntuaciones, ajuste de las matrices de variabilidad a las longitudes de la prueba y ajuste de las cohortes de normalización a las longitudes de la prueba.

Todos los experimentos se han elaborado utilizando el software de voz del ATVS, acerca de cuyo manejo puede encontrarse un tutorial en los anexos de la presente memoria. Además, los resultados obtenidos han contribuido a publicaciones utilizadas en congresos de carácter internacional.

Palabras Clave

Biometría forense, reconocimiento automático de locutor, verificación, variabilidad, compensación, puntuaciones, normalización, duración, longitud, entrenamiento, test, sesión, *eigenvoices, eigenchannels*, modelo Gaussiano, adaptación, *factor analysis*.

Abstract

This project is centered in the study of the application of Factor Analysis based techniques in an automatic speaker recognition system. For that purpose, different experiments in controlled databases (NIST SRE evaluations) and forensic databases (Ahumada III) are carried out in order to make comparisons between the results obtained.

The final objective is achieving a decrement of the error rates or, at least, realizing a study which observes the impact of the different techniques utilized in the final results.

The methods used in the elaboration of these experiments are constituted by session and speaker variability compensation, scores normalization and adaptation of the variability matrices and normalization cohorts to the experiment longitudes.

Every experiment has been elaborated using the ATVS voice software, about whom a tutorial can be found in this memory annexes. Additionally, the results obtained have contributed in a publication of an international workshop.

Key Words

Forensic biometrics, automatic speaker recognition, verification, variability, compensation, scores, normalization, duration, length, training, test, session, *eigenvoices*, *eigenchannels*, Gaussian model, adaptation, factor analysis.

Índice de Contenidos

Agradecimientos i
Resumen iv
Palabras Clave iv
Abstract v
Key Words v
Índice de Contenidos vii
Índice de Figuras xi
Índice de Tablas xvi
1. Introducción 1
1.1. Motivación 1
1.2. Objetivos y Metodología 3
1.3. Organización de la Memoria 3
1.4. Contribuciones 4
2. Sistemas de Reconocimiento Biométrico 7
2.1. Introducción 7
2.2. Breve Cronología Histórica de la Biometría 7
2.3. Reconocimiento Biométrico 7
2.4. Rasgos Biométricos 10
2.4.1. Huella Dactilar 10
2.4.2. Iris 10
2.4.3. Retina 11
2.4.4. Geometría de la Mano 11
2.4.5. Firma 11
2.4.6. Escritura 11
2.4.7. Dinámica de Tecleo 12
2.4.8. Oreja 12
2.4.9. Dientes 12
2.4.10. Termografía Facial 12
2.4.11. Voz 13
2.5. Biometría Forense 13
2.6. Funcionamiento de un Sistema de Reconocimiento Biométrico 16
2.7. Modos de Operación 17
2.7.1. Sistemas de Reconocimiento en Modo Registro 17
2.7.2. Sistemas de Reconocimiento en Modo Identificación 18
2.7.3. Sistemas de Reconocimiento en Modo Verificación 19
2.8. Evaluación de Sistemas Biométricos 19
2.8.1. Factores que afectan al rendimiento 20
2.8.2. Adquisición de Datos 20
2.8.3. Rendimiento del Sistema 21
2.9. Normalización de Puntuaciones 24
3. Sistemas de Reconocimiento Automático de Locutor 27
3.1. Introducción 27
3.2. Información de Identidad en la Señal de Voz 28
3.2.1. Tipos de Reconocedores 28

3.2.2. Niveles de Identidad 29
3.3. Descripción de un Sistema de Verificación de Locutor 30
3.4. Extracción de Características 32
3.4.1. Coeficientes LPCC 32
3.4.2. Coeficientes MFCC 33
3.5. Técnicas Empleadas en Reconocimiento de Locutor 35
3.5.1. Reconocimiento de Locutor Dependiente de Texto 36
3.5.2. Reconocimiento de Locutor Independiente de Texto 36
3.6. Modelos y Clasificadores 37
3.6.1. Detección de la Razón de Verosimilitud 37
3.6.2. Modelos de Mezclas Gaussianas (GMM) 38
3.6.3. Máquinas de Vectores Soporte (SVM) 42
3.7. Variabilidad de Sesión 44
3.7.1. Cepstral Mean Normalization 45
3.7.2. Filtrado RASTA 45
3.7.3. Feature Warping 45
3.7.4. Feature Mapping 45
3.7.5. Joint Factor Analysis 45
4. Técnicas Basadas en Factor Analysis aplicado al Reconocimiento de Locutor 47
4.1. Introducción 47
4.2. Bases del Modelo Joint Factor Analysis 48
4.2.1. Eigenvoices 48
4.2.3. Eigenchannels 48
4.2.4. Joint Factor Analysis 49
5. Bases de Datos y Protocolos de Evaluación 51
5.1. Bases de Datos para Reconocimiento de Locutor 51
5.2. Protocolos de Evaluación 53
5.2.1. Evaluaciones NIST 53
6. Experimentos 57
6.1. Introducción 57
6.2. Efecto de la Compensación de Variabilidad en Entornos Controlados 57
6.2.1. Introducción 57
6.2.2. Sistema de Partida 58
6.2.3. Rendimiento del Sistema Base 58
6.2.4. Rendimiento del Sistema tras Modelar la Variabilidad de Locutor 60
6.2.5. Rendimiento del Sistema tras compensar la Variabilidad Intersesión y modelar la Variabilidad de Locutor mediante técnicas de JFA 61
6.2.6. Rendimiento del Sistema tras entrenar el Espacio de Variabilidad de Locutor Ampliado y compensar la Variabilidad Intersesión e Interlocutor mediante técnicas de JFA 63
6.2.7. Comparativa entre las diferentes técnicas 64
6.3. Efecto de la Compensación de Variabilidad en Entornos Forenses 66
6.3.1. Introducción 66

6.3.2. Sistema de Partida 66
6.3.3. Rendimiento del Sistema Base 67
6.3.4. Rendimiento del Sistema tras modelar la Variabilidad de Locutor 68
6.3.5. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA 68
6.3.6. Rendimiento del Sistema tras entrenar el Espacio de Variabilidad de Locutor Ampliado, compensar la Variabilidad Intersesión y modelar la Variabilidad de Locutor mediante JFA 69
6.3.7. Comparativa entre las diferentes técnicas 70
6.4. Efecto del Ajuste de las Cohortes a las Longitudes de la Prueba 71
6.4.1. Introducción 71
6.4.2. Rendimiento del Sistema Base tras aumentar las Cohortes de Normalización 72
6.4.3. Rendimiento del Sistema tras Modelar la Variabilidad de Locutor, Compensar la Variabilidad de Canal y aumentar las Cohortes de Normalización 73
6.4.4. Rendimiento del Sistema tras ajustar los estadísticos de las Cohortes de Normalización a las Longitudes de la Prueba 74
6.4.5. Rendimiento del Sistema tras ajustar las Matrices de Variabilidad a las longitudes de la prueba 76
6.5. Efecto del Número de Direcciones de Máxima Variabilidad Utilizadas 78
6.5.1. Introducción 78
6.5.2. Rendimiento del Sistema tras modelar la Variabilidad de Locutor para diferente número de eigenvoices 79
6.5.3. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando únicamente el número de eigenvoices 83
6.5.4. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando únicamente el número de eigenchannels 87
6.5.5. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando eigenchannels y eigenvoices 91
6.5.5.1. Rendimiento con 50 eigenvoices y eigenchannels variables 91
6.5.5.2. Rendimiento con 100 eigenvoices y eigenchannels variables 92
6.5.5.3. Rendimiento con 150 eigenvoices y eigenchannels variables 93
6.5.5.4. Rendimiento con 200 eigenvoices y eigenchannels variables 94
6.5.5.5. Rendimiento con 250 eigenvoices y eigenchannels variables 95
6.5.5.6. Rendimiento con 300 eigenvoices y eigenchannels variables 96

6.5.5.7. Comentarios 97
6.6. Efecto del Número de Direcciones de Máxima Variabilidad Utilizadas en Entornos Conocidos 98
6.6.1. Introducción 98
6.6.2. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando únicamente el número de eigenvoices 98
6.6.3. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando únicamente el número de eigenchannels 102
6.6.4. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando eigenchannels y eigenvoices 106
6.6.4.1. Rendimiento con 50 eigenvoices y eigenchannels variables 106
6.6.4.2. Rendimiento con 100 eigenvoices y eigenchannels variables 107
6.6.4.3. Rendimiento con 150 eigenvoices y eigenchannels variables 108
6.6.4.4. Rendimiento con 200 eigenvoices y eigenchannels variables 109
6.6.4.5. Rendimiento con 250 eigenvoices y eigenchannels variables 110
6.6.4.6. Rendimiento con 300 eigenvoices y eigenchannels variables 111
6.6.4.7. Comentarios 112
7. Conclusiones y Trabajo Futuro 113
7.1. Conclusiones 113
7.1.1. Efecto de la compensación de variabilidad 113
7.1.2. Efecto del ajuste de las cohortes a las longitudes de la prueba 113
7.1.3. Efecto del número de direcciones de máxima variabilidad utilizadas 114
7.2. Trabajo futuro 114
Referencias 117
Glosario I
Presupuesto III
Pliego de Condiciones V

Índice de Figuras

Figura 1-1. Aplicaciones del Reconocimiento de Voz 1
Figura 1-2. Ejemplos de Reconocimiento de Habla y Reconocimiento de Locutor 2
Figura 2-1. Distribución en el mercado de los diferentes rasgos biométricos en 2009 10
Figura 2-2. Transferencia de la evidencia 14
Figura 2-3. Ejemplos de marcas en escenario forense 15
Figura 2-4. Aproximación de James L. Wayman de la estructura del procedimiento biométrico 16
Figura 2-5. Diagrama de funcionamiento de un sistema biométrico en modo registro 17
Figura 2-6. Diagrama de funcionamiento de un sistema biométrico en modo identificación 18
Figura 2-7. Diagrama de funcionamiento de un sistema biométrico en modo verificación 19
Figura 2-8. Ejemplo de obtención de Equal Error Rate 22
Figura 2-9. Densidades y distribuciones de probabilidad de usuarios legítimos e impostores 23
Figura 2-10. Ejemplo de curvas DET para diferentes sistemas 24
Figura 3-1. Resumen de características desde el punto de vista de su interpretación física 29
Figura 3-2. Niveles de identidad 30
Figura 3-3. Diagrama de bloques de un sistema de reconocimiento de locutor 30
Figura 3-4. Diagrama de bloques de la fase de entrenamiento de un sistema de verificación de locutor 31
Figura 3-5. Diagrama de la fase de test de un sistema de verificación de locutor 31
Figura 3-6. Diagrama de bloques de una parametrización cepstral basada en LPC 32
Figura 3-7. Diagrama de bloques de una parametrización cepstral basada en banco de filtros 33
Figura 3-8. Ejemplo de enventanado 33
Figura 3-9. Enventanado de señal de audio con ventana Hamming 34
Figura 3-10. Sistema de verificación de locutor basado en razón de verosimilitud 37
Figura 3-11. Distribución espacial de los coeficientes espectrales y GMM entrenado a partir de la misma 39
Figura 3-12. Ejemplo de adaptación de un modelo de locutor 41
Figura 3-13. Ejemplo de vectores soporte 42
Figura 3-14. Ejemplo de SVM visto de forma espacial 43
Figura 6-1. Rendimiento del sistema base con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos 59

Figura 6-2. Rendimiento del sistema base con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 10s-10s para locutores masculinos 59
Figura 6-3. Rendimiento del sistema tras modelar la variabilidad de locutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos 60
Figura 6-4. Rendimiento del sistema tras modelar la variabilidad de locutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 10s-10s para locutores masculinos 61
Figura 6-5. Rendimiento del sistema tras compensar la variabilidad de canal y modelar la de locutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos 62
Figura 6-6. Rendimiento del sistema tras compensar la variabilidad de canal y modelar la de locutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 10s- 10s para locutores masculinos 62
Figura 6-7. Rendimiento del sistema tras entrenar el espacio de variabilidad de locutor ampliado y compensar la variabilidad intersesión e interlocutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos 63
Figura 6-8. Rendimiento del sistema tras entrenar el espacio de variabilidad de locutor ampliado y compensar la variabilidad intersesión e interlocutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos 64
Figura 6-9. Rendimiento del sistema tras aplicar las diferentes técnicas de compensación de variabilidad. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos 65
Figura 6-10. Rendimiento del sistema tras aplicar las diferentes técnicas de compensación de variabilidad. Evaluación sobre datos NIST SRE 2008 tarea 10sec-10sec para locutores masculinos 65
Figura 6-11. Rendimiento del sistema base con normalización de puntuaciones. Evaluación sobre datos procedentes de Ahumada III 67
Figura 6-12. Rendimiento del sistema tras realizar modelado en locutor con normalización de puntuaciones. Evaluación sobre datos procedentes de Ahumada III 68
Figura 6-13. Rendimiento del sistema tras compensar la variabilidad intersesión e interlocutor mediante JFA con normalización de puntuaciones. Evaluación sobre datos procedentes de Ahumada III 69
Figura 6-14. Rendimiento del sistema tras entrenar el espacio de variabilidad ampliada, compensar en canal y modelar en locutor mediante JFA y con normalización de puntuaciones. Evaluación sobre datos procedentes de Ahumada III 70
Figura 6-15. Rendimiento del sistema tras aplicar las diferentes técnicas de compensación de variabilidad sin normalizar puntuaciones. Evaluación sobre datos procedentes de Ahumada III 71
Figura 6-16. Rendimiento del sistema base con normalización NIST SRE 2005. Base de Datos Ahumada III 72
Figura 6-17. Rendimiento del sistema base tras aumentar las cohortes de normalización con datos de NIST SRE 2004. Base de Datos Ahumada III 73

Figura 6-18. Rendimiento del Sistema tras aplicar JFA antes y después de aumentar las cohortes de normalización. Base de Datos Ahumada III 74
Figura 6-19. Rendimiento del sistema sin compensar y tras aplicar JFA, tras el ajuste de las longitudes de los estadísticos en las cohortes de normalización (en todas y únicamente en z-norm). Base de Datos Ahumada III 76
Figura 6-20. Rendimiento del Sistema tras ajustar las matrices de variabilidad de sesión a las longitudes de la prueba y aplicar JFA. Base de Datos Ahumada III 77
Figura 6-21. Rendimiento del Sistema tras ajustar las matrices de variabilidad de sesión y locutor a las longitudes de la prueba y aplicar JFA. Base de Datos Ahumada III 77
Figura 6-22. Rendimiento del sistema tras ajustar las matrices de variabilidad así como las cohortes de normalización. Base de Datos Ahumada III 78
Figura 6-23. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 50 eigenvoices. Base de Datos Ahumada III 80
Figura 6-24. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 100 eigenvoices. Base de Datos Ahumada III 80
Figura 6-25. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 150 eigenvoices. Base de Datos Ahumada III 81
Figura 6-26. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 200 eigenvoices. Base de Datos Ahumada III 81
Figura 6-27. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 250 eigenvoices. Base de Datos Ahumada III 82
Figura 6-28. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 300 eigenvoices. Base de Datos Ahumada III 82
Figura 6-29. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con diferente número de eigenvoices. Gráfica comparativa. Base de Datos Ahumada III 83
Figura 6-30. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 50 eigenvoices y compensar la Variabilidad de Canal. Base de Datos Ahumada III 84
Figura 6-31. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 100 eigenvoices y compensar la Variabilidad de Canal. Base de Datos Ahumada III 85
Figura 6-32. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 150 eigenvoices y compensar la Variabilidad de Canal. Base de Datos Ahumada III 85
Figura 6-33. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 200 eigenvoices y compensar la Variabilidad de Canal. Base de Datos Ahumada III 86
Figura 6-34. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 250 eigenvoices y compensar la Variabilidad de Canal. Base de Datos Ahumada III 86
Figura 6-35. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 300 eigenvoices y compensar la Variabilidad de Canal. Base de Datos Ahumada III 87

Figura 6-36. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 10 eigenchannels. Base de Datos Ahumada III 88
Figura 6-37. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 20 eigenchannels. Base de Datos Ahumada III 89
Figura 6-38. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 30 eigenchannels. Base de Datos Ahumada III 89
Figura 6-39. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 40 eigenchannels. Base de Datos Ahumada III 90
Figura 6-40. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 50 eigenchannels. Base de Datos Ahumada III 90
Figura 6-41. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 50 eigenvoices y eigenchannels variables. Base de Datos Ahumada III 92
Figura 6-42. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 100 eigenvoices y eigenchannels variables. Base de Datos Ahumada III 93
Figura 6-43. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 150 eigenvoices y eigenchannels variables. Base de Datos Ahumada III 94
Figura 6-44. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 200 eigenvoices y eigenchannels variables. Base de Datos Ahumada III 95
Figura 6-45. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 250 eigenvoices y eigenchannels variables. Base de Datos Ahumada III 96
Figura 6-46. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 300 eigenvoices y eigenchannels variables. Base de Datos Ahumada III 97
Figura 6-47. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 50 eigenvoices y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv- 1conv 99
Figura 6-48. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 100 eigenvoices y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv- 1conv 99
Figura 6-49. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 150 eigenvoices y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv- 1conv 100
Figura 6-50. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 200 eigenvoices y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv- 1conv 100

Figura 6-51. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 250 eigenvoices y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv- 1conv 101
Figura 6-52. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 300 eigenvoices y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv- 1conv 101
Figura 6-53. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con número variable de eigenvoices y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv-1conv 102
Figura 6-54. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 10 eigenchannels. Base de Datos NIST SRE 2008 tarea 1conv-1conv 103
Figura 6-55. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 20 eigenchannels. Base de Datos NIST SRE 2008 tarea 1conv-1conv 104
Figura 6-56. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 30 eigenchannels. Base de Datos NIST SRE 2008 tarea 1conv-1conv 104
Figura 6-57. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 40 eigenchannels. Base de Datos NIST SRE 2008 tarea 1conv-1conv 105
Figura 6-58. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 50 eigenchannels. Base de Datos NIST SRE 2008 tarea 1conv-1conv 105
Figura 6-59. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 50 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 107
Figura 6-60. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 100 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 108
Figura 6-61. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 150 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 109
Figura 6-62. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 200 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 110
Figura 6-63. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 250 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 111
Figura 6-64. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 300 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 112

Índice de Tablas

Tabla 2-1. Ejemplos de rasgos biométricos 9
Tabla 2-2. Características de la voz 13
Tabla 2-3. Factores ambientales que afectan a cada tipo de sistema 20
Tabla 4-1. Descomposición del modelo JFA 50
Tabla 5-1. Condiciones de entrenamiento y test en la evaluación NIST SRE 2006 54
Tabla 5-2. Condiciones de entrenamiento y test en la evaluación NIST SRE 2008 55
Tabla 6-1. Rendimiento del Sistema Base antes y después de aumentar las cohortes de normalización. Base de Datos Ahumada III 72
Tabla 6-2. Rendimiento del Sistema aplicando JFA antes y después de aumentar las cohortes de normalización. Base de Datos Ahumada III 74
Tabla 6-3. Rendimiento del sistema sin compensar y tras aplicar JFA, antes y después del ajuste de las longitudes de los estadísticos en las cohortes de normalización. Base de Datos Ahumada III 75
Tabla 6-4. Rendimiento del sistema sin compensar y tras aplicar JFA, ajustando las longitudes de los estadísticos en la cohorte de z-norm. Base de Datos Ahumada III 75
Tabla 6-5. Rendimiento del Sistema tras modelar la Variabilidad de Locutor variando el número de eigenvoices. Base de Datos Ahumada III 79
Tabla 6-6. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenvoices. Base de Datos Ahumada III 84
Tabla 6-7. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels. Base de Datos Ahumada III 88
Tabla 6-8. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 50 eigenvoices. Base de Datos Ahumada III 91
Tabla 6-9. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 100 eigenvoices. Base de Datos Ahumada III 92
Tabla 6-10. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 150 eigenvoices. Base de Datos Ahumada III 93
Tabla 6-11. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 200 eigenvoices. Base de Datos Ahumada III 94
Tabla 6-12. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 250 eigenvoices. Base de Datos Ahumada III 95

Tabla 6-13. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 300 eigenvoices. Base de Datos Ahumada III 96
Tabla 6-14. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 98
Tabla 6-15. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels. Base de Datos NIST SRE 2008 tarea 1conv-1conv 103
Tabla 6-16. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 50 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 106
Tabla 6-17. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 100 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 107
Tabla 6-18. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 150 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 108
Tabla 6-19. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 200 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 109
Tabla 6-20. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 250 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 110
Tabla 6-21. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de eigenchannels para 300 eigenvoices. Base de Datos NIST SRE 2008 tarea 1conv-1conv 111

1. Introducción

1.1. Motivación

En la actualidad el uso de técnicas de reconocimiento biométrico va tomando cada vez más importancia debido a la necesidad creciente de contar con herramientas poderosas que incrementen la seguridad en la identificación de personas [8, 9, 10].

El reconocimiento de voz es una de las tecnologías biométricas más populares gracias al desarrollo de la telefonía y de las redes informáticas [2], y muy aceptada socialmente ya que se trata de una técnica no invasiva que cuenta con numerosas aplicaciones (figura 1-1) y cuyo proceso de adquisición no supone grandes costes [1].

Figura 1-1. Aplicaciones del Reconocimiento de Voz.

El reconocimiento de voz se divide en dos áreas: reconocimiento de habla y reconocimiento de locutor. Mientras que el área de reconocimiento de habla concierne la extracción del mensaje lingüístico de una locución, el reconocimiento de locutor trata la extracción de la identidad de la persona que pronuncia la locución [2].

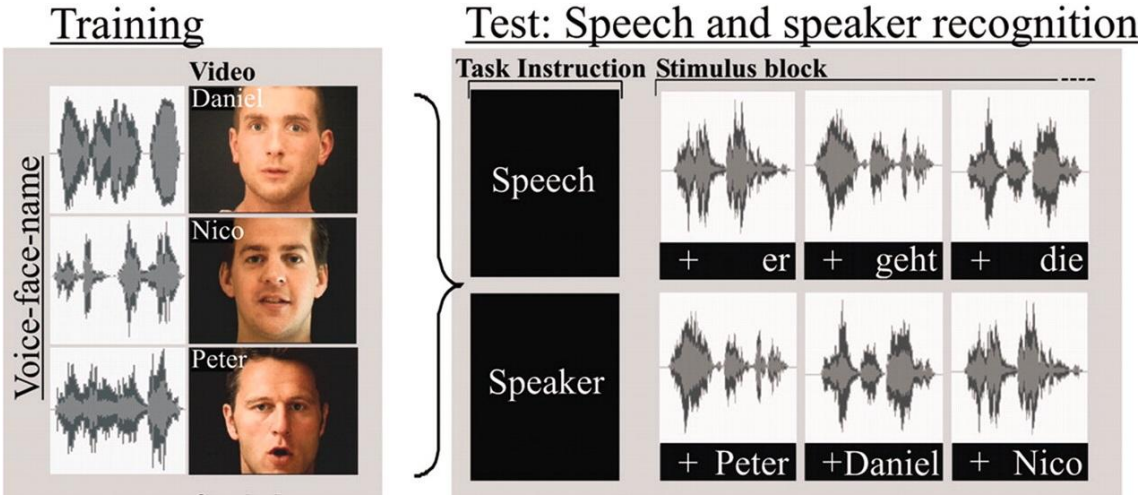


Figura 1-2. Ejemplos de Reconocimiento de Habla y Reconocimiento de Locutor [11].

El eje central de este proyecto es el reconocimiento de locutor, puesto que es el área de interés en entornos forenses, por lo que se detalla a continuación el funcionamiento de los sistemas que lo emplean:

Dadas dos muestras de voz, el objetivo de un sistema automático de reconocimiento de locutor es establecer una medida de verosimilitud entre las identidades asociadas a ambas muestras [1, 2, 3].

Dicha verosimilitud es obtenida mediante diversos tipos de procesado/modelado matemático aplicados a distintas características extraídas de la señal de voz a diferentes niveles (acústico, fonético…).

Es bien conocido, que una de las mayores causas de degradación en este tipo de sistemas se debe a la variabilidad acústica existente entre diferentes muestras asociadas a mismas identidades (variabilidad intersesión) [4]. Esta variabilidad es causada por muy diversos factores como son: el tipo de canal utilizado (telefonía fija, micrófono), ruido ambiente, distintos entornos de grabación, etc. Además, este tipo de degradación es más acuciante en entornos forenses, donde la variabilidad presentada en las muestras es a menudo extrema debido a las condiciones de adquisición [5]. En este sentido, el uso de técnicas de compensación de variabilidad de sesión, basadas en *Factor Analysis*, en el área, ha conducido a una significativa reducción de las tasas de error de los sistemas automáticos [4,6].

Durante el desarrollo de este Proyecto Fin de Carrera (PFC) se explorará y evaluará el uso de este tipo de técnicas en escenarios forenses reales integradas en sistemas de reconocimiento basados en GMMs [7], donde la compensación de variabilidad de sesión es clave y supone un desafío en el trabajo diario de los laboratorios de criminalística.

1.2. Objetivos y Metodología

Los objetivos principales que se persigue con este proyecto son:

- Explorar la algoritmia y funcionamiento de los sistemas de reconocimiento automático de locutor basados en *Gaussian Mixture Models* (GMM) y aplicados en entornos forenses.
- Estudiar, implementar e integrar diferentes técnicas de compensación de variabilidad basadas en *Joint Factor Analysis* (JFA) en sistemas de reconocimiento automático de locutor basados en GMMs, así como estudiar su funcionamiento en entornos forenses reales.

Para lograrlos se ha seguido la siguiente metodología y plan de trabajo, en varias fases:

- **Documentación:** antes de comenzar a trabajar con los sistemas de reconocimiento de locutor es necesario formarse al respecto para lo que es útil estudiar la bibliografía básica (publicaciones científicas y libros) sobre el estado del arte actual en biometría, técnicas de reconocimiento de locutor independientes de texto, GMMs, compensación de variabilidad… Así como documentarse sobre las bases de datos que se utilizarán (NIST y Ahumada).
- **Estudio del software:** se han analizado los algoritmos de reconocimiento de locutor ya desarrollados por el ATVS y se han realizado diversas pruebas sobre bases de datos derivadas de evaluaciones NIST (*National Institute of Standards and Technology*) SRE 2006 y 2008, y Ahumada III.
- **Investigación:** se ha realizado un estudio e implementación de diversas técnicas de compensación de variabilidad de sesión en entornos controlados y forenses.
- **Evaluación de resultados y elaboración de la memoria:** se ha realizado un análisis de los resultados obtenidos en las pruebas llevadas a cabo así como una comparativa entre las diferentes técnicas utilizadas según los resultados arrojados por las mismas. Estos análisis, junto con la revisión del estado del arte y un estudio completo del proyecto llevado a cabo, servirán para elaborar la memoria que pone puto y final al proyecto fin de carrera.

1.3. Organización de la Memoria

La presente memoria se encuentra estructurada de la siguiente forma:

En el **capítulo 2** se realiza una revisión del *estado del arte* en el campo de la biometría, dado que los sistemas de reconocimiento de locutor analizados en este proyecto se encuentran englobados dentro de los sistemas de reconocimiento biométrico. Por ello, en este capítulo se analiza de forma genérica el concepto de biometría, los rasgos biométricos, las propiedades que deben cumplir… Además, se realiza una introducción al concepto de biometría forense, y se entra en detalle a explicar los sistemas de reconocimiento biométrico, su funcionamiento, sus modos de operación y la evaluación de su rendimiento.

Una vez introducidos los sistemas de reconocimiento biométrico, pasamos a explicar los sistemas de reconocimiento automático de locutor en el **capítulo 3**. En este capítulo se profundiza en la señal de voz y la información que esta conlleva, se explica cómo extraer las características que vamos a analizar de la misma, las diferentes técnicas en reconocimiento de locutor y los tipos de clasificadores dando especial importancia a los modelos de mezclas Gaussianas (GMM), sobre el cual versa este proyecto. Además, se define el concepto de variabilidad de sesión así como las diferentes técnicas para tratar con ella, prestando mayor atención a las basadas en *Factor Analysis* puesto que en ellas se basa el trabajo realizado.

Para terminar con el *estado del arte*, en el **capítulo 4** se realiza un análisis más detallado de las técnicas basadas en *Factor Analysis* aplicado al reconocimiento de locutor, concretamente del modelo *Joint Factor Analysis* en el cual se centra el presente proyecto.

Una vez completos los apartados teóricos, se pasa a la parte experimental del proyecto:

En el **capítulo 5** se describen las bases de datos de reconocimiento de locutor empleadas para la realización de los experimentos y los protocolos de evaluación utilizados para llevar a cabo los mismos. Estos datos son esenciales para la evaluación del rendimiento del sistema. Hablaremos aquí de las bases de datos Ahumada, esenciales en el reconocimiento forense de locutor, y de las evaluaciones NIST SRE.

El **capítulo 6** compone el eje central del proyecto puesto que en él se muestran los experimentos realizados con cada base de datos y en determinadas condiciones de compensación de variabilidad de sesión, además de los resultados obtenidos y las comparaciones realizadas.

Por último, el **capítulo 7** muestra las conclusiones extraídas dados los resultados así como las líneas de trabajo futuro para continuar con la investigación en este ámbito del reconocimiento de locutor.

1.4. Contribuciones

El presente proyecto fin de carrera ha contribuido con el grupo de reconocimiento biométrico ATVS y la comunidad científica en los siguientes aspectos:

- Estudio del *estado del arte* de los sistemas de reconocimiento automático de locutor en entornos forenses.

- Adaptación de las bases de datos a las diferentes pruebas a realizar, aumentando así la cantidad de datos de cara a posteriores investigaciones.

- Adaptación de cohortes de normalización ó compensación de variabilidad a las longitudes de las pruebas.

- Estudio del software de reconocimiento de locutor del ATVS.

- Análisis de las diferentes técnicas de compensación de la variabilidad de sesión.

- Análisis del impacto de la adaptación de las cohortes de normalización a las longitudes de las pruebas.

- Algunos de los resultados obtenidos en la elaboración del proyecto han contribuido en la elaboración de artículos aceptados en congresos científicos de carácter internacional.

2. Sistemas de Reconocimiento Biométrico

2.1. Introducción

Este proyecto se basa en un sistema de reconocimiento automático de locutor en entornos forenses. Antes de entrar a explicar en detalle este tipo de sistemas se comenzará de forma más genérica por los sistemas de reconocimiento biométrico, y se realizará una revisión del estado del arte de los mismos.

2.2. Breve Cronología Histórica de la Biometría

- Siglo VIII: se encuentran huellas dactilares en China en documentos escritos [8].
- Año 1686: Marcello Malpighi realiza un estudio sobre las huellas dactilares en su tratado sobre las capas de la piel [8, 11].
- Año 1856: sir William Herschel valida un contrato mediante huella dactilar [8, 11].
- Año 1880: Henry Faulds muestra la utilidad de las huellas en escenas de crímenes [8, 11].
- Año 1890: Alphonse Bertillon estudia la mecánica del cuerpo y las medidas para identificar delincuentes [12].
- Año 1941: Murray Hill inicia el estudio de la identificación de voz [8].
- Década 1960-1970: se desarrolla la autenticación de firma [12].
- Años 70: primeros sistemas de reconocimiento automático de huellas dactilares [8].
- Año 1986: sir Alec Jeffreys utiliza el ADN para identificar a un criminal [8].

2.3. Reconocimiento Biométrico

La biometría es una ciencia que se dedica al estudio estadístico de las características cuantitativas de los seres vivos [8]. Recientemente el término se aplica también a la tecnología de identificación basada en el análisis de características y/o comportamiento de las personas, conocidos en conjunto como rasgos biométricos. Es un excelente sistema de autenticación que se aplica en muchos procesos debido a dos razones fundamentales: la seguridad y la comodidad.

El paradigma de la identificación personal es la autenticación de una entidad concreta relacionada con la persona. Dicha entidad es algo que el usuario tiene (problema: puede perderlo), algo que el usuario sabe (problema: puede olvidarlo) o algo que el usuario es (rasgo personal). No se puede distinguir un usuario de un impostor únicamente por posesión y/o conocimiento.

El reconocimiento biométrico se basa en la autenticación de la identidad mediante la identificación de rasgos de las personas. Dichos rasgos pueden ser fisiológicos o relacionados con la conducta.

Los rasgos fisiológicos son características intrínsecas a la naturaleza física de las personas, como por ejemplo: rostro, huella dactilar, retina, iris, rayas de la mano, geometría de la mano, poros de la piel, olor corporal, venas, ADN, termografía facial…

Los rasgos conductuales son características relacionadas con el comportamiento de los seres humanos, como por ejemplo: escritura, modo de teclear, forma de caminar…

La voz es un caso particular de rasgo biométrico porque se trata a la vez de una característica física (es propia de cada persona) y conductual (se ve afectada por emociones, lo que la hace variable).

En la tabla 2-1 se muestran algunos ejemplos de los principales rasgos biométricos y en la figura 2-1 se muestra de forma cuantitativa la presencia en el mercado de las diferentes tecnologías biométricas durante el año 2009 [15].

Para que los rasgos biométricos puedan utilizarse en identificación deben poseer una serie de propiedades ideales [16]:

- **Universalidad:** característica poseída por todas las personas.
- **Unicidad:** personas diferentes tienen distintos rasgos, es decir, es una característica distintiva.
- **Permanencia:** el rasgo debe ser invariante a corto plazo.
- **Perennidad:** el rasgo debe ser perpetuo.
- **Mensurabilidad:** el rasgo puede medirse de forma cuantitativa.

Además, los sistemas de reconocimiento biométrico deben cumplir características adicionales:

- **Rendimiento:** precisión, velocidad y robustez en el proceso de identificación.
- **Aceptabilidad:** grado de aceptación/rechazo personal y social del sistema.
- **Evitabilidad:** capacidad de eludir el sistema mediante procedimientos fraudulentos (sistemas robustos frente a posibles ataques).

Las principales ventajas de los rasgos biométricos con respecto a otros elementos de identificación/autenticación son las siguientes:

- Un rasgo biométrico es una manifestación tangible de lo que uno es.
- No se pueden sustraer, olvidar o descolocar.
- Combinadas con posesión (por ejemplo una llave o tarjeta) y/o conocimiento (una clave) son potentes herramientas de identificación personal.

El riesgo principal es la suplantación de identidad mediante la imitación o reproducción del rasgo a reconocer.

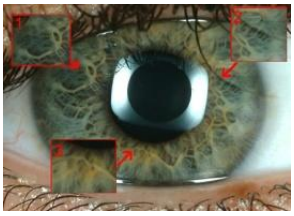
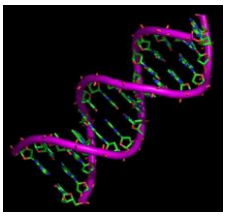
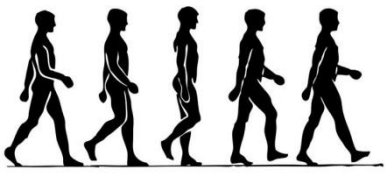
Rasgos Fisiológicos	Rasgos de Comportamiento
Huella 	**Escritura** 
Cara 	**Firma** 
Iris 	**Modo de teclear** 
ADN 	**Modo de andar** 
Voz 	

Tabla 2-1. Ejemplos de rasgos biométricos.

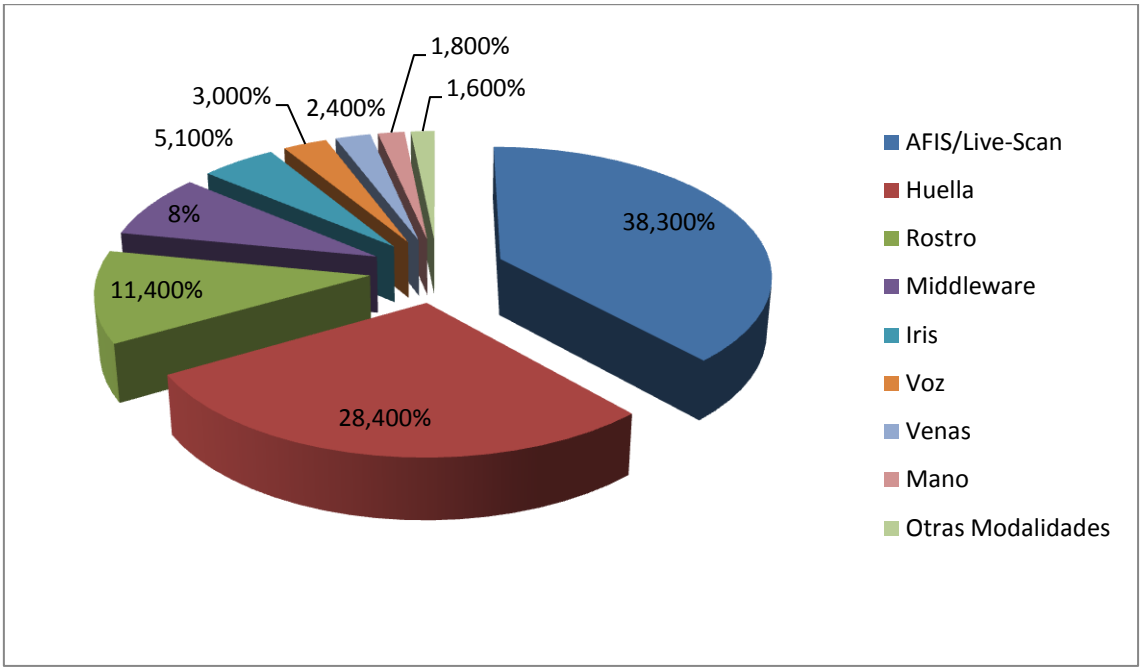


Figura 2-1. Distribución en el mercado de los diferentes rasgos biométricos en 2009 [15].

2.4. Rasgos Biométricos

A continuación se procede a detallar brevemente los diferentes rasgos biométricos mencionados con anterioridad.

2.4.1. Huella Dactilar

Es uno de los rasgos biométricos de mayor uso en el mercado (ver Figura 2-1) gracias a su alta aceptación popular y la facilidad de adquisición. Como ventajas frente a otros rasgos biométricos se encuentran la invariabilidad temporal y la variedad infinita del autentificador [23]. Su desventaja principal es la posibilidad de falsificación de forma relativamente sencilla.

2.4.2. Iris

El iris es un tejido pigmentado de alta movilidad protegido de agentes externos por la córnea, que al ser transparente permite que éste sea visible desde el exterior [24]. Sus principales características como rasgo biométrico son una muy alta unicidad y la estabilidad con el tiempo. Los problemas que puede presentar su adquisición son la oclusión del párpado, el ruido introducido por las pestañas, la falta de resolución o el movimiento del ojo .

2.4.3. Retina

El reconocimiento biométrico mediante retina se basa en la utilización de patrones de la red vascular alrededor del nervio óptico [10]. Es una técnica muy precisa y la retina es un rasgo muy estable y difícil de falsificar, lo que hace al sistema de reconocimiento idóneo. Sin embargo, las técnicas de obtención de los patrones son muy complejas e invasivas, por lo que la tecnología no está muy desarrollada.

2.4.4. Geometría de la Mano

La mano tiene diferentes características que permiten diferenciar a unas personas de otras [25], como pueden ser longitud, anchura de la palma, largo y grosor de los dedos, articulaciones… Además, es un rasgo que cuenta con unicidad y estabilidad aceptables y los sistemas de reconocimiento basados en la geometría de la mano tienen un coste bajo. Por el contrario, los sistemas de adquisición son bastante grandes y como rasgo es poco discriminativo [10], lo que hace que la técnica cuente con un elevado número de detractores.

2.4.5. Firma

La biometría basada en firma manuscrita hace uso de la información instantánea proporcionada por la tableta de firma (modo on-line) o de la imagen de la firma cuando no se tiene acceso al acto de firma (modo off-line) . Actualmente es muy común en verificación de la identidad de las personas. Sus ventajas son la alta aceptación como medio de autenticación, la baja invasividad del método de obtención, el hecho de que el entorno no influye en el proceso de adquisición y su gran disponibilidad en un amplio abanico de aplicaciones comerciales. Como inconvenientes nos encontramos una alta variabilidad intra-clase, variabilidad temporal inter-sesión y posibilidad de falsificación.

2.4.6. Escritura

La biometría de la escritura se basa en la identificación de escritor mediante el análisis de documentos escritos a mano [26]. Las formas y relaciones de los trazos de la escritura son utilizadas como características biométricas para autentificar e identificar a un individuo. Tiene gran relevancia en aplicaciones jurídicas y policiales.

2.4.7. Dinámica de Tecleo

El teclado es uno de los principales mecanismos de interacción de una persona con el ordenador, así como el que genera un mayor flujo de información entre usuario y ordenador, es por ello que puede utilizarse como medio de reconocimiento del usuario en procesos que requieren su autenticación [27].

La dinámica de tecleo se centra en las técnicas que identifican en qué medida existe una cierta regularidad en el modo de teclear de un usuario. Es muy útil en aplicaciones de seguridad informática y no requiere de hardware especial, pero también es un proceso complejo que involucra muchos factores característicos del usuario (además del aspecto físico es una capacidad que surge de la propia dinámica cerebral).

2.4.8. Oreja

Las orejas tienen ciertas ventajas sobre otros rasgos ya que poseen una estructura estable que cambia muy poco con la edad [18]. No se ven afectadas por cambios en la expresión facial y son un factor significativo desde el punto de vista de la identificación gracias a sus múltiples valles y crestas. Su principal desventaja es el ruido que puede introducir el cabello en la obtención de las imágenes. Actualmente el reconocimiento biométrico basado en oreja está poco desarrollado.

2.4.9. Dientes

Los dientes, y en particular las radiografías dentales, son de gran utilidad como medio de identificación en ámbitos forenses [28]. Generalmente se utilizan para identificar cadáveres. Las características utilizadas son el número de dientes presentes, la orientación de dichos dientes y las restauraciones dentales.

2.4.10. Termografía Facial

Consiste en la medida de los patrones infrarrojos de la emisión de calor de la cara, causado por el flujo de sangre bajo la piel. Se trata de una tecnología no invasiva ya que entre otras cosas no requiere contacto físico, puede hacerse a distancia y es accesible a la mayoría de los individuos.

2.4.11. Voz

La voz es el rasgo biométrico en que se centra este proyecto, por lo que se hablará de ella con más detalle posteriormente. El hecho de que las personas seamos capaces de identificar locutores a partir de sus voces es debido a que la señal de voz conlleva información de identidad del hablante. Es un rasgo muy aceptado en la sociedad para el reconocimiento de individuos y combina tanto características fisiológicas como de comportamiento. Su principal desventaja radica en que el hecho de verse afectada por factores conductuales como emociones o estado de salud hace más difícil su reconocimiento con respecto a otros rasgos biométricos.

En la tabla2-2, adaptada de [16] se muestran las diferentes características de un sistema biométrico y el grado en que la voz posee cada una de las mismas.

Universalidad	Unicidad	Permanencia	Mensurabilidad
Media	Baja	Baja	Media
Rendimiento	**Aceptabilidad**		**Evitabilidad**
Bajo	Alta		Alta

Tabla 2-2. Características de la voz [16].

2.5. Biometría Forense

En este apartado realizaremos un breve análisis descriptivo sobre biometría en el ámbito forense. Un amplio y extenso estudio sobre el reconocimiento de locutor en entornos forenses puede encontrarse en [19].

La ciencia forense se refiere a la aplicación de principios científicos y métodos técnicos a la investigación en relación con actividades criminales, para establecer la existencia de un delito y determinar la identidad de los culpables así como su *modus operandi* [29].

Una de las ramas de la ciencia forense es la criminalística [19], profesión y disciplina científica orientada al reconocimiento, identificación, individualización y evaluación de la evidencia física mediante la aplicación de ciencias naturales a los problemas científico-legales.

Según el principio de Locard [21], *todo contacto deja una marca*: o el malhechor deja signos en la escena del crimen o se lleva consigo (en su cuerpo o en la ropa) indicadores de dónde ha estado o qué ha hecho.

La evidencia forense es la relación entre una marca cuya fuente es desconocida y otro material originado por una fuente conocida, ambos relacionados de algún modo con un delito [20].

Existen dos tipos de material:

- Material recuperado, muestra o marca: se transfiere desde la persona implicada a la escena del crimen (por ejemplo la grabación de una cámara de seguridad) ó bien al contrario (muestras de ADN encontradas en la ropa del sospechoso).
- Material de control: muestras de origen conocido.

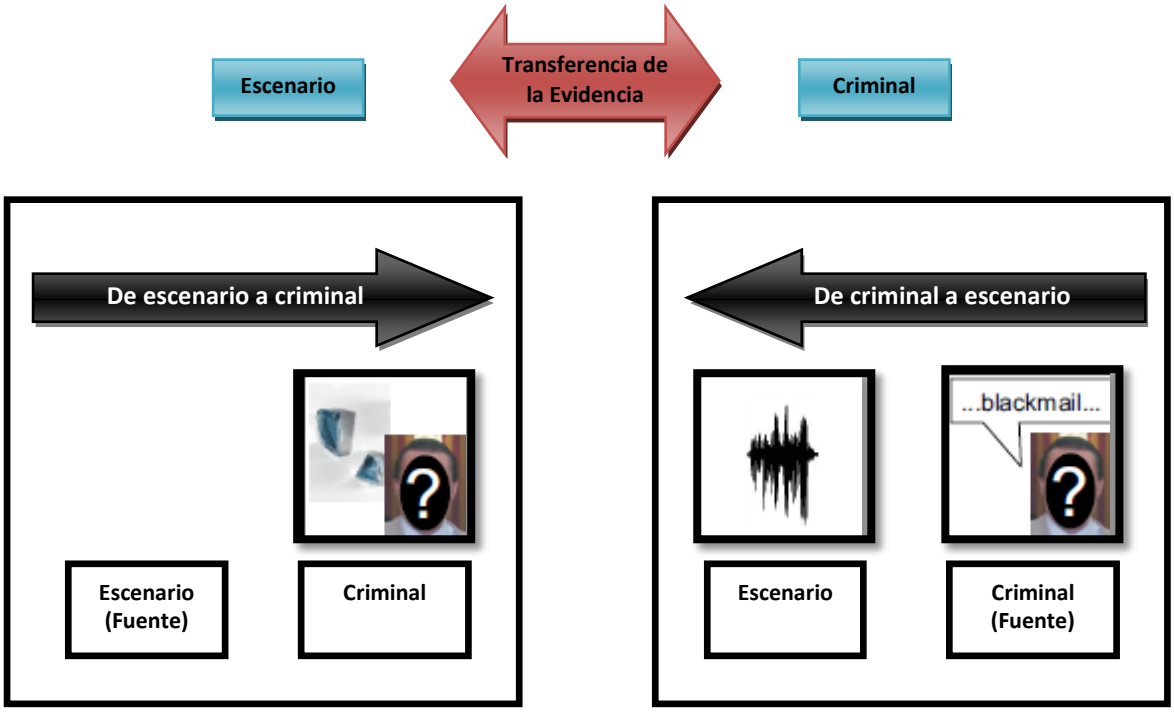


Figura 2-2. Transferencia de la evidencia [19].

La identificación forense es el proceso que busca conseguir la individualización. Aquí el rol del científico forense es la evaluación de la contribución de la evidencia forense de forma que pueda contestarse a la pregunta: *¿pertenece el material incriminatorio de origen desconocido a una fuente conocida?*

El término biometría se refiere a la identificación de un individuo basada en sus características distintivas, por lo que los sistemas biométricos actuales en la ciencia forense se orientan al filtrado de candidatos potenciales y a la verificación 1 a 1 realizada por un especialista forense entrenado en esa disciplina. Se dan los siguientes *casos típicos* [29]:

- Identificación de un individuo: Un conjunto de características biométricas pertenecientes a un individuo desconocido se compara con un conjunto de referencia de individuos conocidos.
- Identificación de la fuente: Un conjunto desconocido de características biométricas encontrado en circunstancias de interés para una investigación se compara con un conjunto de referencia de individuos conocidos basado en las características disponibles.

- Relación entre distintas marcas biométricas que pueden pertenecer a la misma fuente: Una comparación entre conjuntos desconocidos que resulta en una posible detección de incidentes relevantes.

En los dos últimos casos las características biométricas en estudio pueden haberse dejado en escenarios relevantes a una investigación. La investigación forense tiene como uno de sus objetivos principales encontrar marcas que asocien un individuo con un evento que está siendo investigado. Dichas marcas pueden haber sido dejadas por el sospechoso durante el evento o ser encontradas después en el sospechoso. Ejemplos de estas marcas son huellas de los dedos, orejas o pies tras el apoyo de los mismos en distintas superficies, así como grabaciones de teléfonos interceptados o cámaras de seguridad.

En este proyecto se propone el uso de reconocimiento automático de locutor para identificación forense [66] [67]. El reconocimiento automático de locutor, que detallaremos más adelante, se define como el uso de una máquina para reconocer personas a partir de la voz del locutor. Un sistema de reconocimiento automático de locutor es capaz de comparar locuciones, generando una medida de similitud entre ellas. En un caso forense que comprende locuciones de control y recuperadas, esta medida de similitud puede verse como evidencia forense.

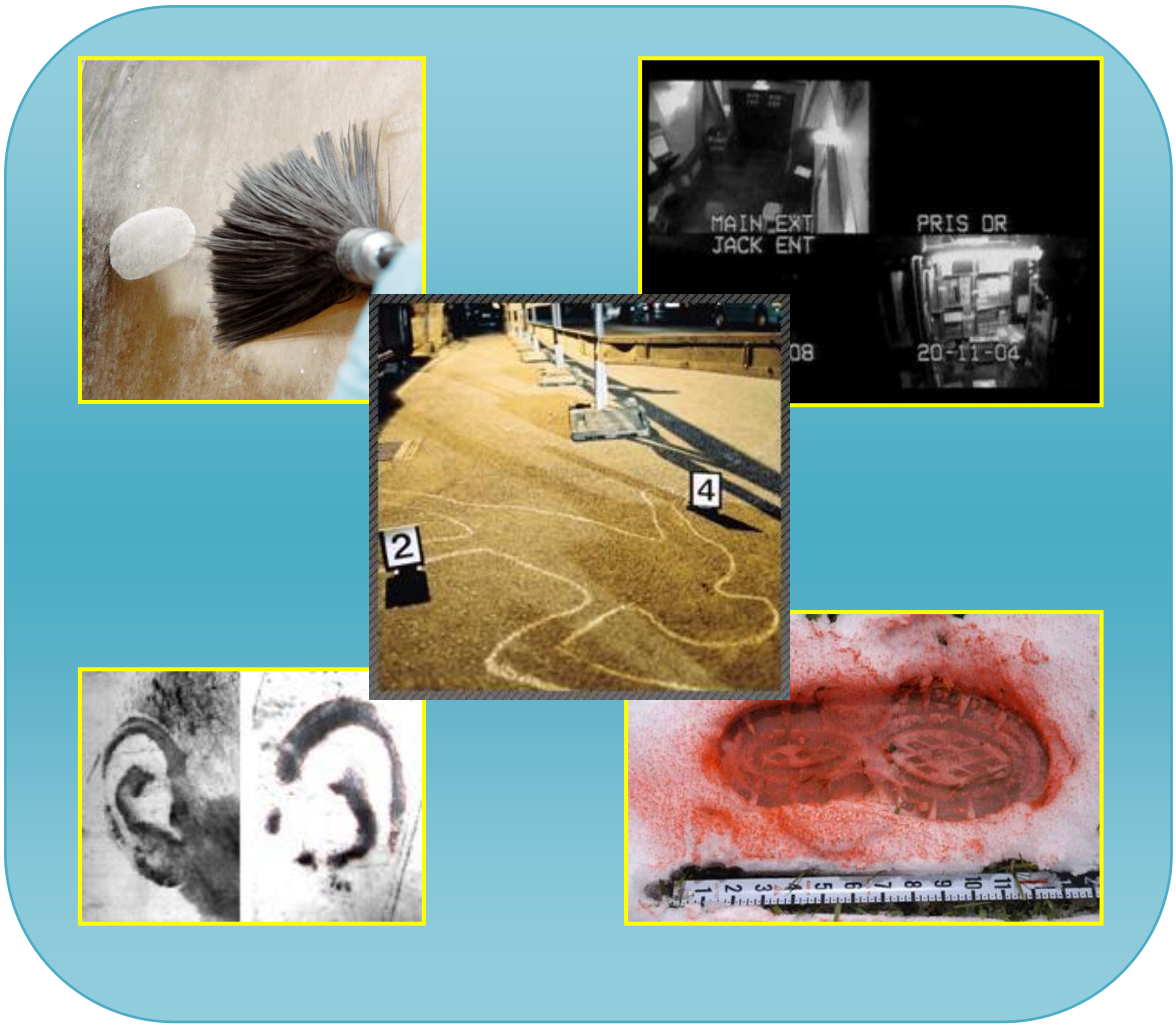


Figura 2-3. Ejemplos de marcas en escenario forense.

2.6. Funcionamiento de un Sistema de Reconocimiento Biométrico

Un sistema de reconocimiento biométrico es un sistema reconocedor de patrones que contiene un registro o base de datos de características extraídas a partir de rasgos biométricos y que, ante la entrada de un nuevo rasgo, procede a comparar sus características con las de la base de datos generando una puntuación que será comparada con un umbral preestablecido, dando lugar a la aceptación o rechazo del rasgo como perteneciente a un usuario [8, 17].

A continuación podemos ver un diagrama del modelo de un sistema general de identificación biométrica [8, 22], compuesto por cinco subsistemas:

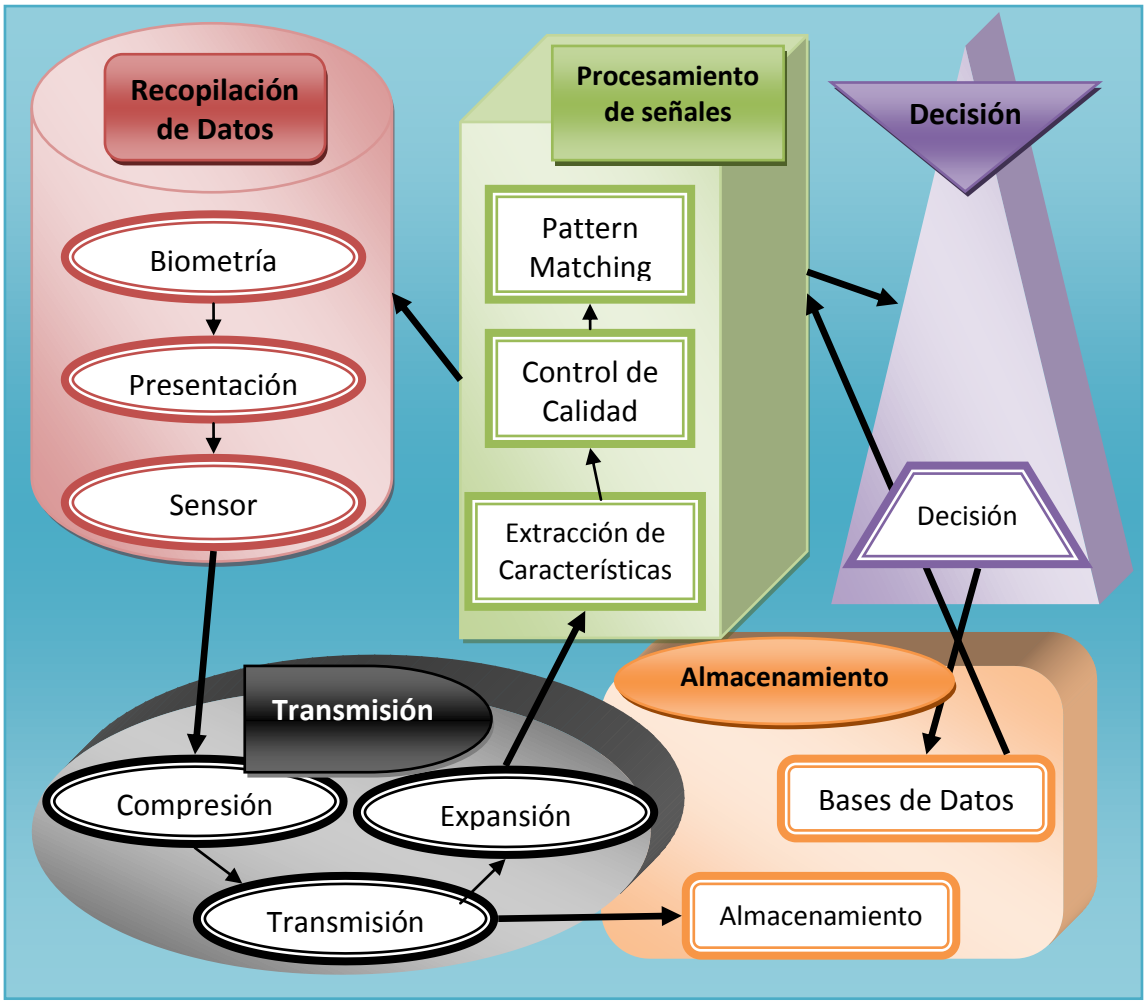


Figura 2-4. Aproximación de James L. Wayman de la estructura del procedimiento biométrico

2.7. Modos de Operación

Existen tres modos de operación en un sistema biométrico: registro, identificación y verificación. El modo registro es una fase previa común a los otros modos, en los que sí se considera que el sistema está en funcionamiento.

2.7.1. Sistemas de reconocimiento en modo registro

En este modo se procede a dar de alta a los usuarios en el sistema, para lo que se extrae la característica y se almacena en el sistema junto con la información del usuario, creando así la base de datos.

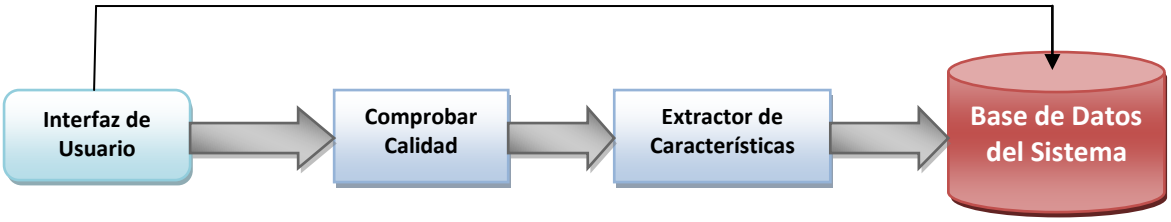


Figura 2-5. Diagrama de funcionamiento de un sistema biométrico en modo registro [16].

2.7.2. Sistemas de reconocimiento en modo identificación

El objetivo es clasificar una realización determinada de un rasgo biométrico de identidad desconocida como perteneciente a uno de entre un conjunto de N posibles individuos. Se diferencian dos posibles casos:

- Identificación en conjunto cerrado: el individuo a reconocer existe en el sistema, pertenece al grupo de usuarios del mismo, por tanto hay N posibles decisiones de salida.
- Identificación en conjunto abierto: puede que el individuo que queremos identificar no pertenezca al grupo de usuarios por lo que existe la posibilidad de no poder clasificar la realización de entrada como perteneciente a las N posibles.

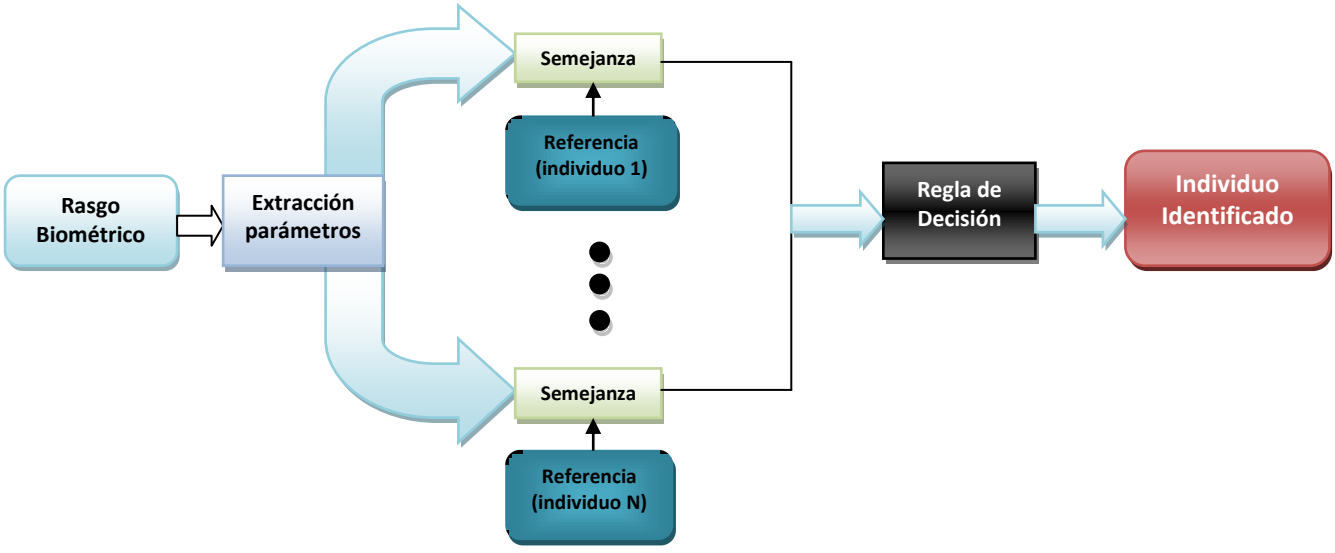


Figura 2-6. Diagrama de funcionamiento de un sistema biométrico en modo identificación.

2.7.3. Sistemas de reconocimiento en modo verificación

Sistemas que toman dos entradas (una realización del rasgo biométrico a identificar y una solicitud de identidad) y cuya salida puede ser aceptación o rechazo, es decir, clasificación del individuo como usuario auténtico o impostor.

La decisión de aceptar o rechazar la entrada depende de si el valor de parecido o probabilidad obtenido supera o no un determinado umbral de decisión.

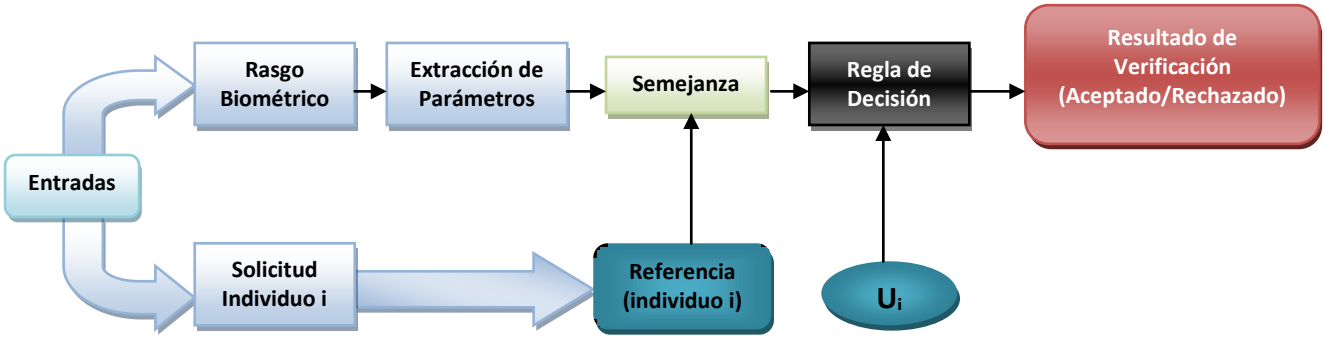


Figura 2-7. Diagrama de funcionamiento de un sistema biométrico en modo verificación.

2.8. Evaluación de Sistemas Biométricos

Evaluando un sistema estimamos sus aptitudes y su rendimiento, por tanto, para realizar una evaluación completa hay que tener en cuenta todos los aspectos, desde la adquisición de los datos hasta la integración del sistema [30]. Entre los puntos importantes a analizar destacan:

- Rendimiento del sistema respecto a su función (reconocimiento de personas).
- Seguridad, integridad y confidencialidad de los datos manejados por el sistema.
- Fiabilidad, disponibilidad y mantenimiento de la aplicación informática.
- Costes y beneficios del producto.
- Aceptación y/o facilidad de manejo por parte del usuario.
- Cuestiones legales.

El punto que se estudiará en este proyecto es el primero, por lo que nuestra evaluación de un sistema biométrico se centrará en el rendimiento del mismo.

Existen tres tipos de evaluación [22]:

- Evaluación tecnológica: es la más general, mide el estado de la tecnología, determina el progreso que ha logrado la misma e identifica los enfoques más prometedores.
- Evaluación de escenario: mide el rendimiento del sistema para un escenario prototipo que modela o simula un determinado campo de aplicación, para determinar si la tecnología cumple los requisitos de funcionamiento para esa aplicación.
- Evaluación operacional: realizada para un sistema concreto en un escenario prototipo, en un entorno de uso totalmente real y para una población determinada. El objetivo es analizar si el sistema biométrico cumple los requisitos de una determinada aplicación.

2.8.1. Factores que afectan al rendimiento

Podemos distinguir dos grandes grupos: los propios de la tecnología empleada y los ajenos a la misma que afectan en cierto grado a todos los sistemas. En el primer grupo destacan los factores ambientales y en el segundo el tiempo transcurrido entre la inscripción y la prueba, y la composición de la población incluida en el estudio (edad, género, origen étnico, número de usuarios,...). Los factores del primer grupo pueden verse en la tabla 3-1 y los del segundo se detallarán en el siguiente apartado.

	Iris	Caras	Huellas dactilares		Manos	Voz
			Sensor óptico	Sensor CMOS		
Luz ambiente	X	X	X		X	
Ruido ambiente						X
Temperatura			X	X	X	
Ruido electromagnético	X	X	X	X	X	X
Humedad ambiental			X	X		
Suciedad y contaminantes	X	X	X	X	X	
Variaciones de voltaje	X	X	X	X	X	X
Golpes y vibraciones	X	X	X	X	X	X

Tabla 2-3. Factores ambientales que afectan a cada tipo de sistema [30].

2.8.2. Adquisición de Datos

En cuanto a los datos sobre los que vamos a medir el rendimiento, debemos tener en cuenta algunos factores de carácter general:

- No se recomienda el uso de muestras creadas artificialmente ya que los resultados obtenidos no se podrán extrapolar a las condiciones reales de uso. Por ejemplo, no son convenientes las voces generadas de forma digital mediante un sintetizador. En cuanto

a las condiciones de adquisición no es aconsejable la simulación de ruido o de variaciones en la adquisición de la muestra.

- Hay que ser cuidadoso con los errores como dobles inscripciones, inconsistencias muestras-individuo, muestras incorrectas, etc.
- Minimizar la intervención humana, que puede añadir subjetividad y errores en la adquisición. Cuanto más automatizado esté el proceso más objetivos, libres de errores y cercanos a la situación real de uso serán los datos capturados.

En la adquisición de datos son importantes el entorno y la composición de la población. Ambos dependen del tipo de evaluación a realizar, como en nuestro caso es una evaluación tecnológica, nos interesa que sean suficientemente representativos y genéricos, como para poder probar de manera objetiva las capacidades reales de los distintos algoritmos.

También es importante el tiempo transcurrido entre la inscripción y la operación, ya que cuando mayor es, menor es el rendimiento del sistema. Esto es debido a que la mayoría de los rasgos biométricos sufren cambios con el tiempo.

2.8.3. Rendimiento del Sistema

Para evaluar y mejorar el rendimiento de un sistema biométrico necesitamos obtener una medida objetiva y cuantitativa del mismo. Con esta medida podremos comparar diferentes sistemas implementados en términos de rendimiento.

Los errores que comete un sistema biométrico pueden proporcionarnos una magnitud que defina el rendimiento del sistema de forma representativa. Son diferentes según el modo de operación, por lo que nos centraremos en los errores en verificación dado que es el modo de trabajo utilizado en este proyecto.

Para evaluar el rendimiento de un sistema en modo verificación, se utilizan dos medidas (tipos de errores) principales [1]:

- Probabilidad de Falsa Aceptación (P_{FA}): es la probabilidad de que se acepte erróneamente una muestra de entrada (impostor) como perteneciente a un individuo de la base de datos.
- Probabilidad de Falso Rechazo (P_{FR}): es la probabilidad de que el sistema rechace a un usuario genuino.

El punto en el que ambas probabilidades coinciden es la tasa EER (*Equal Error Rate*) y nos proporciona un único valor para definir el rendimiento del sistema.

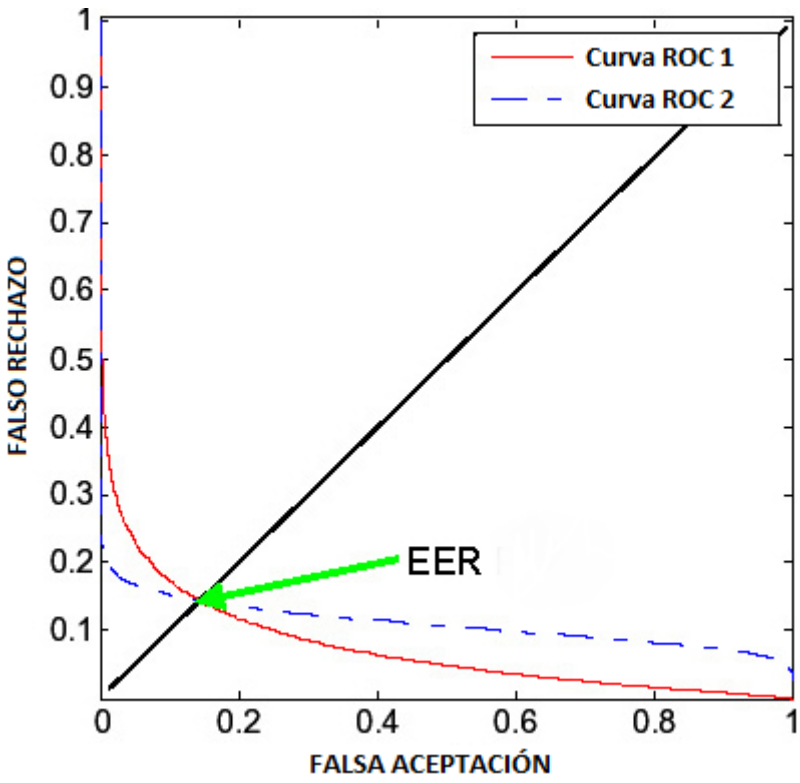


Figura 2-8. Ejemplo de obtención de *Equal Error Rate* [58].

Estos tipos de error dependen del umbral de decisión utilizado. Con un umbral de decisión bajo, el sistema tiende a aceptar a todos los usuarios dando lugar a pocos "falsos rechazos" y muchas "falsas aceptaciones". Por otra parte, con un umbral de decisión alto, el sistema rechazará a la mayoría de los usuarios por lo que se producirán muy pocas "falsas aceptaciones" y muchos "falsos rechazos".

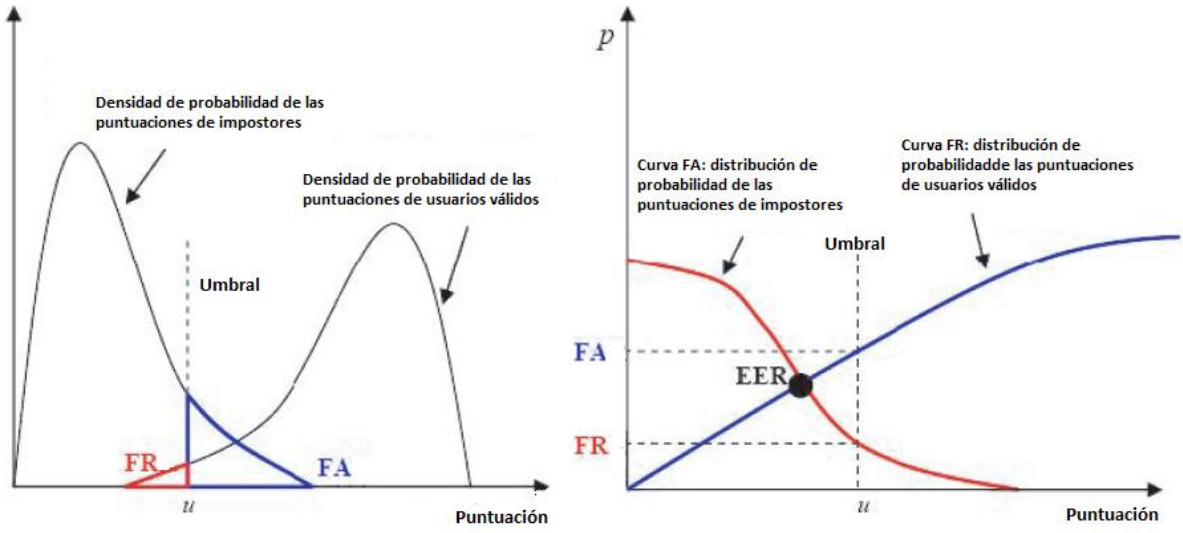


Figura 2-9. Densidades y distribuciones de probabilidad de usuarios legítimos e impostores [9].

De forma gráfica, podemos representar el rendimiento mediante dos tipos de curvas [1, 30]:

- Curvas ROC (*Receiver Operating Characteristics*): es una de las formas tradicionales de representar los errores en problemas de clasificación. Muestran la variación de la tasa de falsos positivos y de la tasa de verdaderos positivos con respecto al umbral de decisión.
- Curvas DET (*Detection Error Tradeoff*): curva monótona y decreciente que representa el rendimiento de un sistema dibujando P_{FA} como una función de P_{FR} en una escala de desviación normal. Permiten una comparación visual entre sistemas más clara y fácil de realizar. La distancia entre curvas expresa las diferencias entre rendimientos de manera más significativa. Cuanto más cerca está una curva del origen, más robusto será el sistema frente a errores de clasificación.

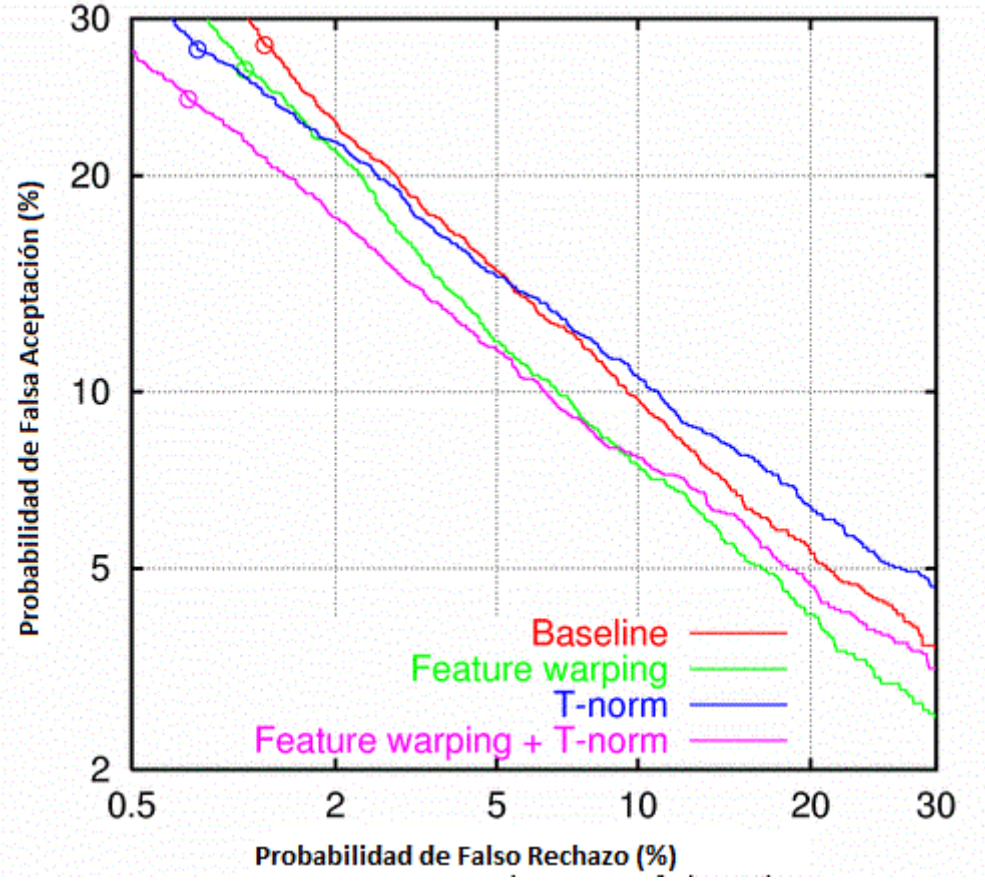


Figura 2-10. Ejemplo de curvas DET para diferentes sistemas [59].

Dibujar las tasas de error como función de un umbral es una buena forma de comparar el potencial de distintos métodos, pero no es suficientemente preciso para la evaluación de sistemas operativos para los que el umbral se define para operar en un punto dado. En este caso, los sistemas se evalúan de acuerdo con una función de coste que tiene en cuenta las dos tasas de error ponderadas con sus respectivos costes, esto es $C = C_{fa}P_{fa} + C_{fr}P_{fr}$.

2.9. Normalización de Puntuaciones

Como se ha mencionado en apartados anteriores, el proceso de decisión implica la comparación de la verosimilitud del modelo de locutor proclamado con la señal de habla entrante mediante un umbral de decisión. Si la verosimilitud es mayor que el umbral, aceptaremos al locutor. Este umbral es compartido por todos los usuarios.

La elección del umbral es problemática, ya que las puntuaciones entre pruebas conllevan una variabilidad que puede provenir de la naturaleza del material de entrenamiento, de la variabilidad intralocutor, canal de transmisión, entorno acústico, variabilidad interlocutor…

Para lidiar con esta variabilidad de las puntuaciones (*scores*), hacer más fácil la elección del umbral de decisión y reducir el desalineamiento entre las distribuciones *target* y *non target* existen las técnicas de normalización de puntuaciones.

La normalización es una técnica que transforma las puntuaciones de salida, de usuarios o de impostores, de forma que se proyecta la misma sobe una función densidad de probabilidad de media cero y varianza unidad, quedando localizadas las distribuciones de uno u otro tipo [38].

Dado que las varianzas se observan en las distribuciones de usuarios (intra-locutor) e impostores (inter-locutor), si normalizamos una de ellas podemos reducir la varianza general de la distribución en el sistema de verificación.

En este caso, vamos a normalizar los impostores, porque sus distribuciones son fáciles de computar utilizando pseudo-impostores y no tenemos las distribuciones de clientes que son las que representan la mayoría de la varianza de la distribución de puntuaciones. La forma de normalizar es la siguiente:

$$norm\big(Lx(y)\big) = \frac{Lx(y) - \mu}{\sigma}$$

Donde $Lx(y)$ es el *score raw*: la puntuación entre el modelo de locutor "x" y la señal de habla entrante "y", μ es la media de la distribución de puntuaciones de impostor y σ es la varianza de la misma.

A continuación se presentan las diferentes técnicas de normalización usadas en este proyecto:

- ***Zero Normalization*** [55]***:*** conocida como Z-Norm. En ella un modelo de locutor se enfrenta a un conjunto de ficheros de test pertenecientes a una población de impostores, resultando una distribución de puntuaciones de impostores. De esta distribución se extraen los parámetros de normalización (media y varianza) y se aplican a las puntuaciones generadas por el sistema de verificación mientras está funcionando. La ventaja principal de este método es que la estimación de los parámetros de normalización puede realizarse *off-line* durante el entrenamiento del modelo.

- ***Test Normalization*** [56]***:*** conocida como T-Norm. Además de enfrentar la señal de habla entrante (fichero de test) al modelo bajo estudio, se enfrenta también a una cohorte de modelos de para estimar la distribución de puntuaciones de impostor y los parámetros de normalización de forma consecutiva. Así se consigue un alineamiento de la distribución de probabilidad *non-target* dependiente del fichero de test a identificar. Si Z-Norm se considera dependiente de locutor, T-Norm es dependiente de test.

- ***ZT-Norm*** [57]***:*** combinación de normalización compuesta por la asociación entre una normalización de "condiciones de aprendizaje" (Z-Norm) y una normalización "basada en test" (T-Norm).

3. Sistemas de Reconocimiento Automático de Locutor

3.1. Introducción

Existe un gran número de rasgos biométricos y por tanto muchos sistemas biométricos basados en los mismos. Los sistemas más comunes se centran en huella, cara y voz. En ese proyecto nos dedicamos a sistemas basados en voz, concretamente en reconocimiento de locutor.

Existen dos factores principales que hacen la voz un rasgo biométrico convincente [1]: el hecho de que el habla es una señal producida de forma natural que los usuarios proporcionan sin considerarlo una amenaza, y el hecho de que los sistemas telefónicos proporcionan una red familiar de sensores para obtener y repartir la señal de voz. Además, el área de reconocimiento de locutor tiene una gran base científica con más de 30 años de investigación, desarrollo y evaluaciones.

Mientras que el área de reconocimiento de voz consiste en la extracción del mensaje textual incluido en una locución, el reconocimiento de locutor se basa en la extracción de la identidad de la persona que pronuncia la locución [2].

Como ya se ha explicado en el capítulo 3 de forma genérica para los sistemas de reconocimiento biométrico, existen diferentes tareas en el reconocimiento de locutor [31] que se diferencian en el tipo de decisión requerida en cada una: identificación, verificación y la más recientemente definida en el NIST, detección.

En identificación de locutor una muestra de voz procedente de un locutor desconocido se enfrenta a un conjunto de modelos de locutor conocidos.

En verificación de locutor comprobamos que un individuo es quien dice ser, por lo que enfrentamos la muestra de voz proporcionada únicamente con el modelo del locutor correspondiente a la identidad proclamada.

En detección de locutor debemos determinar si alguno de un conjunto de locutores conocidos se encuentra presente en una muestra de voz desconocida.

En este proyecto nos centramos en la tarea de verificación de locutor aplicada en ámbito forense, por lo que en ella estarán centrados los sistemas de los que hablaremos a continuación.

3.2. Información de identidad en la señal de voz

La voz es el rasgo biométrico que utilizamos principalmente en la comunicación entre personas, y su producción es un proceso muy complejo que depende de muchas variables a diferentes niveles, desde factores sociolingüísticos (como el nivel de educación, el contexto lingüístico y diferencias dialectales) hasta cuestiones fisiológicas (como la longitud del tracto vocal, su forma, y la configuración dinámica de los órganos articulatorios) [32].

En la señal de voz se encuentra codificada información que permite reconocer al locutor. El objetivo de los sistemas de reconocimiento automático de locutor es centrarse en las características de la señal de voz que permitan individualizar al hablante.

3.2.1. Tipos de Reconocedores

Según el tipo de información utilizada podemos distinguir dos tipos de reconocedores:

- Reconocedores de locutor de alto nivel: las características de alto nivel incluyen la percepción de las palabras y su significado, la sintaxis, la prosodia, dialecto e idiolecto. Por tanto, estos reconocedores se centran en la información procedente de la fase de generación del mensaje en el cerebro. Son sistemas más robustos frente a los efectos de canal y ruido, pero las características son más difíciles de extraer.

- Reconocedores de locutor de bajo nivel: basados en información característica procedente de la fase de producción de voz, es decir, información del espectro de la señal como ancho de banda, periodicidad de tono o frecuencia de resonancia. Son características independientes de texto y lenguaje, más fáciles de extraer y con menor cantidad de datos necesaria para el entrenamiento, pero se ven afectadas por el ruido.

En la figura 3-1 podemos ver un resumen de las características desde un punto de vista diferente, su interpretación física. La elección de las características debe basarse en su poder discriminativo, robustez y practicidad.

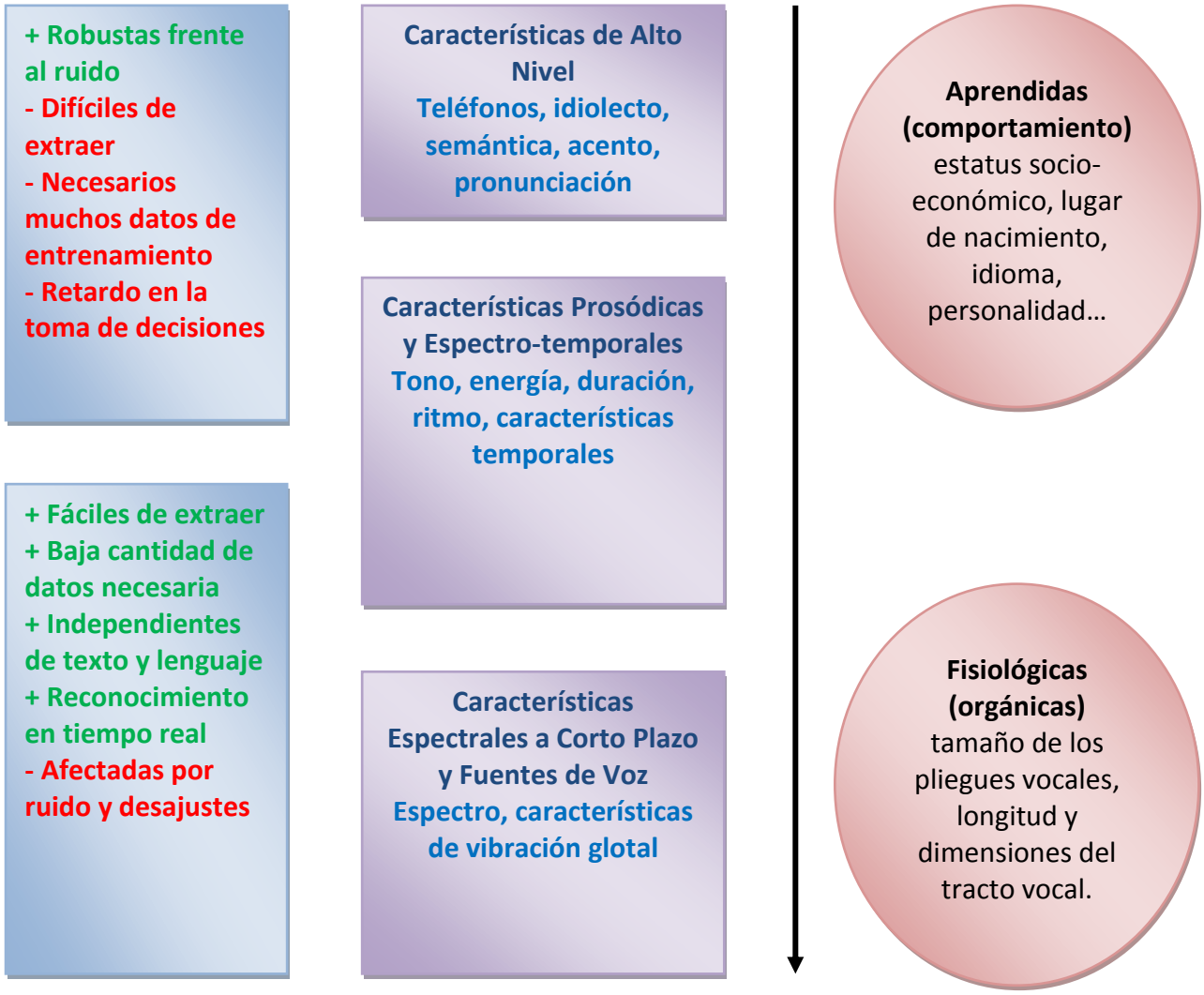


Figura 3-1. Resumen de características desde el punto de vista de su interpretación física [3].

3.2.2. Niveles de Identidad

Dentro de los niveles vistos en el apartado anterior, existen otros subniveles de identidad que permiten realizar una clasificación más precisa de las particularidades de la voz. Del más alto al más bajo, estos niveles son: lingüístico, fonético, prosódico y acústico.

- Nivel lingüístico: características que describen la forma en que el locutor hace uso del lenguaje, que se ven influenciadas por aspectos como la educación, el origen y las condiciones sociológicas del hablante.
- Nivel fonético: está formado por fonemas y secuencias de fonemas.
- Nivel prosódico: lo compone la prosodia, que es una combinación de energía, duración y tono de los fonemas. Es responsable de dotar a la voz de naturalidad y sentido.

- Nivel acústico: en él se encuentran las características espectrales a corto plazo de la señal de voz. Estas características están relacionadas con los órganos que intervienen en la generación del habla.

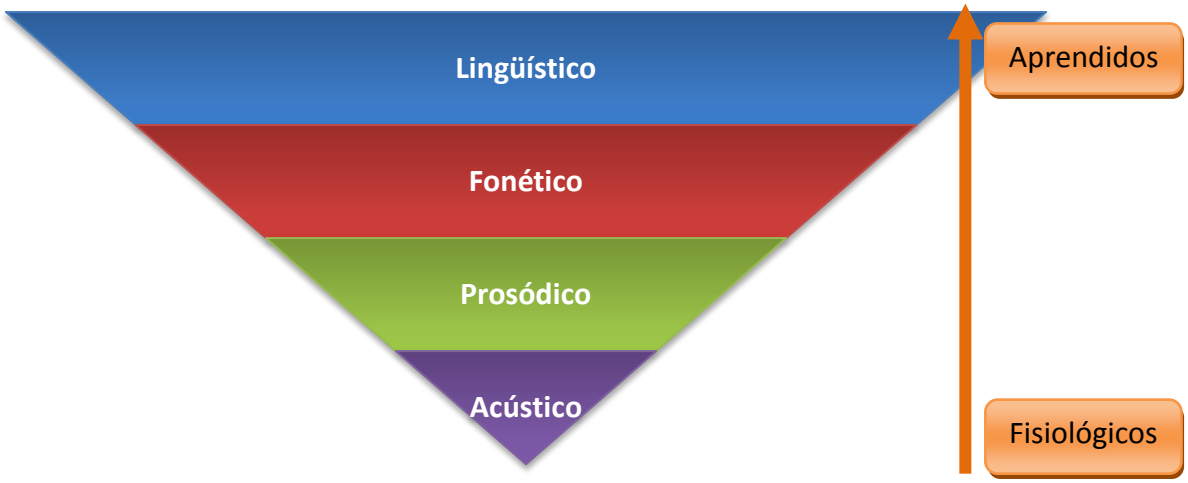


Figura 3-2. Niveles de identidad

3.3. Descripción de un Sistema de Verificación de Locutor

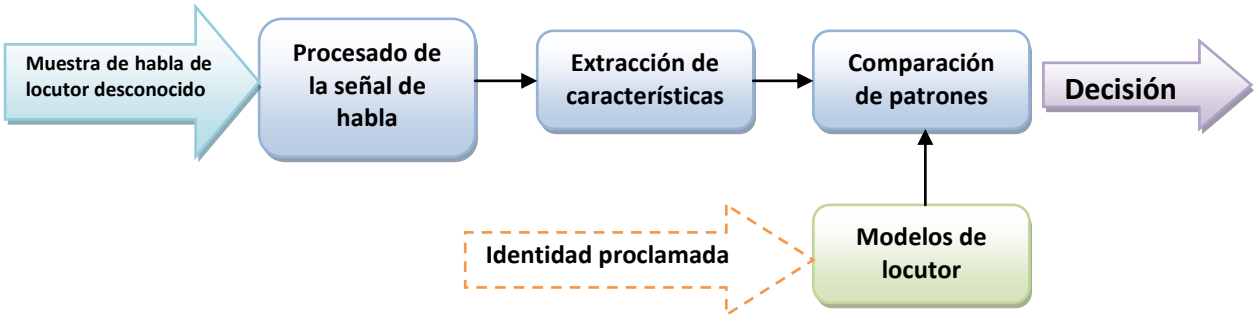


Figura 3-3. Diagrama de bloques de un sistema de reconocimiento de locutor [31].

En la figura 3-3 podemos observar un diagrama de bloques con los elementos básicos de un sistema genérico de reconocimiento de locutor. La entrada del sistema es una muestra de habla de un locutor desconocido. En el caso de verificación de locutor también es una entrada al sistema una proclamación de identidad.

Un sistema de verificación de locutor se compone de dos fases: entrenamiento y test.

En la fase de entrenamiento se extraen los parámetros de la señal de habla para obtener una representación ajustable a un modelado estadístico y después se obtiene el modelo estadístico para los parámetros. Este esquema de funcionamiento que podemos ver en la figura 3-4 se utiliza también en el entrenamiento de modelos *background*.

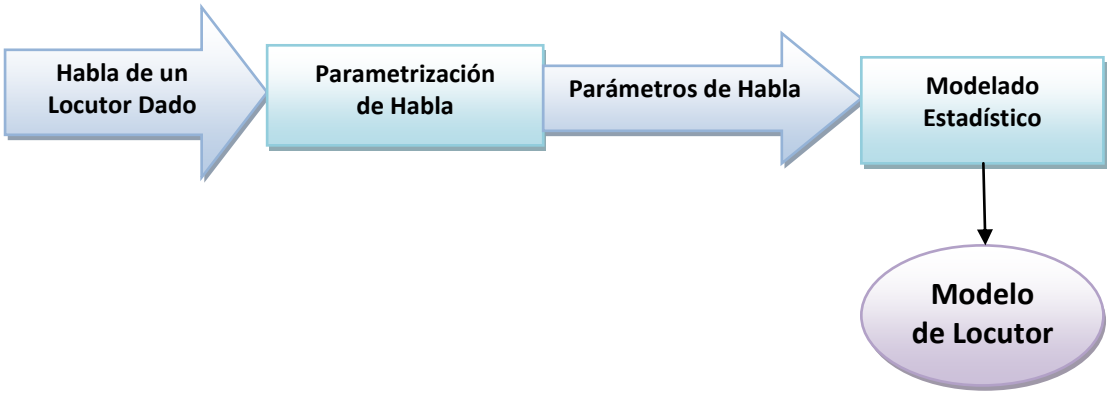


Figura 3-4. Diagrama de bloques de la fase de entrenamiento de un sistema de verificación de locutor [1].

En la fase de test las entradas del sistema son una solicitud de identidad y las muestras de habla de un locutor desconocido. Primero se extraen los parámetros de la señal de habla, y luego el modelo de locutor correspondiente a la identidad solicitada y un modelo *background* son extraídos de un conjunto de modelos estadísticos calculados durante la fase de entrenamiento. Por último se calculan las puntuaciones, se normalizan y se toma la decisión de aceptación o rechazo.

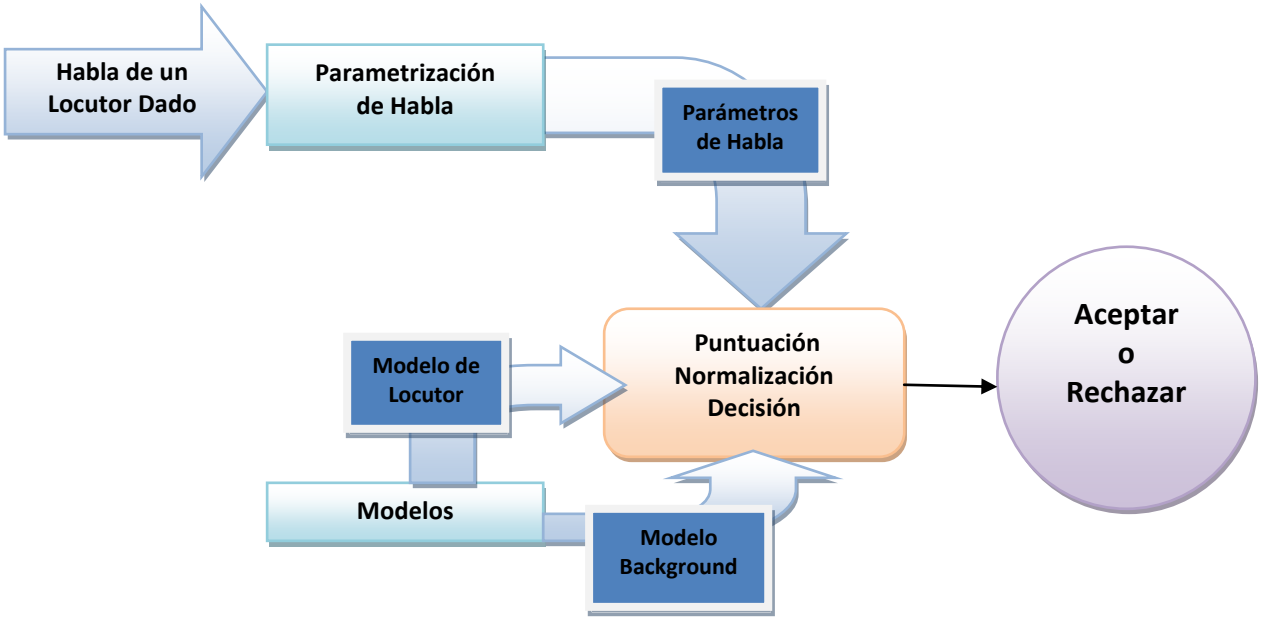


Figura 3-5. Diagrama de la fase de test de un sistema de verificación de locutor [1].

3.4. Extracción de Características

Como se ha visto anteriormente en el apartado 3.2, la señal de voz contiene información de alto y bajo nivel. Los módulos de parametrización de habla del sistema de verificación de locutor se encargan de la tarea de extracción de características, puesto que la parametrización es la conversión de la señal en un conjunto de vectores de características.

Con esta transformación se pretende conseguir una nueva representación de la información más compacta, menos redundante y más útil de cara al modelado estadístico y al cálculo de puntuaciones. A continuación se presentan algunos de los métodos más comunes en la caracterización del locutor.

3.4.1. Coeficientes LPCC (Linear Predictive Cepstral Coefficients)

El análisis LPC se basa en un modelado lineal de la producción de habla. Una muestra de habla posee fuerte correlación con la consecutiva, por lo que puede estimarse cada muestra como combinación lineal de las anteriores.

El proceso es el siguiente: la envolvente de la ventana bajo análisis es estimada mediante un filtro de predicción lineal y se realiza la transformada Cepstral de los coeficientes de dicho filtro generando coeficientes transformados, parte de los cuales serán el vector de parámetros de la ventana.

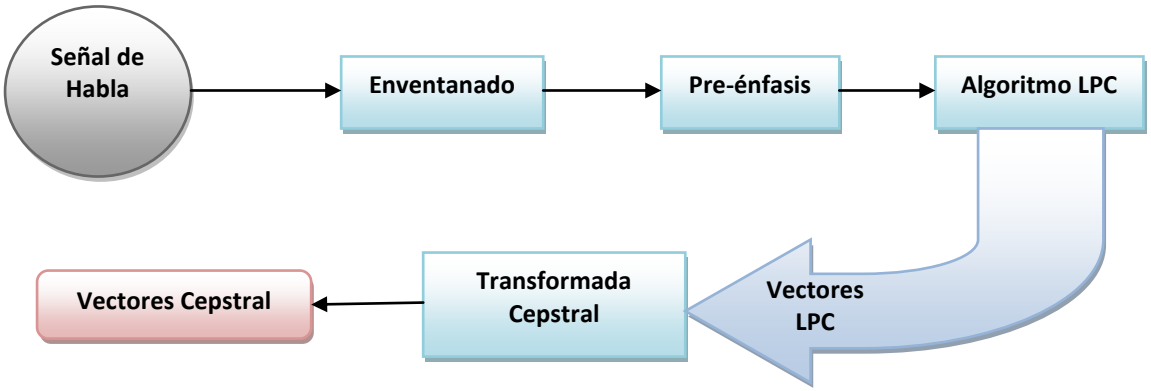


Figura 3-6. Diagrama de bloques de una parametrización cepstral basada en LPC. Figura adaptada de [1].

3.4.2. Coeficientes MFCC (Mel-Frequency Cepstral Coefficients)

En el caso de los MFCCs no se modela la envolvente sino que se extrae una serie de coeficientes procedentes de un banco de filtros *Mel* (escala basada en la percepción logarítmica del oído humano). La transformada cepstral de dichos coeficientes genera unos coeficientes transformados, parte de los cuales serán el vector de parámetros de la ventana.

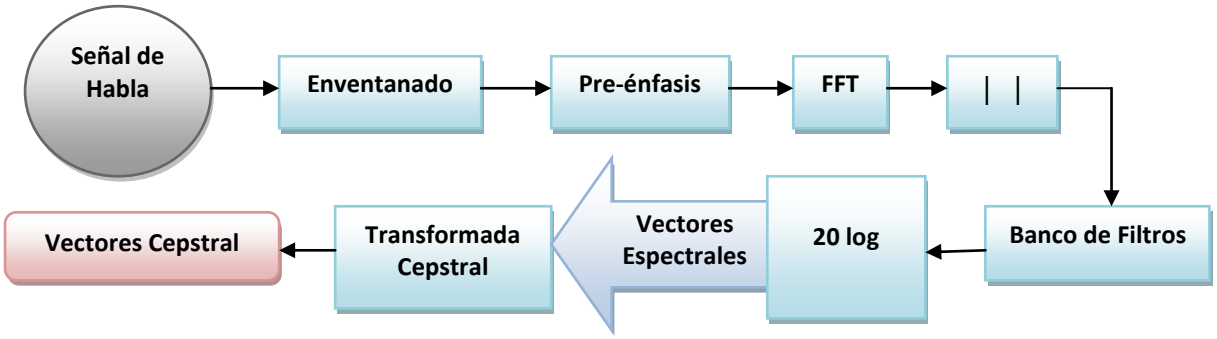


Figura 3-7. Diagrama de bloques de una parametrización cepstral basada en banco de filtros. Figura adaptada de [1].

Los MFCCs son muy populares en procesamiento de audio y habla, y son las características que utilizaremos en el transcurso de este proyecto para procesar nuestros archivos. Por ello, procedemos a explicar su obtención con más detalle a continuación.

Como podemos observar en el diagrama, primero se aplica a la señal una ventana de duración temporal menor a la de la señal original (ventanas de 20-30 ms con retardo de 10 ms entre las mismas).

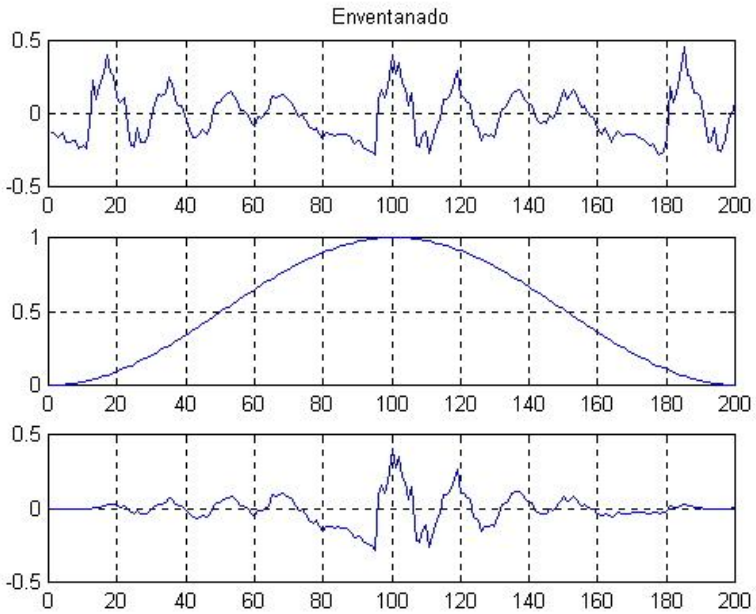


Figura 3-8. Ejemplo de enventanado [60].

Generalmente, en reconocimiento de locutor se utilizan ventanas de Hamming o de Hanning para conseguir un lóbulo principal estrecho así como lóbulos secundarios bajos. Tras el Enventanado la señal quedaría de la siguiente forma:

$$x(n) = s(n) \cdot w(m-n), \qquad n\,\epsilon\,[m-N+1, m]$$

Donde $s(n)$ es la señal de voz original, $w(n)$ es la ventana y N es la duración de la misma.

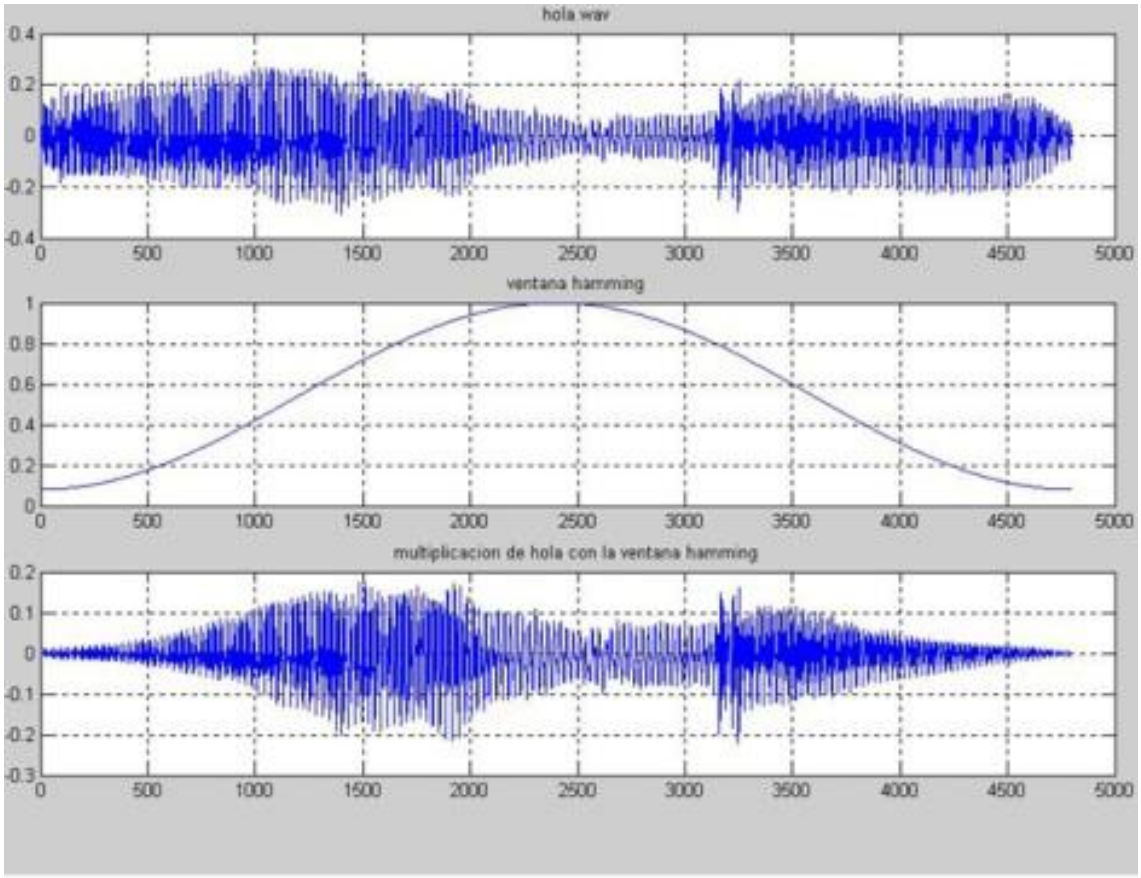


Figura 3-9. Enventanado de señal de audio con ventana Hamming.

Después, la señal enventanada se pre-enfatiza mediante la aplicación de un filtro que aumenta las altas frecuencias del espectro. El filtro utilizado es el siguiente:

$$x(t) = x(t-1) - a \cdot x(t-2), \qquad a\,\epsilon\,[0.95, 0.98]$$

Cabe destacar que el pre-énfasis no es obligatorio y por tanto no siempre se utiliza el filtro definido.

Posteriormente se calcula la transformada rápida de Fourier (FFT o *Fast Fourier Transform*) de la señal obtenida (ya sea tras el enventanado o tras el pre-énfasis) con un número de puntos N

que debe ser potencia de 2 y mayor que el número de puntos en la ventana. N suele tomarse igual a 512 puntos. Una vez tenemos la transformada hallamos su módulo.

Con el objetivo de suavizar el espectro (dado que sólo nos interesa su envolvente y que queremos reducir el tamaño de los vectores espectrales) lo multiplicamos por un banco de filtros *Mel* cuyas frecuencias centrales vienen dadas por:

$$f_{MEL} = 1000 \cdot \frac{log\left(1 + \frac{f_{LIN}}{1000}\right)}{\log 2}$$

Por último calculamos los vectores espectrales tomando el logaritmo de la envolvente espectral y multiplicando por 20 cada coeficiente (para obtener decibelios).

Aplicando la DCT (*Discrete Cosine Transform*) a los vectores espectrales obtenemos los coeficientes cepstral de la siguiente forma:

$$c_n = \sum_{k=1}^{K} S_k \cdot \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 1, 2, \dots, L$$

Donde K es el número de coeficientes log-espectrales calculados previamente, S_k son los coeficientes log-espectrales y L es el número de coeficientes cepstral que queremos calcular, siendo $L \leq K$.

A partir de los coeficientes cepstral también es posible incorporar información dinámica en los vectores, de forma que sepamos cómo varían en el tiempo. Eso se consigue con los coeficientes Δ y ΔΔ, que son aproximaciones polinomiales de las derivadas de primer y segundo orden y dependen de la velocidad y aceleración con las que varían los coeficientes cepstral.

3.5. Técnicas Empleadas en Reconocimiento de Locutor

El reconocimiento de locutor tiene multitud de aplicaciones, en el ámbito de la seguridad (como controles de acceso), servicios personalizados de usuario, aplicaciones forenses, aplicaciones de vigilancia, comerciales… Y todas ellas pueden llevarse a cabo mediante dos sistemas de reconocimiento de locutor: dependientes de texto e independientes de texto.

3.5.1. Reconocimiento de Locutor Dependiente de Texto

El reconocimiento de locutor dependiente de texto utiliza el contenido léxico del habla para llevar a cabo su función [32]. Caracteriza una tarea, como verificación o identificación, en la que el conjunto de palabras (léxico) utilizado durante la fase de test es un subconjunto de las palabras utilizadas en la fase de entrenamiento [33]. Es por esto que los sistemas que emplean estos reconocedores son más fáciles de engañar.

En los sistemas de reconocimiento de locutor dependientes de texto, el solapamiento entre las fases de entrenamiento y test permite una gran precisión con una cantidad limitada de datos (menos de 8 segundos de habla).

Sus principales aplicaciones son comerciales: controles de acceso, aparatos electrónicos, marcación por voz, domótica… En ellas el texto que hay que repetir suele ser una clave de usuario.

Los reconocedores de locutor dependientes de texto pueden utilizar diferentes técnicas de modelado, pero la más común es HMM (*Hidden Markov Models* o Modelos Ocultos de Markov) de la que hablaremos en el siguiente apartado.

3.5.2. Reconocimiento de Locutor Independiente de Texto

En los reconocedores de locutor independientes de texto el léxico utilizado en entrenamiento y el utilizado en la fase de test son totalmente diferentes. Las aplicaciones basadas en este tipo de reconocedores tienen el desafío adicional de operar con poco o ningún control sobre el comportamiento del usuario [2], ya que éste no es cooperativo como en el caso dependiente de texto. Es necesario utilizar técnicas de compensación de la variabilidad de los entornos acústicos y los canales.

Los reconocedores de este tipo pueden utilizarse en las mismas aplicaciones que los dependientes de texto, pero son mucho más comunes en aplicaciones forenses de identificación y verificación de locutor. En general, son más utilizados que los anteriores ya que el hecho de no restringir la locución a un texto concreto los hace válidos para un mayor número de aplicaciones.

En este tipo de reconocedores se utilizan los siguientes clasificadores: GMM (*Gaussian Mixture Models* o modelos de mezclas Gaussianas), SVM (*Support Vector Machines* o máquinas de vectores soporte) y supervectores (híbridos GMM-SVM).

3.6. Modelos y Clasificadores

Un sistema de reconocimiento de locutor toma una decisión de autentificación de individuos en función de unos parámetros que tienen que ser extraídos de las señales de voz. Para ello existen sistemas que realizan modelos de locutor a partir de las características de las locuciones de forma que puedan compararse con modelos de los archivos de test, obteniendo así la puntuación necesaria para la toma de decisiones.

En este apartado se realizará una revisión del estado del arte en modelado estadístico, viendo los diferentes esquemas y su utilidad actual.

3.6.1. Detección de la Razón de Verosimilitud

Dado un segmento de habla, Y, y un locutor hipotético, S, en verificación de locutor debemos determinar si Y ha sido dicho por S. Nos encontramos con dos hipótesis:

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

La prueba óptima para decidir entre las dos hipótesis es un test de la razón de verosimilitud:

$\frac{p\,(Y \mid H_0)}{p\,(Y \mid H_1)}$, donde si es mayor que un umbral de decisión (Θ) se acepta H_0 y si es menor que Θ se aceptará H_1, siendo $p\,(Y \mid H_0)$ la función densidad de probabilidad para la hipótesis H_0 evaluada para el segmento de habla observado Y, ó verosimilitud de la hipótesis H_0 dado el segmento de habla. De la misma forma se define $p\,(Y \mid H_1)$.

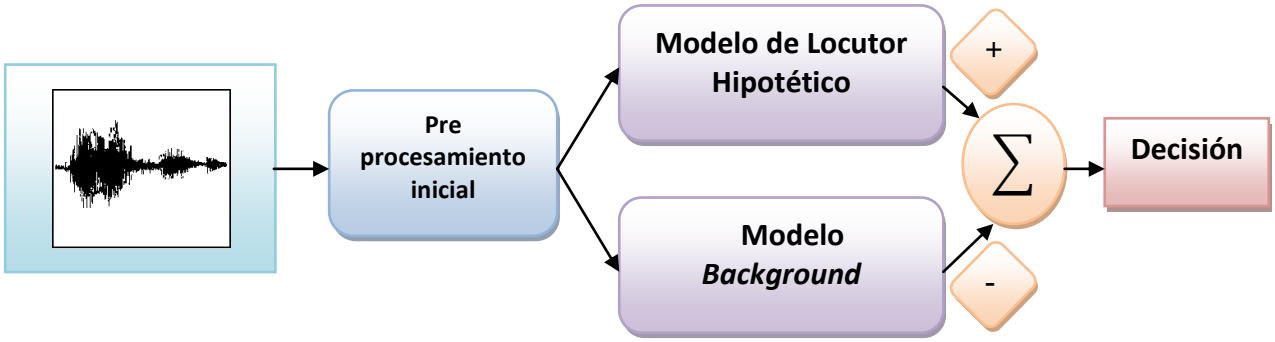


Figura 3-10. Sistema de verificación de locutor basado en razón de verosimilitud [1].

El procesamiento de interfaz extrae características de la señal de habla que portan información dependiente del locutor, y en él se utilizan además técnicas para minimizar efectos confusos de esas características. La salida de esta etapa es una secuencia de vectores de características que representan el segmento de test, $X = \{\overrightarrow{x_1}, \overrightarrow{x_2}, \dots, \overrightarrow{x_T}\}$ donde $\overrightarrow{x_t}$ es un vector de características con $t \in [1, 2, \dots, T]$. Estos vectores de características se utilizan para calcular las verosimilitudes de H_0 y H_1. El modelo λ_{hyp} representa H_0, que caracteriza al locutor hipotético S en el espacio de características de $\vec{x}$,y el modelo $\lambda_{\overline{hyp}}$ la hipótesis alternativa, H_1; por lo que la razón de verosimilitud estadística queda como $\dfrac{p\,(X \mid \lambda_{hyp})}{p\,(X \mid \lambda_{\overline{hyp}})}$.

El logaritmo de este estadístico nos proporciona la razón de verosimilitud logarítmica:

$\Lambda(x) = \log p(X \mid \lambda_{hyp}) - p\,(X \mid \lambda_{\overline{hyp}})$.

3.6.2. Modelos de Mezclas Gaussianas (GMM)

El modelo de mezclas Gaussianas parte de la base de que mediante la combinación de Gaussianas podemos reproducir cualquier distribución de probabilidad por complicada que sea. Es un modelo estocástico que se ha convertido en el método de referencia por excelencia en el reconocimiento de locutor. Un paso importante para implementar el detector de la razón de verosimilitud vista en el apartado 3.6.1 es la selección de la función verosimilitud $p(X|\lambda)$. Cuando disponemos de características continuas en reconocimiento de locutor independiente de texto, donde no hay conocimiento a priori de lo que el locutor va a decir, la función de verosimilitud más exitosa es GMM [1, 2, 7].

Un GMM puede considerarse una extensión del modelo VQ en el que las celdas se solapan. En este caso el vector de características no se asigna a la celda más cercana pero tiene una probabilidad de haberse originado en cada celda [3].

Un modelo de mezclas Gaussianas (figura 3-14) se compone de una mezcla finita de componentes Gaussianas multivariadas. Denotando el GMM como λ, podemos caracterizarlo mediante su función densidad de probabilidad ó verosimilitud:

$$p\,(\overrightarrow{x} \mid \lambda) = \sum_{k=1}^{K} P_k \; N\,(\overrightarrow{x} \mid \vec{\mu}_k, \vec{\Sigma}_k)$$

Donde K es el número de componentes Gaussianas, P_k es la probabilidad prior de la componente k-ésima ó peso de la mezcla (que satisface $\sum_{k=1}^{K} P_k = 1$ y $P_k \geq 0$) y $N\,(\overrightarrow{x} \mid \vec{\mu}_k, \vec{\Sigma}_k)$ es la función densidad, una combinación de densidades Gaussianas uni-modales cada una parametrizada por un vector media $\vec{\mu}_k$ y una matriz de covarianzas $\vec{\Sigma}_k$:

$$N\,(\overrightarrow{x} \mid \vec{\mu}_k, \vec{\Sigma}_k) = \frac{1}{2\pi^{D/2} |\vec{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\overrightarrow{x} - \vec{\mu}_k)^T \, \vec{\Sigma}_k \, (\overrightarrow{x} - \vec{\mu}_k)}$$

El vector media tiene dimensión Dx1 y la matriz de covarianzas (que suele utilizarse en forma diagonal por motivos computacionales) DxD, siendo D la dimensión del vector de características $\overrightarrow{x}$.

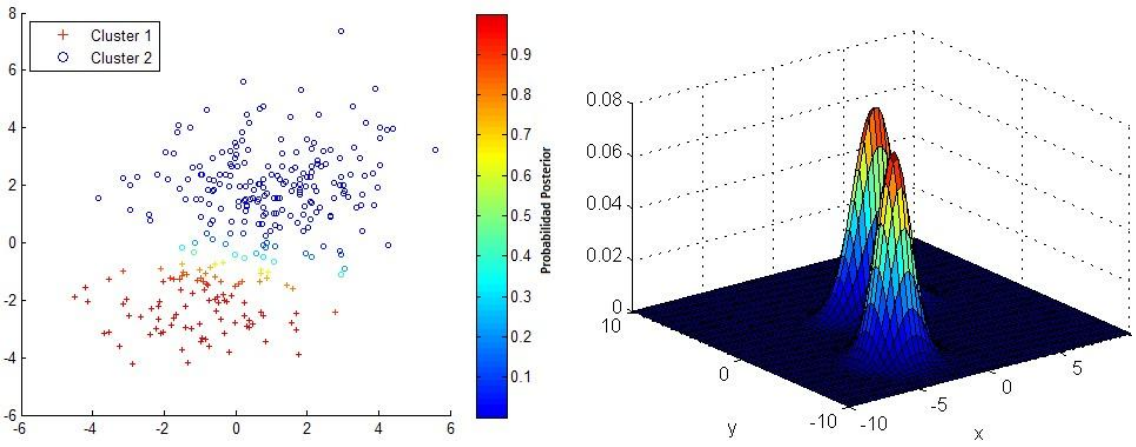


Figura 3-11. Distribución espacial de los coeficientes espectrales y GMM entrenado a partir de la misma [68].

Una vez tenemos los vectores de entrenamiento, los parámetros del modelo de máxima verosimilitud se estiman utilizando el algoritmo iterativo de Maximización de la Esperanza (EM o *Expectation-Maximization*). Éste algoritmo va refinando los parámetros del GMM de forma iterativa para aumentar la verosimilitud del modelo estimado para los vectores de características observados, hasta que llegamos a una convergencia.

Entrenar un GMM consiste en la estimación de los parámetros $\lambda = \{P_k, \overrightarrow{\mu_k}, \overrightarrow{\Sigma_k}\}$ de una muestra de entrenamiento, para lo que la aproximación básica es la estimación de máxima verosimilitud. Para trabajar de forma más cómoda computacionalmente, se utiliza la verosimilitud logarítmica, que, asumiendo independencia entre vectores de características, se calcula de la siguiente forma:

$$\log p(X\,|\lambda) = \frac{1}{T}\sum_{t} \log p\,(\overrightarrow{x_t}|\lambda)$$

Donde $p(\overrightarrow{x_t}|\lambda)$ es la verosimilitud vista anteriormente.

En las aplicaciones de voz es importante adaptar los modelos acústicos a las nuevas condiciones de operación debido a la variabilidad de los datos causada por diferentes locutores, entornos, estilos de habla y demás. En el reconocimiento de locutor basado en GMM se entrena primero un modelo universal o *background* (UBM) con el algoritmo EM a partir de datos de cientos de horas de habla obtenidos de un elevado número de locutores. El UBM representa una distribución de vectores de características independiente de locutor. Cuando se registra un nuevo locutor en el sistema, los parámetros del UBM se adaptan a la

distribución de características de este nuevo locutor. El modelo adaptado se utiliza como el modelo de ese locutor. El proceso de adaptación (generalmente mediante estimación MAP) se lleva a cabo como un proceso de estimación en dos pasos:

En el primer paso las estimaciones de los estadísticos suficientes de los datos de entrenamiento del locutor se computan para cada mezcla en el UBM.

En el segundo paso los "nuevos" estadísticos suficientes se combinan con los "viejos" de los parámetros de mezcla del UBM usando un coeficiente de mezcla dependiente de los datos, que está diseñado de tal forma que las mezclas con altas cuentas de datos del locutor se tienen más en cuenta en las estimaciones de los "nuevos" estadísticos suficientes y las mezclas con bajas cuentas de datos del locutor se tienen más en cuenta en los "viejos" estadísticos suficientes para la estimación final de parámetros.

Veamos cómo se obtienen los vectores adaptados: dado un modelo *background* y vectores de entrenamiento para el locutor hipotético, determinamos el alineamiento probabilístico para los mismos en los componentes de mezcla del UBM:

$$P(k|\overrightarrow{x_t}) = \frac{P_k N\left(\overrightarrow{x_t}\,|\vec{\mu}_k, \vec{\Sigma}_k\right)}{\sum_{m=1}^{K} P_m N\left(\overrightarrow{x_t}\,|\vec{\mu}_m, \vec{\Sigma}_m\right)}$$

Utilizando los datos que tenemos podemos calcular los estadísticos suficientes para los parámetros de peso, media y varianza:

$$n_k = \sum_{t=1}^{T} P(k|\overrightarrow{x_t})$$

$$E_k(\vec{x}) = \frac{1}{n_k}\sum_{t=1}^{T} P(k|\overrightarrow{x_t})\,\overrightarrow{x_t}$$

$$E_k(\vec{x}^2) = \frac{1}{n_k}\sum_{t=1}^{T} P(k|\overrightarrow{x_t})\,\overrightarrow{x_t}^2$$

Estos nuevos estadísticos suficientes se utilizan para actualizar los viejos para la mezcla "k" y crear así parámetros adaptados de la siguiente forma:

$$\boldsymbol{w_k} = \left[\frac{\alpha_k n_k}{T} + (1-\alpha_k) w_k\right]\gamma$$

$$\overrightarrow{\boldsymbol{\mu_k}} = \alpha_k E_k(\vec{x}) + (1-\alpha_k)\overrightarrow{\mu_k}$$

$$\overrightarrow{\sigma_k}^2 = \alpha_k E_k(\vec{x}^2) + (1-\alpha_k)\left(\overrightarrow{\sigma_k}^2 + \overrightarrow{\mu_k}^2\right) - \overrightarrow{\boldsymbol{\mu_k}}^{\boldsymbol{2}}$$

El factor de escala γ se computa sobre todos los pesos de mezcla adaptados para asegurarse de que su sumatorio es la unidad. El coeficiente de adaptación que controla el balance entre las estimaciones viejas y las nuevas se define como:

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

Siendo r un factor de relevancia que se fija previamente y controla el efecto de las muestras de entrenamiento en el modelo resultante con respecto al UBM.

El hecho de utilizar un coeficiente de adaptación dependiente de los datos (figura 3-15) permite adaptar los parámetros de forma dependiente de la mezcla. El factor de relevancia es un modo de controlar cuántos datos nuevos deben observarse en una mezcla antes de que los nuevos parámetros comiencen a reemplazar a los viejos.

En el modo de reconocimiento, el modelo adaptado mediante MAP y el UBM están emparejados, por lo que el reconocedor suele conocerse comúnmente como GMM-UBM (*Gaussian Mixture Model – Universal Background Model*). La puntuación depende tanto del modelo *target* como del modelo *background* a través de la razón media de verosimilitud logarítmica:

$$\Lambda_{avg}(X, \lambda_{target}, \lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T}\{\log p(x_t|\lambda_{target}) - \log p(x_t|\lambda_{UBM})\}$$

Esta razón mide la diferencia entre los modelos *target* y *background* en la generación del conjunto de observaciones.

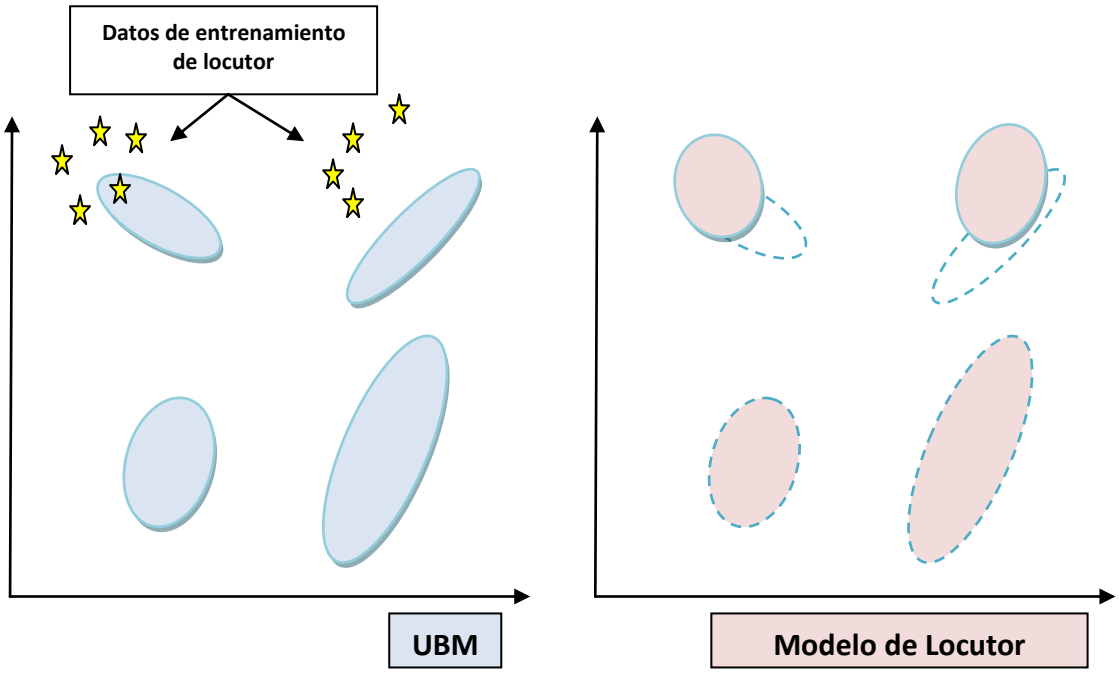


Figura 3-12. Ejemplo de adaptación de un modelo de locutor. Adaptada de [7].

En la figura 3-12 podemos observar a la izquierda cómo los vectores de entrenamiento se mapean probabilísticamente en las mezclas del UBM. A la derecha los parámetros de mezcla adaptados se derivan utilizando los estadísticos de los nuevos datos y los parámetros de

mezcla del UBM. Por tanto puede observarse que es un proceso de adaptación dependiente de los datos.

Las ventajas de usar un GMM como función verosimilitud son que no tiene coste computacional, que está basado en un modelo estadístico bien conocido, y que, para tareas independientes de texto, es insensible a los aspectos temporales del habla, modelando sólo la distribución subyacente de las observaciones acústicas de un locutor.

3.6.3. Máquinas de Vectores Soporte (SVM)

A diferencia de GMM que es generativo, SVM es un clasificador discriminativo (modela el límite entre un locutor y un conjunto de impostores) muy potente que ha comenzado a utilizarse cada vez con mayor frecuencia en los últimos tiempos [2,3]. El objetivo de un SVM es clasificar patrones asignándolos a su clase correspondiente.

Estos clasificadores se han aplicado con éxito utilizando tanto características espectrales como prosódicas o de alto nivel. También se han combinado con GMM para incrementar la precisión, obteniendo buenos resultados.

Como podemos ver en la figura 3-13, un SVM modela la frontera de decisión entre dos clases como un hiperplano. De forma más intuitiva, en la figura 3-14 se observa la transformación sufrida en el espacio de entrada tras modelar su espacio de características mediante SVM.

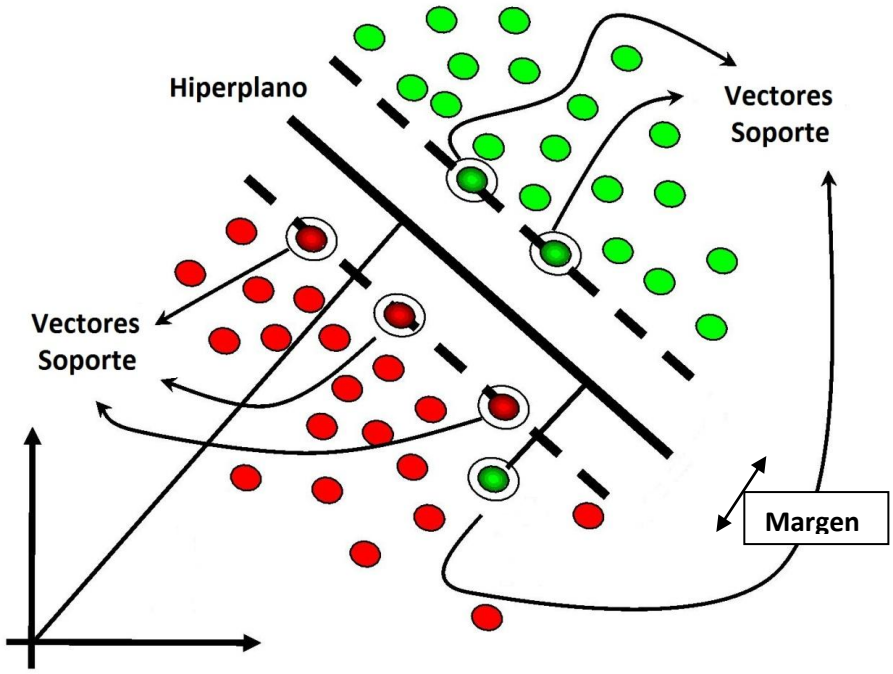


Figura 3-13. Ejemplo de vectores soporte [62].

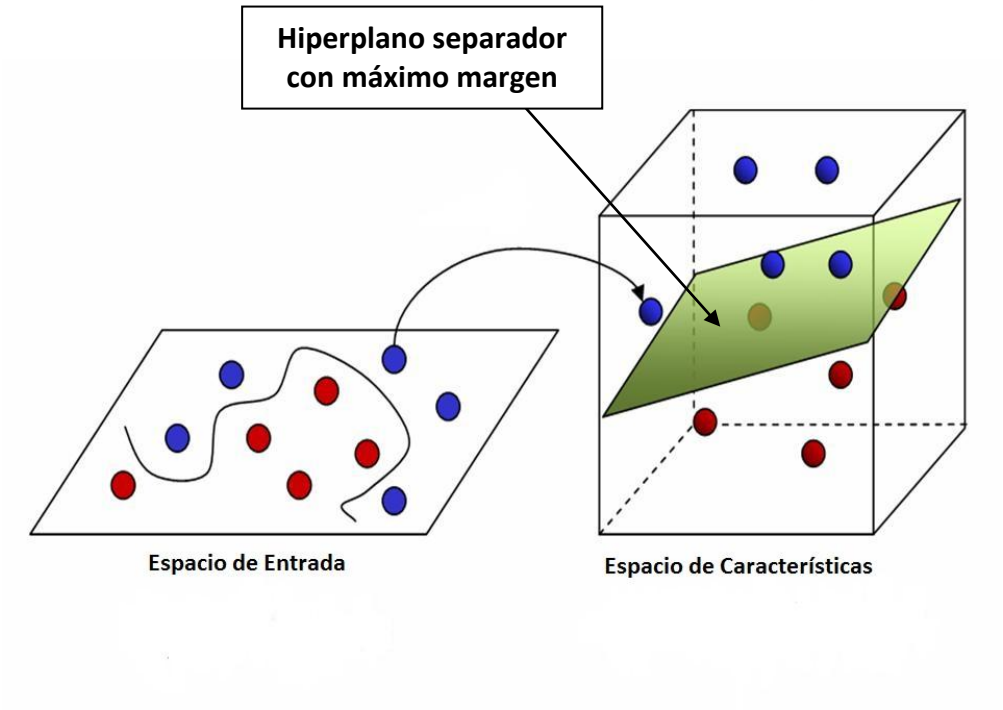


Figura 3-14. Ejemplo de SVM visto de forma espacial [63].

En verificación de locutor una de las clases está formada por los vectores de entrenamiento del locutor *target* y la otra por los vectores de entrenamiento de una población de impostores (*background*). El optimizador SVM encuentra un hiperplano que maximiza el margen de separación entre las dos clases.

De forma teórica, un SVM modela dos clases usando sumatorios de una función *kernel*:

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

En la ecuación t_i son las salidas ideales (-1 ó 1 dependiendo de la clase del vector soporte correspondiente), $\sum_{i=1}^{N} \alpha_i t_i = 0$, y $\alpha_i > 0$. $\vec{x}_i$ son los vectores soporte y se obtienen del conjunto de entrenamiento mediante un proceso de optimización.

Del resultado de la función $f(\vec{x})$ comparado con un umbral obtenemos la decisión de clasificación.

En el ámbito de la verificación de locutor tenemos que tener en cuenta que SVM es un clasificador de dos clases por lo que entrenamos un modelo *target* para el locutor y lo enfrentamos contra un conjunto de locutores de muestra que tengan características de la población de impostores (este conjunto es el *background*).

3.7. Variabilidad de Sesión

Para un locutor dado, los supervectores estimados de diferentes locuciones de entrenamiento no suelen ser son los mismos prácticamente en ninguna ocasión, especialmente cuando esas muestras de entrenamiento han sido extraídas de diferentes canales. Este hecho da lugar a la variabilidad de sesión, de la cual podemos diferenciar dos tipos: variabilidad intersesión y variabilidad intrasesión.

Definimos la variabilidad intersesión como la diferencia existente entre las diferentes muestras de habla asociadas a una misma identidad. Esta variabilidad es una de las principales causantes de degradación de los sistemas de reconocimiento de locutor [4].

La variabilidad intrasesión es la diferencia observable dentro de una misma muestra perteneciente a un locutor.

Además, podemos encontrar otro tipo de variabilidad que también afecta a los sistemas de reconocimiento de locutor: la variabilidad de locutor, que podemos dividir en variabilidad interlocutor e intralocutor.

La variabilidad interlocutor se define como la diferencia entre muestras idénticas pertenecientes a distintos individuos mientras que la variabilidad intralocutor es la diferencia que observamos en una misma locución pronunciada por el mismo locutor en diferentes situaciones.

Este proyecto se centra principalmente en la variabilidad intersesión, a la que también llamaremos variabilidad de canal o variabilidad de sesión. Dicha variabilidad es un reto desde los inicios del reconocimiento de locutor. Para enfrentarnos a la degradación generada por la misma en los sistemas se ha demostrado que es útil modelar la variabilidad de sesión [49]. Una forma de modelarla es compensando en canal y locutor.

La compensación es necesaria para asegurarse de que los datos de prueba y/o entrenamiento obtenidos de distinto canal pueden ser apropiadamente enfrentados entre sí. Existen técnicas que modelan explícitamente la variabilidad de canal consiguiendo así una compensación efectiva [4].

Algunas de estas técnicas, orientadas a minimizar el efecto del ruido y/o perturbaciones introducidas en el canal de transmisión comprenden CMN (*Cepstral Mean Normalization*), filtrado RASTA, FW (*Feature Warping*), FM (*Feature Mapping*) y JFA *(Joint Factor Analysis*).

3.7.1. *Cepstral Mean Normalization*

CMN es una técnica que trata de suprimir el efecto del canal en el dominio *cepstral* mediante la eliminación de la media ponderada de cada coeficiente calculada a partir de la locución [50]. Sustrayendo el vector media los conjuntos de características obtenidos de diferentes canales llegan a media cero y el efecto de canal se reduce. De forma similar, las varianzas de las características pueden ser igualadas dividiendo cada una entre su desviación estándar. Se asume que el efecto de canal es constante durante toda la locución [3].

3.7.2. Filtrado RASTA

El filtrado RASTA (RelAtive SpecTrAl filtering) consiste en la aplicación de un filtro paso banda en el domino *cepstral*, bajo la hipótesis de que cualquier componente de variación demasiado lenta o rápida no pertenece a la señal de habla [51]. Es un método independiente de la señal, al contrario que CMN.

3.7.3. *Feature Warping*

Método que pretende modificar la distribución de características a corto plazo para seguir una distribución de referencia. Para ello, se ajusta la función de distribución acumulada de las características para que coincida con la de referencia (por ejemplo, una gaussiana de media cero y varianza unidad) [52].

3.7.4. *Feature Mapping*

FM es un método de normalización supervisado que transforma un espacio de características dependiente de canal en uno independiente de canal de forma que se reduzca la variabilidad del mismo. Para lograrlo se utiliza un conjunto de GMMs dependientes de canal adaptados de un modelo independiente de canal [53].

3.7.5. *Joint Factor Analysis*

Técnica desarrollada recientemente que modela las direcciones de máxima variabilidad interlocutor e intralocutor de las características extraídas de la señal de habla [54]. Reduce de forma significativa la influencia del canal en las locuciones y en ella se basa este proyecto fin de carrera, por lo que el capítulo 4 versa en su totalidad sobre JFA.

4. Técnicas Basadas en *Factor Analysis* Aplicado al Reconocimiento de Locutor

4.1. Introducción

En los últimos años, los sistemas de verificación de locutor independiente de texto pertenecientes al *estado del arte* se han centrado en el uso de técnicas basadas en *Factor Analysis*. Estas técnicas han ganado terreno a otras gracias a su habilidad para tratar con la variabilidad de sesión.

En este capítulo analizaremos el modelo *Joint Factor Analysis* (en adelante JFA) cuyas características principales proponen una nueva forma de lidiar con la variabilidad y marcan la diferencia entre JFA y otras aproximaciones existentes antes de su aparición. A continuación detallaremos estas ideas básicas de JFA:

- Se considera la variabilidad, tanto de sesión como de locutor, como una fuente continua. Esto es así porque inicialmente se consideraba como discreta dado que se disponía de menor cantidad de datos, pero actualmente si nos paramos a pensar en cuántos tipos de transductores, canales de transmisión, situaciones conversacionales o entornos existen, son imposibles de enumerar. Por ello es más lógico tratar la variabilidad como continua y no como discreta.

- El modelo es capaz de modelar de forma explícita e independiente la variabilidad de sesión (inter-sesión) y de locutor (inter-locutor), objetivo que hasta JFA no se había conseguido puesto que no había una frontera clara entre ambas variabilidades y se trataba simplemente de compensar los efectos de sesión.

- Gran parte de la variabilidad continua se encuentra restringida en un subespacio de dimensionalidad mucho menor que el espacio del modelo. Esta idea, que partió como una hipótesis y fue demostrada empíricamente con posterioridad, marca el desarrollo del modelo JFA.

4.2. Bases del Modelo *Joint Factor Analysis*

4.2.1. Eigenvoices

En el campo del reconocimiento de habla y de locutor nos referimos a los vectores que mejor representan los componentes de variación más importantes entre locutores como *eigenvoices*.

El modelo fue presentado para lidiar con el problema de la adaptación de locutor en las tareas de reconocimiento de habla en las que había muy pequeñas cantidades de datos específicos de locutor [45], como reconocimiento de dígitos o de letras. Aproximaciones que utilizaban algoritmos como MAP o MLLR [3] fallaban para cantidades de datos disponibles muy limitadas.

Introduciendo un subespacio de variabilidad de locutor de baja dimensionalidad previamente entrenado (espacio de locutor) la cantidad de parámetros libres para ser estimados mediante adaptación de locutor se reduce drásticamente y se puede alcanzar una adaptación robusta incluso para pequeñas cantidades de datos disponibles. Como contrapartida, si la estimación del espacio de locutor no representa la variabilidad de locutor de forma apropiada, el sistema fallará.

La aproximación de *eigenvoices* se introdujo en el campo de la identificación y verificación de locutor como un reemplazo a la adaptación MAP en modelos GMM en los casos en los que la cantidad de datos de entrenamiento de locutor disponibles eran escasos [46]. El método *eigenvoice MAP* se resume de la siguiente forma: cada locutor está representado por un punto en un espacio de alta dimensionalidad, su supervector media. Dado un conjunto de estos supervectores media, el espacio de locutor se entrena estimando las direcciones de máxima variación del conjunto de datos. Una vez estimado, los locutores, tanto de entrenamiento como de test, se representan en este espacio, y la puntuación se lleva a cabo mediante el cálculo de las distancias de las representaciones de subespacio de locutores.

Posteriormente se consideró una distribución de probabilidades *a priori* para el supervector de locutor dentro de la estimación de modelo *eigenvoice* [47]. Esta aproximación se conoce como *eigenvoice MAP* porque se utiliza una estimación *maximum posteriori* en lugar de *maximum likelihood*.

4.2.3. Eigenchannels

Bajo la idea de adaptar un modelo de locutor a un canal dado de la misma forma que un modelo independiente de locutor se adapta a un locutor dado, la aproximación *eigenchannel MAP* consta exactamente de los mismos principios que la *eigenvoice MAP* salvo que en este caso se necesita un subespacio intersesión [48].

Eigenchannel MAP se utilizó como una técnica para enfrentarse a la variabilidad intersesión en el momento del reconocimiento. Una vez los modelos de locutor habían sido adaptados de un modelo independiente de locutor (UBM) a un locutor target a través de MAP clásico o *eigenvoice MAP*, fueron adaptados sucesivamente a los efectos de sesión de cada locución de prueba. Así, el modelo target se desplaza al tipo de canal específico de la locución de test, evitando el posible desalineamiento de canal entre las locuciones de entrenamiento y test.

4.2.4. *Joint Factor Analysis*

En este apartado se realizará una breve descripción del modelo JFA, para descripciones más detalladas del mismo existen publicaciones de carácter científico como [47] [48] [65].

La cuestión principal a resolver tras los estudios anteriores consiste en modelar de forma separada la variabilidad de sesión y de locutor. Para ello JFA tiene en cuenta la hipótesis de que un supervector media de locutor está formado por dos componentes, una relacionada con la variabilidad de locutor y otra con la variabilidad intersesión.

El modelo JFA considera la variabilidad de un supervector Gaussiano como una combinación lineal de las componentes de locutor y canal. Comenzaremos analizando la componente dependiente de locutor.

Dado un sistema GMM clásico con C componentes y F dimensiones, y un modelo UBM independiente de locutor entrenado previamente con supervector de medias μ (de dimensión CFx1), tras la adaptación MAP clásica la forma del supervector de medias dependiente de locutor para el locutor *s* es:

$$\mu_s = \mu + Dz_s$$

Donde el término Dz_s representa el desplazamiento del offset de la media μ como resultado de la adaptación MAP, y está formado por una matriz diagonal D de dimensiones CFxCF y el vector z_s de dimensiones CFx1.

Análogamente, pero teniendo en cuenta que la varianza de la distribución para los supervectores media μ_s está restringida en un subespacio del espacio supervector, obtenemos:

$$\mu_s = \mu + Vy_s$$

Donde V es una matriz de bajo rango y dimensión CDxR que contiene la varianza del locutor, e y_s son los pesos que representan al locutor s en el subespacio de variabilidad de locutor expandido por V. Variando y_s, el espacio completo de supervector puede sufrir modificaciones en las direcciones determinadas por las columnas de V.

JFA une ambas ideas de forma que se deriva una componente del modelo supervector de medias de locutor dependiente de locutor como:

$$\mu_s = \mu + Vy_s + Dz_s$$

En esta expresión la varianza del conjunto de μ_s se explica ahora por V y D, y aunque la variabilidad de locutor se supone restringida en el espacio extendido por V, el término Dz_s permite posibles desplazamientos en todas las direcciones del espacio supervector.

En JFA las componentes del vector y_s suelen llamarse *speaker factors* (factores de locutor) porque representan la variabilidad de locutor dentro del subespacio de variabilidad de locutor V.

Una vez establecida la componente dependiente de locutor, analizaremos la componente dependiente de sesión del supervector media del locutor. En JFA se 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* de la siguiente forma:

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

Donde U tiene el mismo papel que V, representando al subespacio de variabilidad de sesión en lugar del de locutor. La componente de x_h son *channel factors* (factores de canal) y, a diferencia de los factores de locutor, dependen estrictamente de la locución *h* en lugar del locutor.

Resumiendo, JFA se representa de forma matricial como sigue:

$$\mu_{sh} = \mu + Vy_s + Dz_s + Ux_h$$

En la siguiente tabla podemos observar una descripción de cada componente del modelo JFA.

Término	Descripción	Dimensiones
μ	Media de los nuevos modelos. Suele ser el supervector media de locutor del UBM.	CFx1
V	Subespacio de variabilidad de locutor. Matriz de bajo rango.	$CFxR_s$
D	Término residual de locutor. Matriz diagonal.	CFxCF
U	Subespacio de variabilidad de sesión. Matriz de bajo rango.	$CFxR_c$
$\mathbf{y_s}$	Factores de locutor.	R_sx1
$\mathbf{z_s}$	Término residual de locutor.	CFx1
$\mathbf{X_h}$	Factores de canal.	R_cx1

Tabla 4-1. Descomposición del modelo JFA.

5. Bases de Datos y Protocolos de Evaluación

En este capítulo detallaremos las bases de datos utilizadas en el proyecto, así como los protocolos experimentales seguidos para la realización de las diferentes pruebas, en las que debemos utilizar bases de datos de condiciones similares a las utilizadas en la fase de funcionamiento del sistema, para que el comportamiento del mismo sea coherente y obtengamos tasas de error similares.

Las bases de datos utilizadas en este proyecto componen el conjunto de datos necesarios para el funcionamiento del sistema de reconocimiento de locutor. Estos datos son los requeridos para las fases de desarrollo, entrenamiento y evaluación. Es indispensable tener una gran cantidad de los mismos, así como que recojan la mayor variabilidad posible tanto de locutores como de condiciones de habla.

Los protocolos son pautas y medidas objetivas normalizadas que permiten saber de qué forma hemos llegado a los resultados obtenidos, de manera que puedan compararse con los de otros experimentos realizados en sistemas de reconocimiento de terceros.

5.1. Bases de Datos para Reconocimiento de Locutor

En la elaboración de las pruebas experimentales se han utilizado como bases de datos principales las obtenidas de diferentes evaluaciones NIST SRE y la base de datos forense Ahumada.

Con estas bases de datos conseguimos el objetivo propuesto de recoger la mayor cantidad de factores de variabilidad posible de la señal de voz así como una gran población de locutores. A continuación se procede a detallar las diferentes bases de datos utilizadas en la elaboración del proyecto [19]:

- **AHUMADA:** base de datos registrada por el grupo ATVS que contiene habla en español grabada de la línea telefónica y dos tipos de micrófonos bajo condiciones controladas. El habla ha sido recolectada de casos forenses reales. Contiene variabilidad en el estilo hablado, desde texto leído hasta habla espontánea. Los detalles del corpus de Ahumada pueden encontrarse en [35].

  En la elaboración de este proyecto se ha utilizado concretamente la base de datos **Ahumada III**, compuesta por habla de 61 locutores y registrada utilizando procedimientos y sistemas de la Guardia Civil española [5]. Esta base de datos contiene habla obtenida de llamadas vía teléfono móvil realizadas en España, con variedad en el origen geográfico de los locutores, distintos tipos de dialectos en español y diferentes

condiciones acústicas y emocionales. No existe variación en el género puesto que todos los locutores son masculinos.
En esta base de datos, todos los locutores cuentan con 2 minutos de habla (obtenida de la llamada telefónica) útiles para entrenar el modelo de locutor. Además, existen 10 segmentos de habla para 31 locutores y otros 5 segmentos para 30 locutores, todos de diferentes llamadas, incluidos para los archivos de test. Estos segmentos tienen entre 7 y 25 segundos de habla, con una duración media de 13 segundos.

- **Switchboard 1:** contiene habla conversacional en inglés americano, grabada sobre línea telefónica convencional. En la grabación no se ha considerado la variabilidad dialectal. Sin embargo, contiene habla grabada de diferentes líneas telefónicas y diferentes tipos de terminales como teléfonos de carbón o de tipo electret. Estas fuentes de variabilidad han demostrado afectar de forma importante al desarrollo del reconocimiento.

- **Switchboard 2:** esta base de datos contiene habla conversacional en inglés americano grabada de la línea de teléfono convencional. Como la anterior, contiene variabilidad relativa a diferentes líneas y terminales, pero en un grado superior. Esta base de datos también contiene variabilidad dialectal que fue grabada en tres fases cada una con un contenido dialectal diferente: Fase 1 (inglés americano del medio-Atlántico), Fase 2 (inglés americano del medio-Oeste) y Fase 3 (inglés americano del Sur).

- **MIXER y datos adicionales multilenguaje:** MIXER presenta tres diferencias fundamentales con respecto a las distintas versiones de Switchboard. Primero, la variabilidad de canal y terminal telefónico es significativamente mayor, incluyendo habla grabada sobre teléfonos inalámbricos y redes móviles, en terminales de diversos tipos como manos libres o teléfonos convencionales. En segundo lugar, es multilenguaje, incluyendo habla en inglés americano, español, árabe, chino mandarín y ruso. Por último, se utilizó un protocolo *Fisher* para aleatorizar las conversaciones en la base de datos. Además, para las evaluaciones NIST SRE 2005 y 2006 se grabó una cantidad significativa e nuevos datos de habla de los lenguajes mencionados siguiendo el mismo protocolo, incluyendo variación de dialecto y hablantes no nativos. Los detalles de esta base de datos pueden encontrarse en [36].

- **Baeza:** contiene habla conversacional microfónica y recoge la variabilidad dialectal de diferentes lugares de la geografía española.

La mayoría de las bases de datos vistas han sido empleadas en la parte experimental del proyecto, ya sea para entrenar los modelos UBM, las matrices de *eigenvoices* y *eigenvectors* (para la compensación de variabilidad de sesión empleando técnicas basadas en *Factor Analysis*), o las cohortes de normalización (T-NORM, Z-NORM, ZT-NORM).

Para la realización de las pruebas se han utilizado las bases de datos Ahumada III y NIST SRE 2008 (conversación completa y 10 segundos) para la obtención de modelos y locutores. Sus características específicas se detallarán en siguientes apartados.

5.2. Protocolos de Evaluación

Un protocolo se define como un conjunto de normas y procedimientos. En este caso definimos los protocolos de evaluación como el conjunto de condiciones impuestas a los sistemas utilizados. De esta forma medimos el rendimiento de los sistemas de forma objetiva bajo las mismas condiciones.

Las condiciones bajo las que probamos los sistemas marcan las diferentes pruebas a realizar, las bases de datos a utilizar, condiciones de entrenamiento y *test* (número y tipo de los canales, duración de la conversación, idioma…), etc.

En la realización de este proyecto se ha seguido el protocolo de evaluación propuesto por NIST [37] (*National Institute of Standards and Technology*) aplicado a SRE (*Speaker Recognition Evaluation*).

5.2.1. Evaluaciones NIST

El organismo norteamericano NIST organiza evaluaciones internacionales competitivas de tecnología de reconocimiento de locutor e idioma. En este proyecto nos centramos en las evaluaciones SRE [37] que se encargan del reconocimiento de locutor, son de carácter abierto y de ellas derivan gran parte de las bases de datos utilizadas en los experimentos realizados.

Estas evaluaciones llevan organizándose desde el año 1996 y fueron de carácter anual hasta 2006; después, comenzaron a realizarse cada dos años intercalándose con evaluaciones de reconocimiento de idioma. Durante este periodo de tiempo las bases de datos utilizadas han ido evolucionando, ampliando el número de idiomas hablados por los locutores, el tipo de canal utilizado,…, y por tanto ha aumentado la variabilidad en los enfrentamientos, lo que convierte las evaluaciones en un desafío para los sistemas de reconocimiento que quieran presentarse a las mismas. De esta forma, se trata de mejorar siempre los sistemas para encontrarse en el estado-del-arte y conseguir un mejor rendimiento de cara a la problemática surgida, en constante aumento.

Así, tenemos sistemas en continua evolución que tratan de reflejar de la forma más precisa posible la realidad y se adaptan a las necesidades de la verificación de locutor independiente de texto. En estas evaluaciones se realiza una comparación de los sistemas, de forma que podemos conocer cuáles son más fiables para utilizarlos en entornos con múltiples variabilidades.

El protocolo de evaluación define la medida del rendimiento (función de coste) y los datos sobre los que realizar la evaluación. Es el mismo procedimiento para todos los integrantes y está definido por datos de entrenamiento (modelan la identidad a reconocer) y de test (modelos de locutor generados), así como por datos complementarios como los utilizados para compensación, normalización…

En las evaluaciones NIST SRE la tarea fundamental consiste en la verificación de un individuo a partir de una locución de prueba. Existen diversas *condiciones* en función de la cantidad y tipo de datos disponibles para entrenamiento y test (cuyas combinaciones proporcionan las diferentes tareas): en el caso de entrenamiento desde 10 segundos de habla hasta 8 conversaciones de 5 minutos (aproximadamente 2.5 minutos por cada locutor), y para los datos de test desde 10 segundos hasta 1 conversación de 5 minutos (unos 2.5 minutos de habla por locutor). La *condición* obligatoria en todas las evaluaciones NIST es la denominada 1conv-1conv, es decir, una conversación de habla de entrenamiento y una conversación de prueba.

Estas *condiciones* pueden proporcionarse incluyendo datos de entrenamiento/test en dos canales (4-*wire*), caso en el que el habla de los locutores de la conversación se encuentra en ficheros separados, ó en canales sumados (2-*wire*), caso en el que las conversaciones se encuentran mezcladas en un único fichero de audio.

En la elaboración de este proyecto se han utilizado las evaluaciones de NIST SRE 2006 y 2008, por lo que a continuación se detallan las *condiciones* de entrenamiento y test de cada plan de evaluación en las tablas 5-1 y 5-2.

		Condiciones de Test			
		10 seg 2 canales	**1 conv 2 canales**	**1 conv canales sumados**	**1 conv micrófono auxiliar**
Condiciones de Entrenamiento	**10 segundos 2 canales**	Opcional			
	1 conversación 2 canales	Opcional	Obligatoria	Opcional	Opcional
	3 conversaciones 2 canales	Opcional	Opcional	Opcional	Opcional
	8 conversaciones 2 canales	Opcional	Opcional	Opcional	Opcional
	3 conversaciones canales sumados		Opcional	Opcional	

Tabla 5-1. *Condiciones* de entrenamiento y test en la evaluación NIST SRE 2006 [39].

Condiciones de Entrenamiento		Condiciones de Test			
		10 sec	**short3**	**long**	**summed**
	10 sec	Opcional			
	short2	Opcional	Obligatoria		Opcional
	3conv		Opcional		Opcional
	8conv	Opcional	Opcional		Opcional
	long		Opcional	Opcional	
	3summed		Opcional		Opcional

Tabla 5-2. *Condiciones* de entrenamiento y test en la evaluación NIST SRE 2008 [40].

Donde la condición *short2* consiste en una extracción de 5 minutos de una conversación telefónica en dos canales o un segmento extraído de un medio microfónico de 3 minutos de conversación formada por una entrevista entre el locutor *target* y un entrevistador.

6. Experimentos

6.1. Introducción

En este capítulo se procede a mostrar los diferentes experimentos realizados en la elaboración de este proyecto fin de carrera. Con ellos se ha tratado de estudiar el efecto que produce aplicar técnicas de compensación de variabilidad de sesión (explicadas en el Capítulo 4) en las puntuaciones obtenidas utilizando el sistema de reconocimiento automático de locutor con las diferentes bases de datos que han sido detalladas en el Capítulo 5.

Los experimentos tienen como objetivo probar los efectos de la compensación de la variabilidad intersesión y del modelado de la variabilidad de locutor, así como los efectos de las diferentes normalizaciones en las puntuaciones, de la adaptación de las cohortes de normalización a las longitudes de la prueba, de la adaptación de las matrices de compensación de variabilidad a las longitudes de la prueba y finalmente de la variación del número de *eigenvoices* y/o *eigenchannels*.

El ámbito y la problemática en que se centra el proyecto son forenses, pero con el objetivo de realizar comparaciones y como medio de apoyo se han realizado también experimentos en entornos bien conocidos controlados.

El sistema de reconocimiento de locutor que ha sido utilizado en todos los experimentos es un GMM-UBM de 1024 Gaussianas en un espacio de características de 38 dimensiones.

Todos los experimentos se han realizado utilizando el software de voz del ATVS, en lenguajes C++ y Matlab, realizando ligeras modificaciones al código para adaptarlo a los requisitos de cada prueba.

6.2. Efecto de la Compensación de Variabilidad en Entornos Controlados

6.2.1. Introducción

A continuación se procede a analizar los resultados obtenidos realizando experimentos sobre un sistema de reconocimiento de locutor utilizado en un entorno bien conocido: las evaluaciones NIST SRE. Comenzamos en un entorno controlado para demostrar la validez de los experimentos posteriores, realizados en entornos forenses. De esta forma tenemos una base previa.

Los experimentos han sido realizados variando las longitudes de los segmentos utilizados en entrenamiento y test. Nuestro objetivo es comparar el efecto de la compensación de variabilidad en experimentos en los que se cuenta con datos de suficiente duración para entrenamiento y test con el efecto en experimentos en los que la duración de los datos es escasa.

Para lograr dicho objetivo, se parte de dos tareas diferentes de la evaluación NIST SRE 2008: 1conv-1conv y 10sec-10sec. La primera de ellas consta de locuciones de entrenamiento y test formadas por conversaciones de una duración media de 2.5 minutos y la segunda está formada por locuciones de corta duración, aproximadamente 10 segundos. De este modo se comprueba el efecto que tiene la duración de los datos disponibles de cara a la realización de los experimentos.

La utilidad de estos experimentos se ve reflejada en apartados posteriores, dado que aunque las evaluaciones NIST SRE se realizan en entornos controlados con gran cantidad de datos y larga duración de los mismos, existen otros muchos casos donde la duración de las muestras es mucho menor, como es el caso de los escenarios forenses, en los que se centra este proyecto.

6.2.2. Sistema de Partida

El sistema utilizado para la realización de los experimentos de este apartado es el que comprende la base de datos NIST SRE 2008. Dicha base de datos se ha utilizado para la realización de los estadísticos de entrenamiento y test, tomando únicamente datos de hablantes masculinos.

Para obtener los estadísticos, es necesario realizar un proceso previo de parametrización de los archivos de audio iniciales. Este proceso se ha realizado segmentando los archivos en tramas de 20 ms con solapamiento del 50% (tasa de 10 ms), extrayendo los vectores de características utilizando 19 coeficientes MFCC y normalizando los vectores en canal mediante CMN RASTA *warping*.

El tipo de sistema de reconocimiento de locutor empleado en este y en el resto de experimentos es un GMM-UBM de 1024 Gaussianas en un espacio de características de 38 dimensiones.

La normalización de puntuaciones empleada es TNORM, ZNORM y ZTNORM, utilizando para normalizar cohortes formadas por archivos de las bases de datos de NIST SRE 2004 y 2005.

6.2.3. Rendimiento del Sistema Base

Inicialmente es necesario conocer el rendimiento del sistema sin realizar ningún tipo de compensación, ni de canal ni de locutor. De este modo vemos el efecto logrado a medida

que utilizamos las técnicas de compensación de variabilidad sobre nuestro sistema de referencia.

Tras llevar a cabo 12922 *trials* sobre 648 modelos de locutor para el caso 1conv-1conv y 7799 *trials* sobre 648 modelos de locutor para el caso 10s-10s los resultados obtenidos son los que se muestran en las siguientes curvas DET.

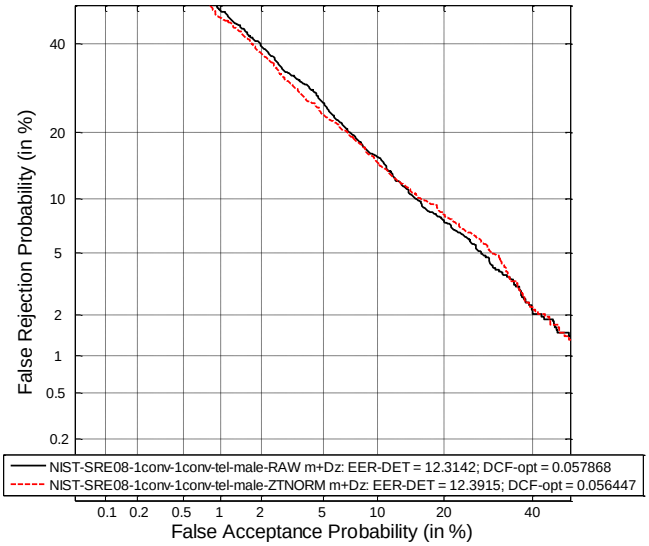


Figura 6-1. Rendimiento del sistema base con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos.

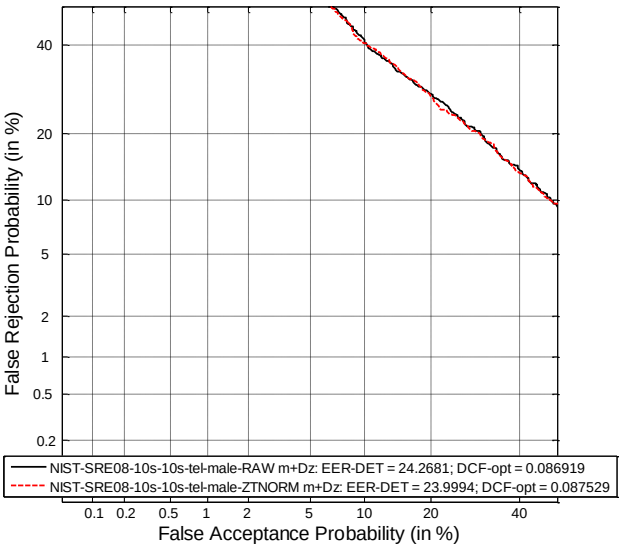


Figura 6-2. Rendimiento del sistema base con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 10s-10s para locutores masculinos.

Como puede observarse, el hecho de disponer de archivos de duración menor aumenta la tasa de error de tal modo que prácticamente la duplica. Esto ocurre, entre otras causas, porque con menor cantidad de datos de entrenamiento los modelos son menos robustos y no reflejan de igual forma la variabilidad existente entre las distintas locuciones y el sistema obtiene un rendimiento más bajo en los enfrentamientos.

6.2.4. Rendimiento del Sistema tras modelar la Variabilidad de Locutor

Utilizando las bases de datos definidas en el capítulo 5, entrenamos una matriz de variabilidad de locutor, V, con 300 *eigenvoices* y 10 iteraciones EM. Aplicamos el subespacio de variabilidad de locutor a los modelos de entrenamiento y a la cohorte de normalización en test.

Tras llevar a cabo 12922 *trials* sobre 648 modelos de locutor para el caso 1conv-1conv y 7799 *trials* sobre 648 modelos de locutor para el caso 10s-10s los resultados obtenidos son los que se muestran en las siguientes curvas DET.

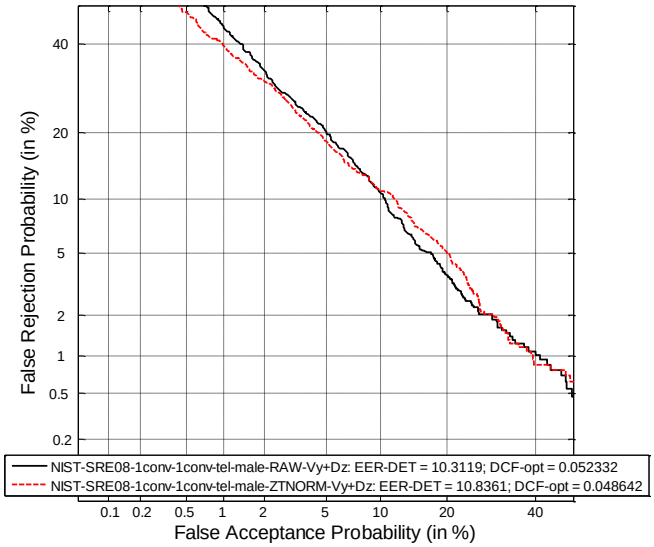


Figura 6-3. Rendimiento del sistema tras modelar la variabilidad de locutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos.

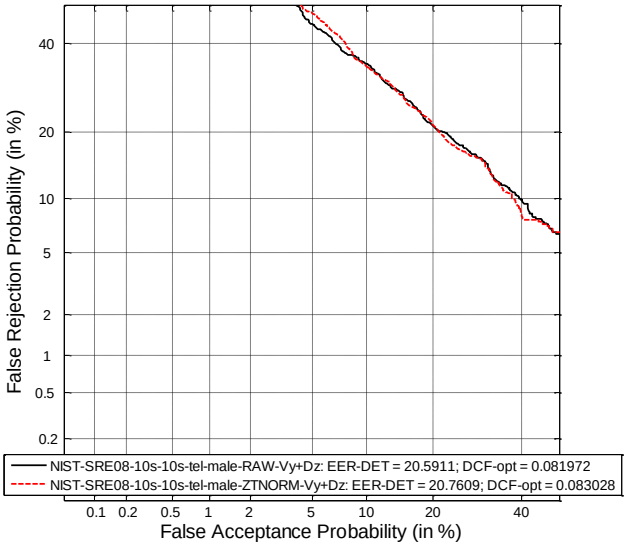


Figura 6-4. Rendimiento del sistema tras modelar la variabilidad de locutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 10s-10s para locutores masculinos.

De nuevo se observa un aumento significativo de la tasa de error en el caso de las locuciones de 10 segundos, pero si comparamos los resultados con los del apartado anterior queda patente una mejora de dichas tasas gracias al modelado de locutor. Esta mejora es más notable en el caso de la tarea 10sec-10sec puesto que se reduce la EER en 4 puntos (mejora del 17%).

6.2.5. Rendimiento del Sistema tras compensar la Variabilidad Intersesión y modelar la Variabilidad de Locutor mediante técnicas de JFA.

Utilizando las bases de datos definidas en el capítulo 5, entrenamos una matriz de variabilidad de locutor, V, con 300 *eigenvoices* y 10 iteraciones EM. Con dicha matriz, modelamos la variabilidad de locutor de los modelos de entrenamiento y de la cohorte de normalización en test.

Con las bases de datos de NIST SRE 2004 y 2005, 50 *eigenchannels* y 10 iteraciones EM, entrenamos una matriz de variabilidad de canal con la que compensamos en canal los estadísticos de entrenamiento y test así como los de las cohortes de t-norm y z-norm.

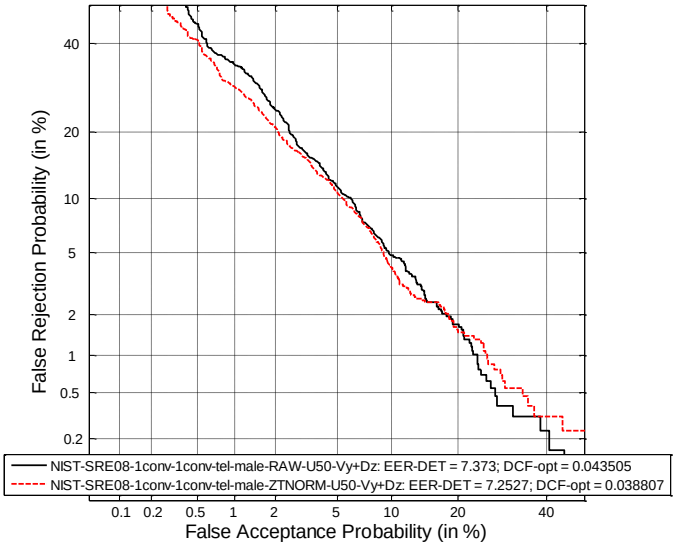


Figura 6-5. Rendimiento del sistema tras compensar la variabilidad de canal y modelar la de locutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos.

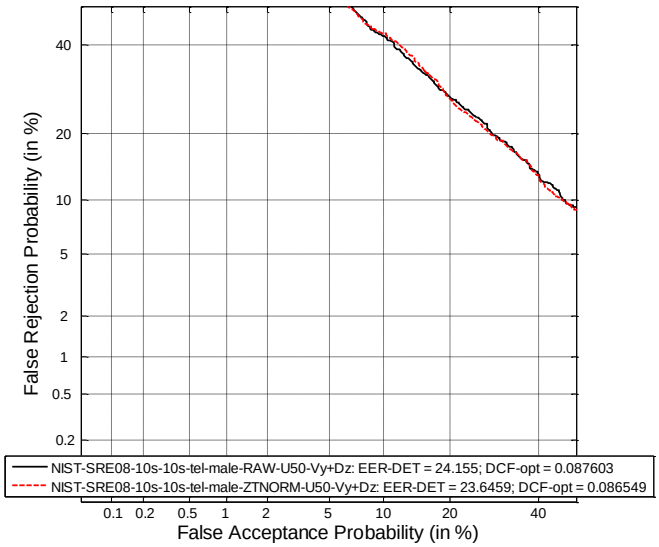


Figura 6-6. Rendimiento del sistema tras compensar la variabilidad de canal y modelar la de locutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 10s-10s para locutores masculinos.

En este caso los resultados toman caminos diferentes:

Para la tarea 1conv-1conv se observa una reducción de la tasa de error del 41% con respecto a los experimentos sin compensación y del 30% con respecto a los resultados arrojados tras modelar la variabilidad de locutor.

Sin embargo, en la tarea 10sec-10sec parece que compensar en canal y modelar en locutor no ofrece una mejora significativa, de hecho existe cierto empeoramiento de los resultados con respecto al apartado anterior, en el que únicamente se entrenaba el subespacio de variabilidad de locutor.

6.2.6. Rendimiento del Sistema tras entrenar el Espacio de Variabilidad de Locutor Ampliado y compensar la Variabilidad Intersesión e Interlocutor mediante técnicas de JFA.

Los experimentos realizados en este apartado son similares a los del apartado anterior puesto que se trata de *Joint Factor Analysis*. La diferencia radica en que en este caso ampliamos el espacio de variabilidad de locutor entrenando la matriz D (residual) con estadísticos obtenidos de la base de datos Baeza (mencionada en el capítulo 5).

A continuación mostramos las gráficas con los resultados obtenidos tras llevar a cabo 12922 *trials* sobre 648 modelos de locutor para el caso 1conv-1conv y 7799 *trials* sobre 648 modelos de locutor para el caso 10s-10s.

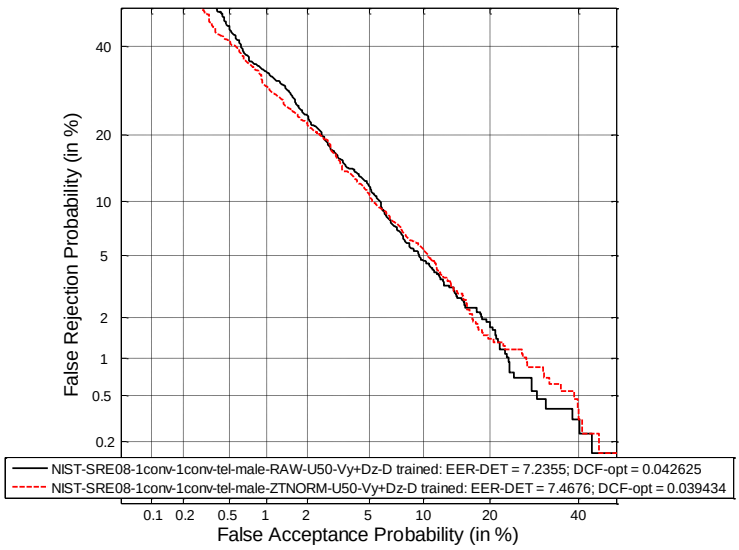


Figura 6-7. Rendimiento del sistema tras entrenar el espacio de variabilidad de locutor ampliado y compensar la variabilidad intersesión e interlocutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos.

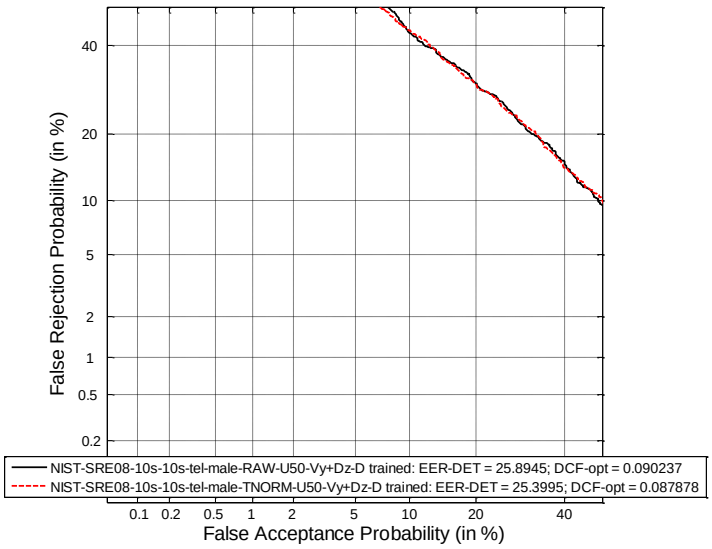


Figura 6-8. Rendimiento del sistema tras entrenar el espacio de variabilidad de locutor ampliado y compensar la variabilidad intersesión e interlocutor con normalización de puntuaciones. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos.

En este caso, la tasa de error resultante es muy similar a la del apartado anterior para la tarea 1conv-1conv y empeora entre 1 y 2 puntos para 10sec-10sec. Estos resultados se deben probablemente al hecho de haber tomado datos de Baeza, una base de datos con diferente formato a aquéllas con las que estamos trabajando, ya sea tanto en modelos y locutores como en matrices de compensación de variabilidad (procedentes de las distintas evaluaciones NIST SRE).

6.2.7. Comparativa entre las diferentes técnicas

Con el propósito de facilitar las conclusiones obtenidas del estudio realizado en el apartado 6.2 se han realizado gráficas conjuntas con los resultados sin normalización de puntuaciones, de forma que se puede observar el efecto que tiene cada técnica de compensación de variabilidad añadida en los resultados finales. A continuación se presentan dichas gráficas.

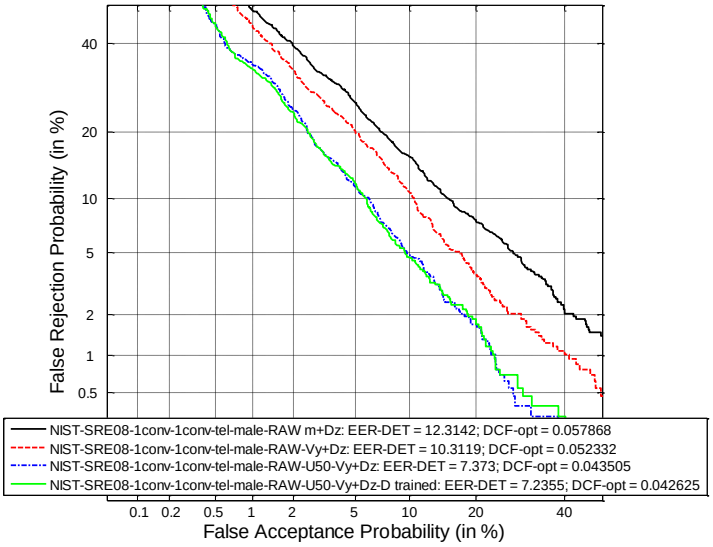


Figura 6-9. Rendimiento del sistema tras aplicar las diferentes técnicas de compensación de variabilidad. Evaluación sobre datos NIST SRE 2008 tarea 1conv-1conv para locutores masculinos.

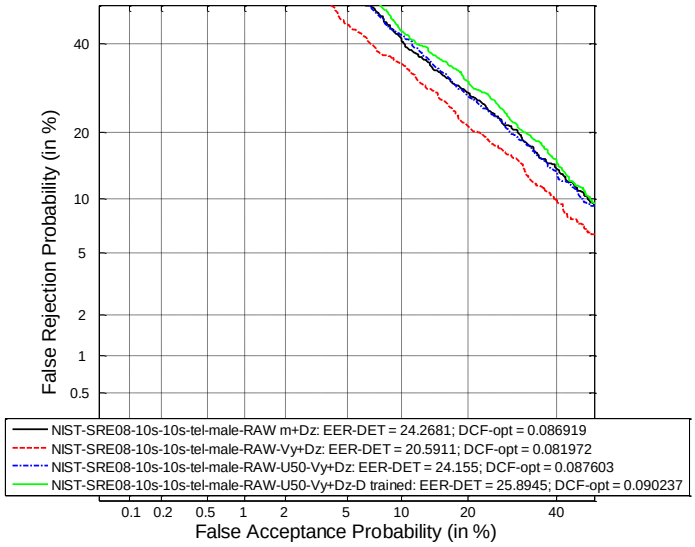


Figura 6-10. Rendimiento del sistema tras aplicar las diferentes técnicas de compensación de variabilidad. Evaluación sobre datos NIST SRE 2008 tarea 10sec-10sec para locutores masculinos.

Como se puede observar en las gráficas, para la tarea 1conv-1conv se produce una degradación progresiva tanto de la tasa de error como de la función de coste, esto es, nuestro sistema mejora su rendimiento a medida que introducimos técnicas de compensación de variabilidad.

Como contrapunto, la tarea 10sec-10sec ofrece resultados poco satisfactorios en tanto en cuanto la técnica JFA no sólo no mejora los resultados de las tasas de error y función de coste, sino que los empeora. Los resultados óptimos para estos experimentos son los obtenidos modelando únicamente la variabilidad de locutor.

6.3. Efecto de la Compensación de Variabilidad en Entornos Forenses

6.3.1. Introducción

En este apartado se trabaja con experimentos en entornos forenses, es decir, casos reales que suponen un desafío para los sistemas de reconocimiento de locutor.

Concretamente hemos trabajado con la base de datos Ahumada III, definida en el capítulo 5, en su tercera versión, que consta de 69 locutores masculinos hispanohablantes.

Los experimentos realizados son idénticos a los ya ejecutados en el apartado 6.2 añadiendo nuevas pruebas de forma que el estudio realizado en nuestro sistema de reconocimiento de locutor sobre esta base de datos forense sea más completo. Las pruebas incluyen aumento de las cohortes de normalización, ajuste de las cohortes de normalización a las longitudes de la prueba, ajuste de las matrices de variabilidad a las longitudes de la prueba, variación del número de *eigenvoices* y variación del número de *eigenchannels*.

6.3.2. Sistema de Partida

El sistema utilizado para la realización de los experimentos de este apartado es el que comprende la base de datos Ahumada III. Dicha base de datos se ha utilizado para la realización de los estadísticos de entrenamiento y test.

Para obtener los estadísticos, es necesario realizar un proceso previo de parametrización de los archivos de audio iniciales. Este proceso se ha realizado segmentando los archivos en tramas de 20 ms con solapamiento del 50% (tasa de 10 ms), extrayendo los vectores de características utilizando 19 coeficientes MFCC y normalizando los vectores en canal mediante CMN RASTA *warping*.

El tipo de sistema de reconocimiento de locutor empleado en los experimentos es un GMM-UBM de 1024 Gaussianas en un espacio de características de 38 dimensiones.

La normalización de puntuaciones empleada es TNORM, ZNORM y ZTNORM, utilizando para normalizar cohortes formadas por archivos de las bases de datos de NIST SRE 2004 y 2005.

6.3.3. Rendimiento del Sistema Base

Antes de comprobar el efecto de las diferentes técnicas de compensación de variabilidad utilizadas o de los diferentes ajustes realizados en las cohortes, es necesario conocer la tasa de error y la función de coste de nuestro sistema. De este modo conocemos el punto de partida del rendimiento y podemos así realizar comparaciones entre los diferentes resultados a medida que vamos ejecutando experimentos.

Tras llevar a cabo 33879 trials sobre 69 modelos de locutor obtenemos los siguientes resultados.

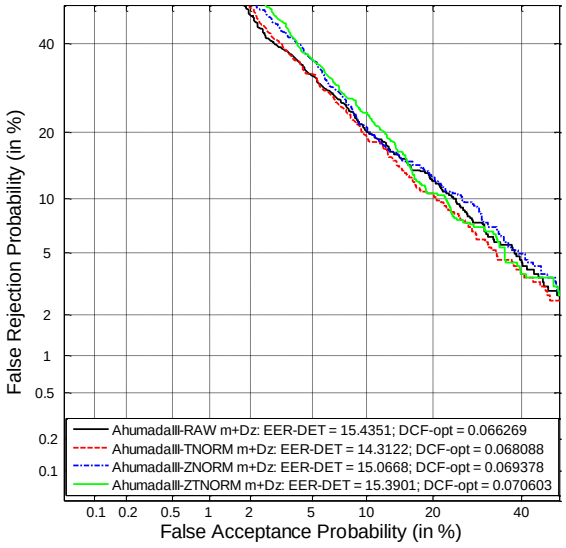


Figura 6-11. Rendimiento del sistema base con normalización de puntuaciones. Evaluación sobre datos procedentes de Ahumada III.

Los resultados otorgan una tasa de error del 15% aproximadamente, un dato comparable al obtenido en la base de datos NIST SRE 2008 para la tarea 1conv-1conv, aunque ligeramente superior. Ocurre así porque ambas bases de datos contienen modelos de longitudes similares (150 segundos para NIST y 120 para Ahumada) pero los archivos de test utilizados en Ahumada III constan de longitudes significativamente menores (13 segundos).

6.3.4. Rendimiento del Sistema tras modelar la Variabilidad de Locutor

La primera técnica de compensación utilizada consiste en entrenar una matriz de variabilidad de locutor a partir de las bases de datos definidas en el capítulo 5. Utilizando 300 *eigenvoices* y 10 iteraciones EM conseguimos la matriz con la que modelaremos la variabilidad de locutor de los modelos de entrenamiento y de la cohorte de normalización en test.

Llevamos a cabo 33879 *trials* sobre 69 modelos de locutor y obtenemos los siguientes resultados.

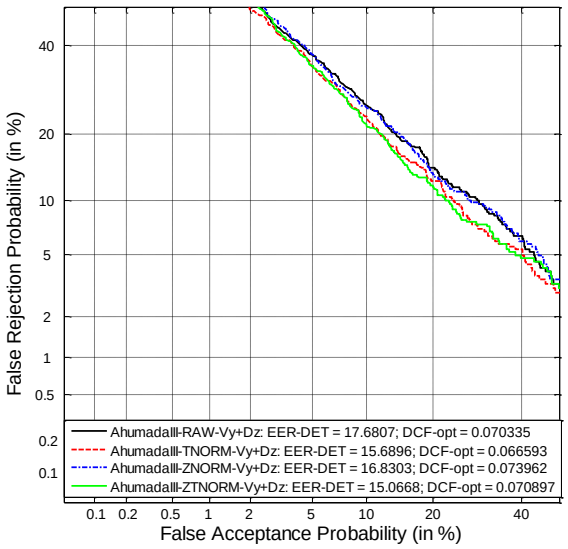


Figura 6-12. Rendimiento del sistema tras realizar modelado en locutor con normalización de puntuaciones. Evaluación sobre datos procedentes de Ahumada III.

En este caso la compensación de variabilidad interlocutor no es muy efectiva puesto que incrementa la tasa de error con respecto al sistema inicial en todos los casos salvo para normalización en cero y en test (ZTNORM) que disminuye ligeramente.

6.3.5. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA

En este apartado utilizaremos el modelo *Joint Factor Analysis* de compensación de variabilidad para tratar de reducir las tasas de error del sistema, así como la función de coste.

Utilizando 300 *eigenvoices* y 10 iteraciones EM entrenamos la matriz con la que modelaremos la variabilidad interlocutor en los modelos de entrenamiento y los de la cohorte de normalización en test.

Con 50 *eigenchannels* y 10 iteraciones EM, entrenamos una matriz de variabilidad de canal con la que compensamos en canal los estadísticos de entrenamiento y test así como los de las cohortes de t-norm y z-norm.

Una vez que tenemos ambas matrices procedemos a ejecutar el experimento que genera los siguientes resultados.

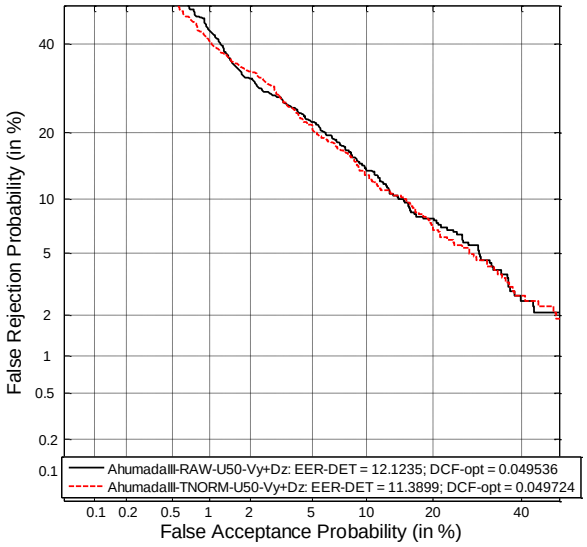


Figura 6-13. Rendimiento del sistema tras compensar la variabilidad intersesión e interlocutor mediante JFA con normalización de puntuaciones. Evaluación sobre datos procedentes de Ahumada III.

Tras la compensación en canal y locutor conseguimos una mejora de 3 puntos en la tasa de error. Al igual que ocurría en las bases de datos NIST SRE, JFA nos proporciona los resultados más satisfactorios tanto en DET como en DCF para Ahumada III.

6.3.6. Rendimiento del Sistema tras entrenar el Espacio de Variabilidad de Locutor Ampliado, compensar la Variabilidad Intersesión y modelar la Variabilidad de Locutor mediante JFA

En este apartado utilizamos de nuevo el modelo JFA pero con un añadido, se amplía el espacio de variabilidad de locutor entrenando una matriz, para lo cual utilizamos la base de datos Baeza (capítulo 5).

Tras entrenar las matrices pertinentes, compensar en canal y modelar en locutor los resultados obtenidos son los que muestra la gráfica a continuación.

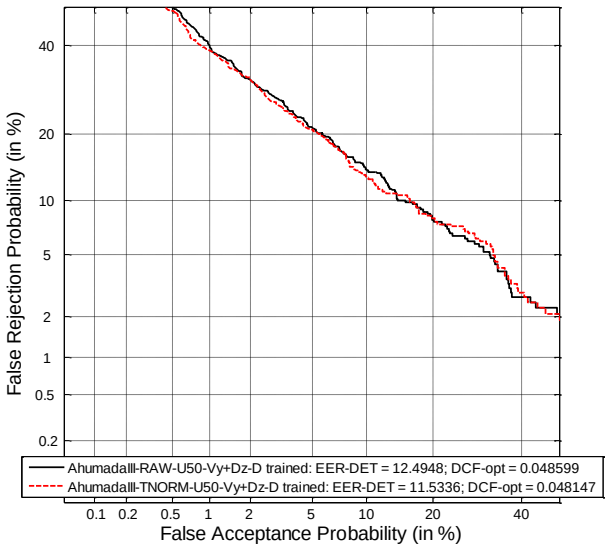


Figura 6-14. Rendimiento del sistema tras entrenar el espacio de variabilidad ampliada, compensar en canal y modelar en locutor mediante JFA y con normalización de puntuaciones. Evaluación sobre datos procedentes de Ahumada III.

Como se observa en la gráfica, con este método no hemos conseguido mejoras en los resultados, salvo un leve descenso de la tasa de error para el caso con normalización ZNORM, pero prácticamente imperceptible.

6.3.7. Comparativa entre las diferentes técnicas

La mejor forma de comparar los resultados arrojados por los distintos experimentos llevados a cabo es resumirlos en una única gráfica. Con ese propósito hemos tomado las puntuaciones RAW (sin normalizar) de cada uno de los apartados anteriores y las hemos representado en la siguiente figura.

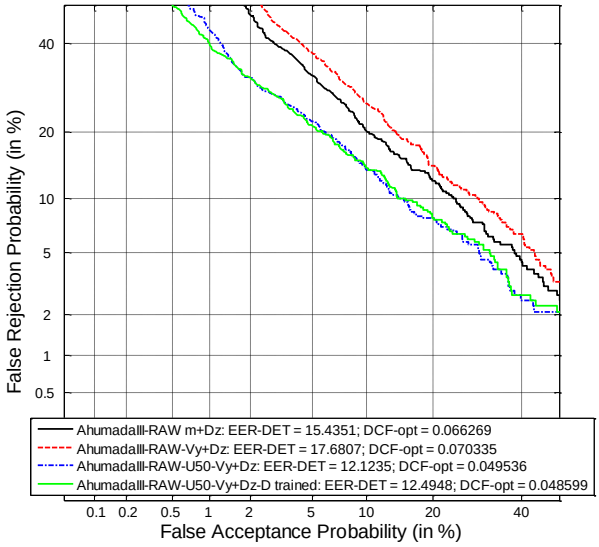


Figura 6-15. Rendimiento del sistema tras aplicar las diferentes técnicas de compensación de variabilidad sin normalizar puntuaciones. Evaluación sobre datos procedentes de Ahumada III.

En la gráfica queda patente el hecho de que los resultados óptimos en este conjunto de experimentos son los obtenidos mediante el uso de técnicas de *Joint Factor Analysis* para compensar en canal y en locutor. En contrapunto están los resultados obtenidos realizando modelado únicamente en la variabilidad de locutor, que ofrecen tasas de error incluso superiores a los experimentos en los que no se han utilizado técnicas de compensación. El motivo principal de que ocurra esto es el desajuste existente entre los datos.

6.4. Efecto del Ajuste de las Cohortes a las Longitudes de la Prueba

6.4.1. Introducción

En los apartados anteriores se han realizado experimentos utilizando bases de datos que constan de archivos con diferentes duraciones. Ahumada III está formada por modelos de 120 segundos y archivos de test de 13 segundos, mientras que las cohortes de NIST SRE, utilizadas para normalizar y para entrenar las matrices de variabilidad, contienen archivos de aproximadamente 150 segundos.

El objetivo de los experimentos que vamos a detallar es comprobar si ajustando las duraciones de los archivos pertenecientes a dichas cohortes a las longitudes de la prueba conseguimos una mejora significativa del rendimiento o, por el contrario, este se mantiene similar o incluso empeora.

6.4.2. Rendimiento del Sistema Base tras aumentar las Cohortes de Normalización

Inicialmente se verá el sistema con el que se ha trabajado sin utilizar ningún tipo de compensación. Para ello contamos con los resultados obtenidos anteriormente y además los obtenidos tras aplicar un aumento a las cohortes de normalización, de forma que se utilizan estadísticos no sólo de la base de datos de NIST SRE 2005 sino también de la evaluación de 2004.

Tras llevar a cabo 33879 *trials* sobre 69 modelos de locutor obtenemos los siguientes resultados.

	RAW	TNORM	ZNORM	ZTNORM
NIST SRE 2005	15.44 / 0.066	14.31 / 0.068	15.90 / 0.073	15.28 / 0.075
NIST SRE 2004 y 2005	15.44 / 0.066	19.23 / 0.067	16.18 / 0.064	14.34 / 0.065

Tabla 6-1. Rendimiento del Sistema Base antes y después de aumentar las cohortes de normalización. Base de Datos Ahumada III.

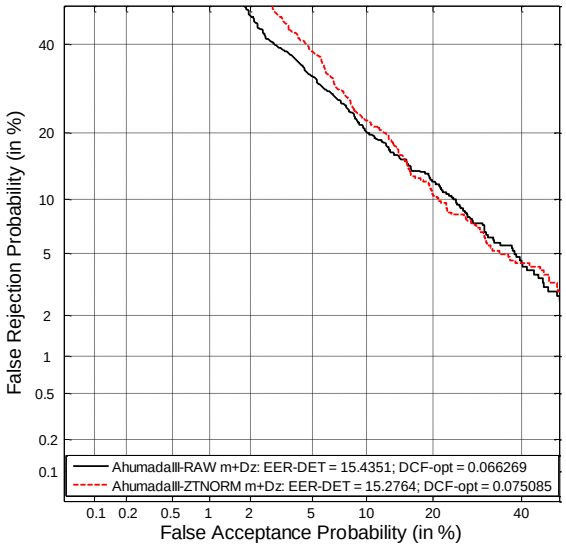


Figura 6-16. Rendimiento del sistema base con normalización NIST SRE 2005. Base de Datos Ahumada III.

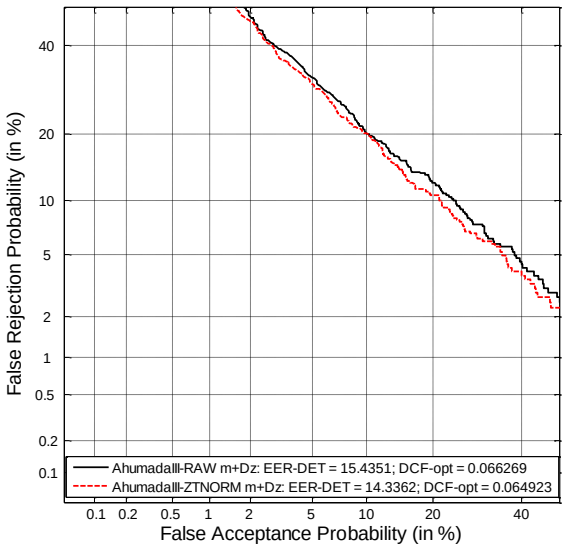


Figura 6-17. Rendimiento del sistema base tras aumentar las cohortes de normalización con datos de NIST SRE 2004. Base de Datos Ahumada III.

Analizando los resultados obtenidos podemos concluir que no es útil aumentar las cohortes de normalización en este caso, puesto que aunque en el caso de normalización en cero y test se obtiene una mejora del rendimiento en un punto, en los casos de normalizaciones de cero o test por separado el rendimiento empeora, especialmente en TNORM, donde la tasa de error aumenta hasta 4 puntos.

El motivo de la obtención de dichos resultados es el hecho de que para normalizar se emplean cohortes formadas por archivos pertenecientes a bases de datos de NIST SRE mientras que para las pruebas se emplea la base de datos Ahumada III, por lo que seguir aumentando las cohortes con archivos tan diferentes a los utilizados en los experimentos no produce mejora alguna.

6.4.3. Rendimiento del Sistema tras Modelar la Variabilidad de Locutor, Compensar la Variabilidad de Canal y aumentar las Cohortes de Normalización

Tras comprobar el efecto del aumento de las cohortes de normalización en el sistema base, se procede a comprobar si es el mismo en un sistema en el que se han utilizado técnicas de *Joint Factor Analysis*. Para ello partimos de los experimentos realizados en el apartado 6.3.5 y aumentamos las cohortes de normalización de puntuaciones utilizando estadísticos de la base de datos NIST SRE 2004.

Los resultados obtenidos de forma comparativa, antes y después de aumentar las cohortes, se muestran a continuación.

	RAW	TNORM	ZNORM	ZTNORM
NIST SRE 2005	12.12 / 0.050	11.39 / 0.050	12.18 / 0.054	11.95 / 0.053
NIST SRE 2004 y 2005	12.12 / 0.050	12.16 / 0.051	16.60 / 0.074	15.90 / 0.073

Tabla 6-2. Rendimiento del Sistema aplicando JFA antes y después de aumentar las cohortes de normalización. Base de Datos Ahumada III.

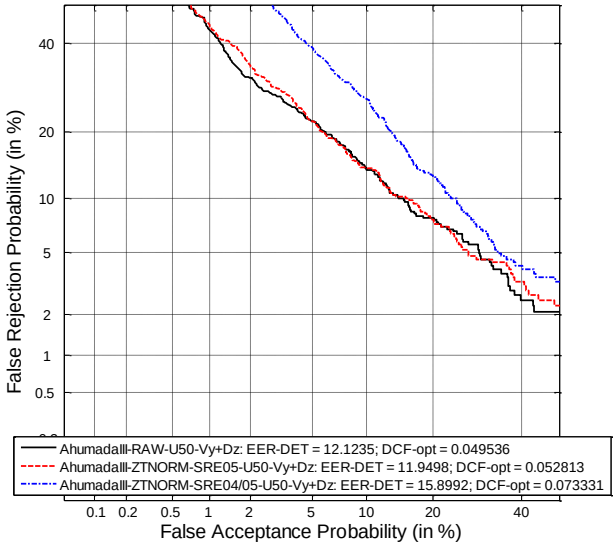


Figura 6-18. Rendimiento del Sistema tras aplicar JFA antes y después de aumentar las cohortes de normalización. Base de Datos Ahumada III.

En este caso queda patente la incapacidad para mejorar las puntuaciones del aumento en el número de estadísticos de las cohortes de normalización, al menos, para las bases de datos utilizadas. Este sistema probablemente sea útil en el caso en que los tipos de archivos utilizados para normalizar sean similares a los utilizados en la prueba, pero no ocurre así para los estadísticos disponibles en nuestro sistema.

6.4.4. Rendimiento del Sistema tras ajustar los estadísticos de las Cohortes de Normalización a las Longitudes de la Prueba

En los experimentos realizados se utilizan archivos pertenecientes a la base de datos Ahumada III mientras que se cuenta con una cohorte de estadísticos de normalización obtenidos de bases de datos NIST SRE. Este hecho implica que la duración de los archivos sea diferente, puesto que los archivos de Ahumada III tienen una duración de 120 segundos para modelos y 13 segundos para test, y en NIST la duración media es de 120 segundos.

Para comprobar en qué grado afecta la variación en las duraciones a los resultados finales, se ha realizado un ajuste de los estadísticos utilizados para normalizar disminuyendo a 120 segundos los empleados para normalización en test y a 13 segundos los empleados para normalización en cero.

Tras llevar a cabo las pruebas pertinentes tanto en el sistema sin compensar como en el sistema tras aplicar JFA se obtienen los siguientes resultados.

	RAW	TNORM	ZNORM	ZTNORM
Sistema Base	15.44 / 0.066	14.31 / 0.068	15.90 / 0.073	15.28 / 0.075
JFA sin ajuste	12.12 / 0.050	11.39 / 0.050	12.18 / 0.054	11.95 / 0.053
Sin Compensar	15.44 / 0.066	14.65 / 0.076	15.69 / 0.073	18.53 / 0.077
JFA	12.12 / 0.050	11.41 / 0.049	11.95 / 0.050	14.17 / 0.059

Tabla 6-3. Rendimiento del sistema sin compensar y tras aplicar JFA, antes y después del ajuste de las longitudes de los estadísticos en las cohortes de normalización. Base de Datos Ahumada III.

Podemos observar que el ajuste de las cohortes no es efectivo, probablemente por la pérdida de datos al cortar los estadísticos, ya que no se realiza un corte selectivo. Adicionalmente, se ha comprobado que es más efectivo ajustar únicamente los estadísticos de z-norm, como se puede ver en los resultados a continuación.

	RAW	TNORM	ZNORM	ZTNORM
Sin Compensar	15.44 / 0.066	14.31 / 0.068	15.69 / 0.073	14.96 / 0.075
JFA	12.12 / 0.050	11.39 / 0.050	11.95 / 0.050	12.01 / 0.051

Tabla 6-4. Rendimiento del sistema sin compensar y tras aplicar JFA, ajustando las longitudes de los estadísticos en la cohorte de z-norm. Base de Datos Ahumada III.

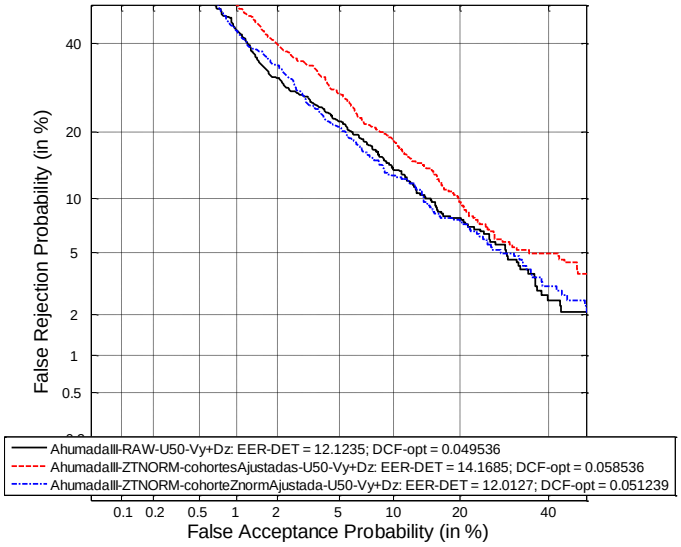


Figura 6-19. Rendimiento del sistema sin compensar y tras aplicar JFA, tras el ajuste de las longitudes de los estadísticos en las cohortes de normalización (en todas y únicamente en z-norm). Base de Datos Ahumada III.

6.4.5. Rendimiento del Sistema tras ajustar las Matrices de Variabilidad a las longitudes de la prueba

Como últimos experimentos ajustando la duración de los estadísticos se han adaptado las matrices utilizadas para compensar la variabilidad. Dichas matrices se entrenan utilizando archivos de diferentes bases de datos, sin ser ninguna de ellas Ahumada III, que es la utilizada para los modelos y ficheros de test utilizados en las pruebas. Por ello, las longitudes de los archivos utilizados son diferentes y el objetivo de este apartado es comprobar si ajustando dichas longitudes se consigue optimizar el rendimiento o, al contrario, empeora.

La primera prueba consiste en ajustar los archivos con los que se entrena la matriz de variabilidad de sesión a 120 segundos y a 13 segundos, para compensar en canal tanto archivos de entrenamiento como de test. Una vez hecho esto procedemos a llevar a cabo JFA obteniendo los siguientes resultados.

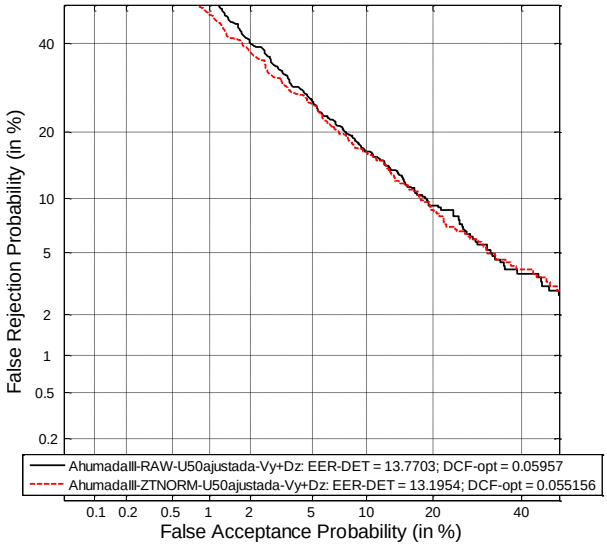


Figura 6-20. Rendimiento del Sistema tras ajustar las matrices de variabilidad de sesión a las longitudes de la prueba y aplicar JFA. Base de Datos Ahumada III.

En los resultados se observa que no se produce una mejora con respecto al sistema antes de ajustar las longitudes a la prueba. Para comprobarlo de forma más competa, se han ajustado también las longitudes en la matriz utilizada para modelar la variabilidad de locutor. Como estas matrices se utilizan en los modelos de locutor, se han ajustado a 120 segundos. Aplicando JFA con ambas matrices restringidas a las nuevas duraciones, se obtienen los resultados que podemos ver a continuación.

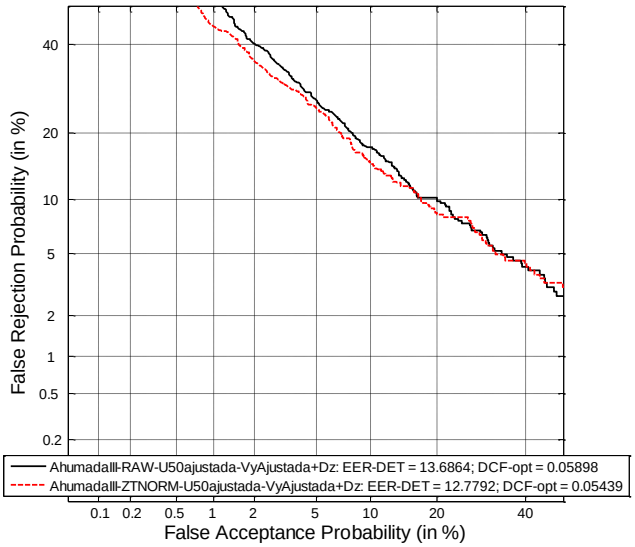


Figura 6-21. Rendimiento del Sistema tras ajustar las matrices de variabilidad de sesión y locutor a las longitudes de la prueba y aplicar JFA. Base de Datos Ahumada III.

En este caso los resultados mejoran levemente con respecto al caso anterior pero siguen siendo peores que en el caso sin ajustar las longitudes en las matrices.

Por último se ha comprobado el efecto combinado de ajustar también las longitudes de las cohortes de normalización en los dos casos previos. En la siguiente gráfica se resumen los cuatro casos vistos en este apartado de forma que se puedan comparar mejor los resultados.

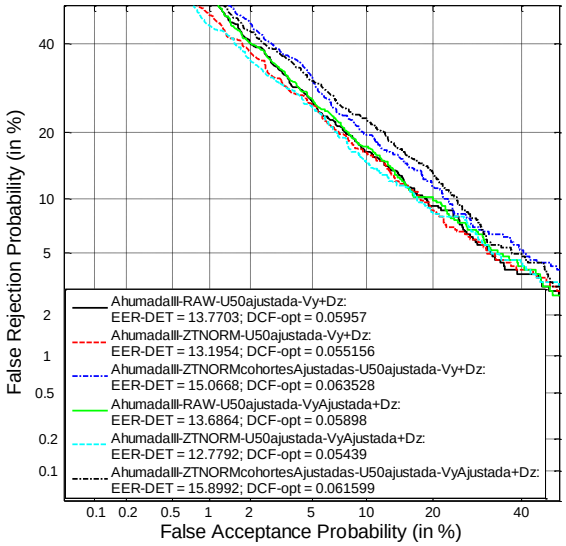


Figura 6-22. Rendimiento del sistema tras ajustar las matrices de variabilidad así como las cohortes de normalización. Base de Datos Ahumada III.

En base a los resultados obtenidos podemos concluir que el caso óptimo es aquél en el que ajustamos las matrices de variabilidad de canal y locutor a las longitudes de la prueba pero sin ajustar las cohortes de normalización de puntuaciones.

6.5. Efecto del Número de Direcciones de Máxima Variabilidad Utilizadas

6.5.1. Introducción

A lo largo de los experimentos realizados en este proyecto se utilizan, en aquellos en los que se aplican técnicas de JFA, matrices de variabilidad, bien de canal, bien de locutor. Dichas matrices están constituidas por los vectores que mejor representan los efectos de la sesión en canal (los llamados *eigenchannels*) y por los que mejor representan los componentes de variación más importantes entre locutores (*eigenvoices*).

En los apartados anteriores, la presencia cuantitativa de estos vectores en las matrices era invariante: 300 *eigenvoices* y 50 *eigenchannels* en todos los casos de estudio. En este apartado se ha tratado de comprobar el efecto que tiene la variación en la cantidad de *eigenfactors* utilizados.

Para cumplir el objetivo, se ha procedido a variar el número de *eigenvoices* entre 50 y 300 en intervalos de 50, y la cantidad de *eigenchannels* entre 10 y 50 en intervalos de 10. Las pruebas se han realizado variando únicamente el número de *eigenvoices*, variando únicamente el número de *eigenchannels* y variando tanto *eigenvoices* como *eigenchannels*.

6.5.2. Rendimiento del Sistema tras modelar la Variabilidad de Locutor para diferente número de *eigenvoices*

Se comienza por entrenar una matriz de variabilidad de locutor a partir de las bases de datos definidas en el capítulo 5. Utilizando diferente número de *eigenvoices* (comenzando por un mínimo de 50 e incrementando 50 en cada caso hasta llegar a 300) y 10 iteraciones EM conseguimos la matriz con la que modelaremos la variabilidad de locutor de los modelos de entrenamiento y de la cohorte de normalización en test.

Llevamos a cabo 33879 *trials* sobre 69 modelos de locutor y obtenemos los siguientes resultados.

NEIGENVOICES	RAW	TNORM	ZNORM	ZTNORM
50	18.64/0.077	17.52/0.072	17.44/0.077	16.11/0.071
100	17.77/0.073	16.73/0.070	17.27/0.075	15.62/0.071
150	17.63/0.072	16.11/0.068	17.02/0.074	15.07/0.071
200	17.38/0.071	15.90/0.068	16.98/0.074	15.28/0.071
250	17.55/0.071	15.85/0.068	16.75/0.074	15.28/0.071
300	17.68/0.070	15.69/0.067	16.83/0.074	15.07/0.071

Tabla 6-5. Rendimiento del Sistema tras modelar la Variabilidad de Locutor variando el número de *eigenvoices*. Base de Datos Ahumada III.

Para comprobar el efecto de un modo más visual se han generado las siguientes gráficas.

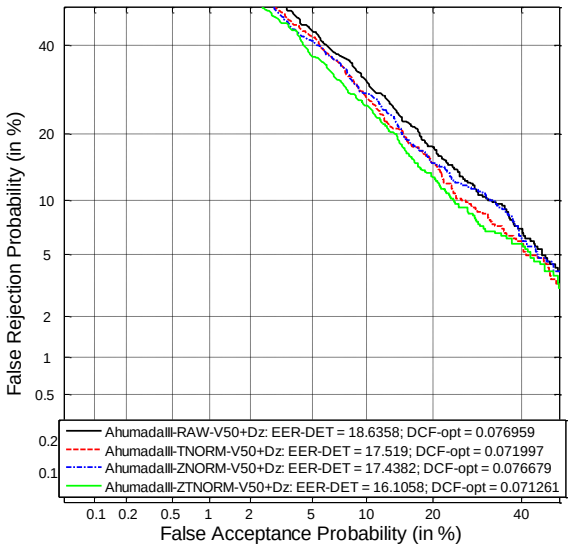


Figura 6-23. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 50 *eigenvoices*. Base de Datos Ahumada III.

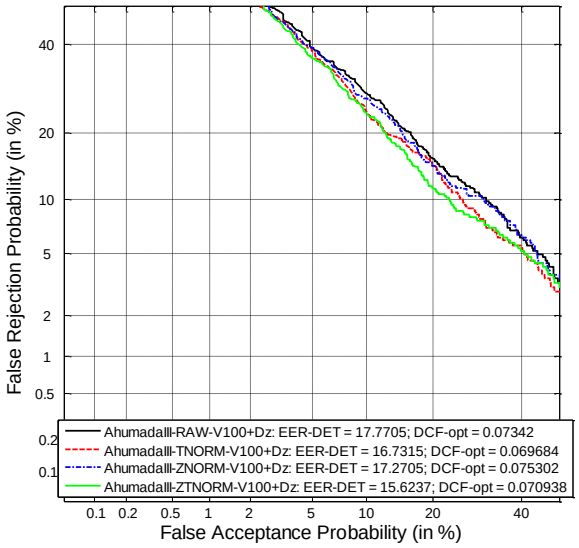


Figura 6-24. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 100 *eigenvoices*. Base de Datos Ahumada III.

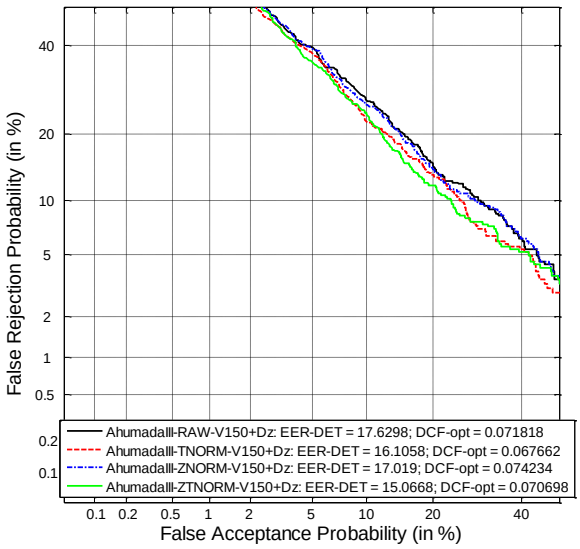


Figura 6-25. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 150 *eigenvoices*. Base de Datos Ahumada III.

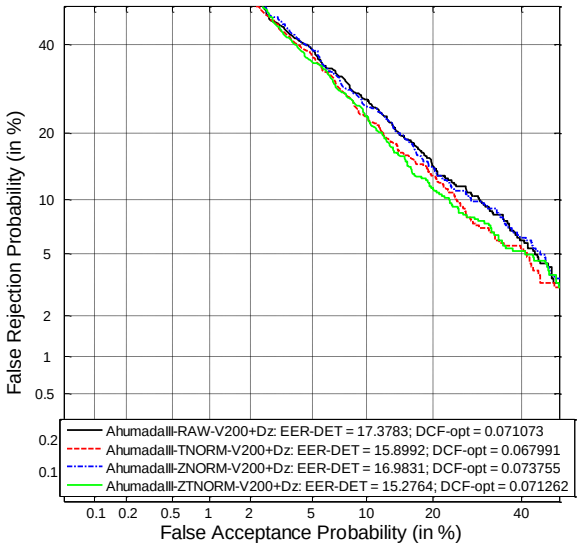


Figura 6-26. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 200 *eigenvoices*. Base de Datos Ahumada III.

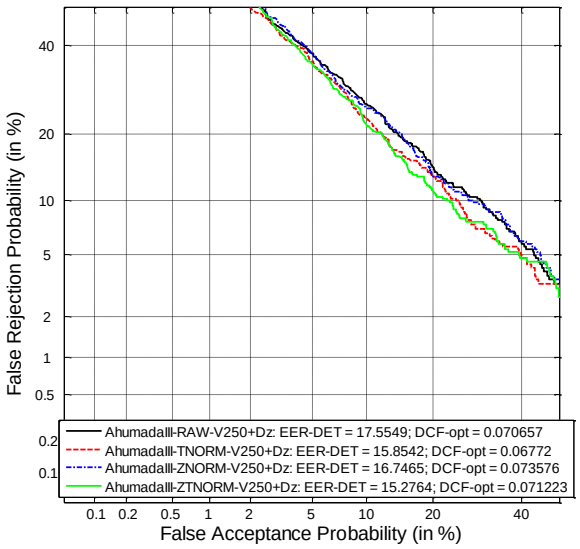


Figura 6-27. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 250 *eigenvoices*. Base de Datos Ahumada III.

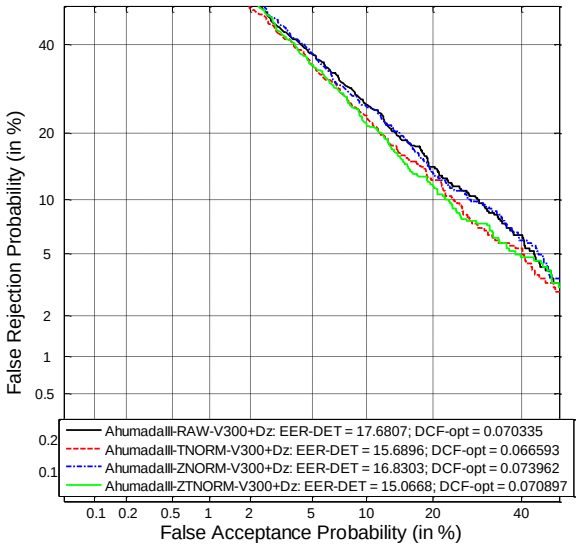


Figura 6-28. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 300 *eigenvoices*. Base de Datos Ahumada III.

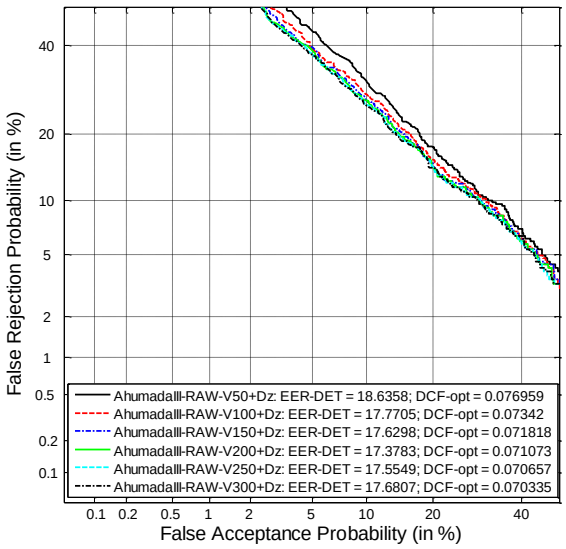


Figura 6-29. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con diferente número de *eigenvoices*. Gráfica comparativa. Base de Datos Ahumada III.

De los resultados anteriores podemos determinar que el resultado óptimo para las puntuaciones RAW (sin normalizar) se obtiene para 200 *eigenvoices*, generándose una tasa de error 1 punto más baja que en el caso de 50 *eigenvoices*. Para normalización en cero y test, la puntuación óptima pertenece al caso de 150 *eigenvoices*. En este sentido podemos concluir que no es necesario llegar a 300 direcciones de máxima variabilidad entre locutores para conseguir los mejores resultados. Así, podemos llegar al máximo rendimiento ahorrando coste computacional.

En cuanto a la variación de los resultados con respecto al número de *eigenvoices* utilizados, no se observan grandes cambios en la tasa de error, y prácticamente ninguno en la función de coste a medida que se van incrementando los *eigenvoices*, lo cual puede deberse a que se está trabajando con archivos de diferente tipo en las bases de datos utilizadas para entrenar la variabilidad y en las bases de datos utilizadas para las pruebas.

6.5.3. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando únicamente el número de *eigenvoices*

En este apartado se realizará tanto modelado de la variabilidad de locutor como compensación de la variabilidad de canal. Para ello se utilizarán técnicas de JFA.

Utilizando un número variable de *eigenvoices* (entre 50 y 300 en incrementos de 50) y 10 iteraciones EM entrenamos la matriz con la que modelaremos la variabilidad interlocutor en los modelos de entrenamiento y los de la cohorte de normalización en test.

Con 50 *eigenchannels* y 10 iteraciones EM, entrenamos una matriz de variabilidad de canal con la que compensamos en canal los estadísticos de entrenamiento y test así como los de las cohortes de t-norm y z-norm.

Los resultados obtenidos tras realizar estas pruebas se resumen en la siguiente tabla, así como en las gráficas a continuación de la misma.

NEIGENVOICES	RAW	TNORM	ZNORM	ZTNORM
50	12.57/0.049	**11.16/0.049**	11.95/0.052	**11.29/0.055**
100	12.13/0.048	**11.13/0.049**	12.62/0.052	**10.91/0.055**
150	11.87/0.049	11.53/0.050	12.37/0.052	11.52/0.054
200	12.16/0.049	11.32/0.049	12.24/0.051	11.56/0.054
250	12.16/0.048	11.32/0.050	11.95/0.051	11.37/0.053
300	12.16/0.048	11.53/0.049	11.74/0.051	10.99/0.053

Tabla 6-6. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenvoices*. Base de Datos Ahumada III.

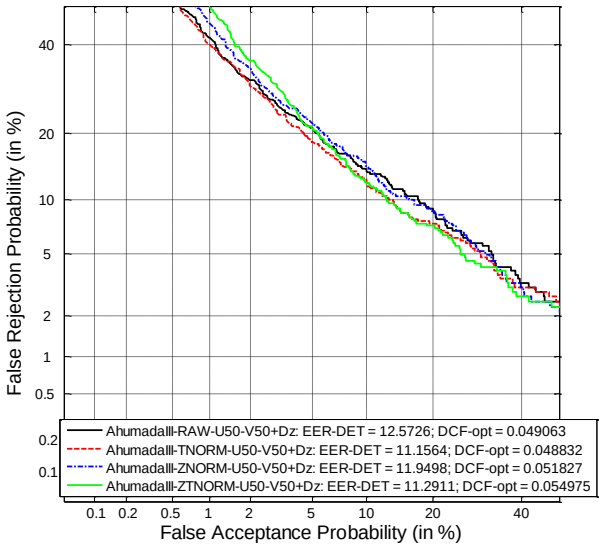


Figura 6-30. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 50 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos Ahumada III.

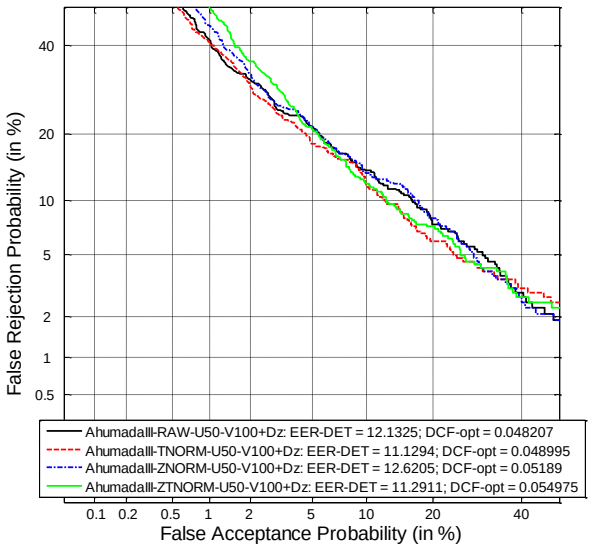


Figura 6-31. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 100 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos Ahumada III.

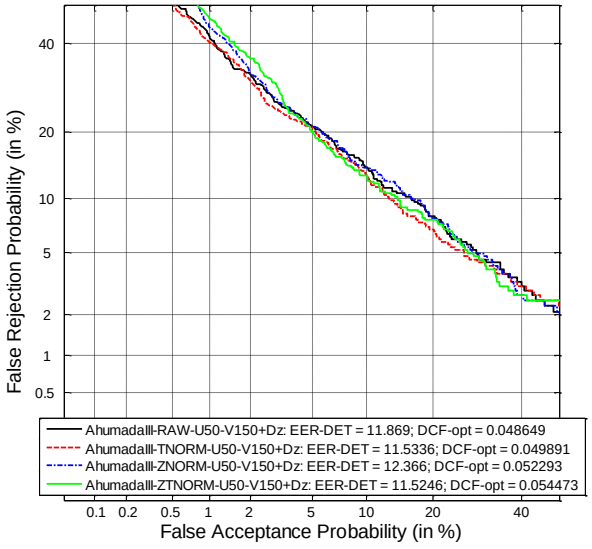


Figura 6-32. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 150 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos Ahumada III.

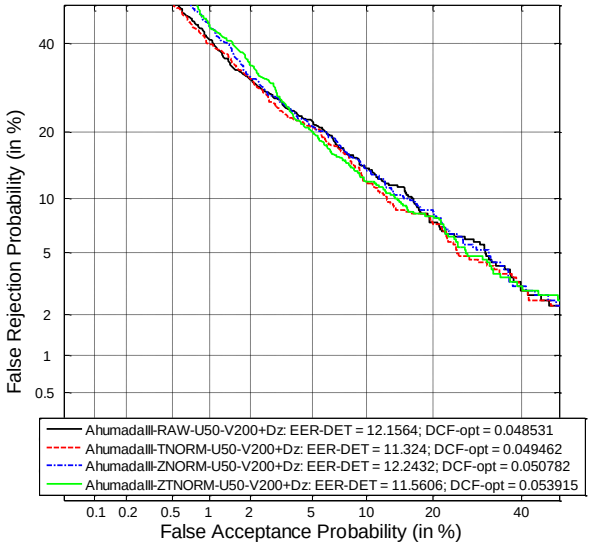


Figura 6-33. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 200 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos Ahumada III.

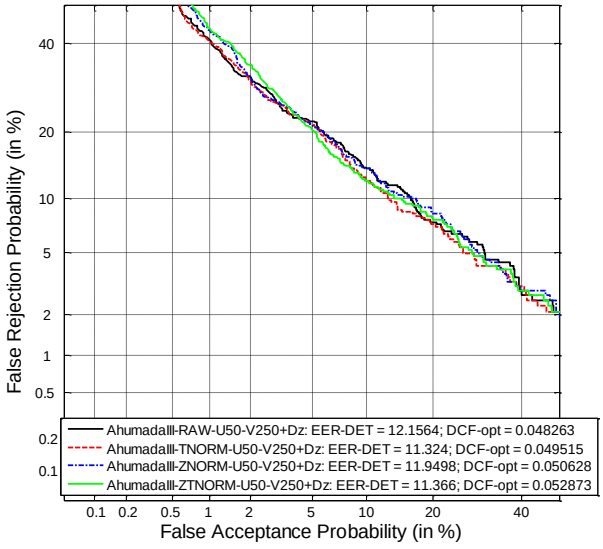


Figura 6-34. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 250 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos Ahumada III.

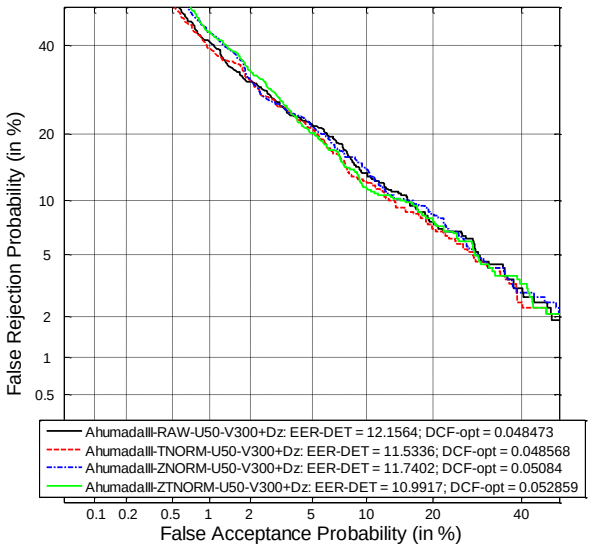


Figura 6-35. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 300 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos Ahumada III.

De nuevo ha sucedido algo similar al apartado anterior, las puntuaciones óptimas se obtienen para 150 *eigenvoices* en el caso sin normalizar y para 100 tras normalizar en cero y test, es decir, sigue sin ser necesario utilizar 300 direcciones de máxima variabilidad de locutor.

Si bien es cierto que, de forma global, las mejores puntuaciones y funciones de coste para todos los tipos de normalización se consiguen en el caso de 300, pero la variación de los resultados es tan poco significativa que no compensa con respecto al coste computacional del entrenamiento de las matrices de variabilidad.

6.5.4. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando únicamente el número de *eigenchannels*

En este apartado se realizará tanto modelado de la variabilidad de locutor como compensación de la variabilidad de canal. Para ello se utilizarán técnicas de JFA.

Utilizando 300 *eigenvoices* y 10 iteraciones EM entrenamos la matriz con la que modelaremos la variabilidad interlocutor en los modelos de entrenamiento y los de la cohorte de normalización en test.

Con un número variable de *eigenchannels* (entre 10 y 50 en incrementos de 10) y 10 iteraciones EM, entrenamos una matriz de variabilidad de canal con la que compensamos en canal los estadísticos de entrenamiento y test así como los de las cohortes de t-norm y z-norm.

Los resultados otorgados por estas pruebas se resumen en la siguiente tabla, así como en las gráficas a continuación de la misma.

NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	12.37/0.055	11.32/0.050	13.81/0.058	12.92/0.056
20	11.25/0.054	**10.49/0.051**	12.78/0.057	11.95/0.057
30	11.85/0.053	11.12/0.050	11.59/0.054	11.95/0.055
40	11.32/0.050	10.93/0.048	11.64/0.053	11.62/0.052
50	12.16/0.048	11.53/0.049	11.74/0.051	10.99/0.053

Tabla 6-7. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels*. Base de Datos Ahumada III.

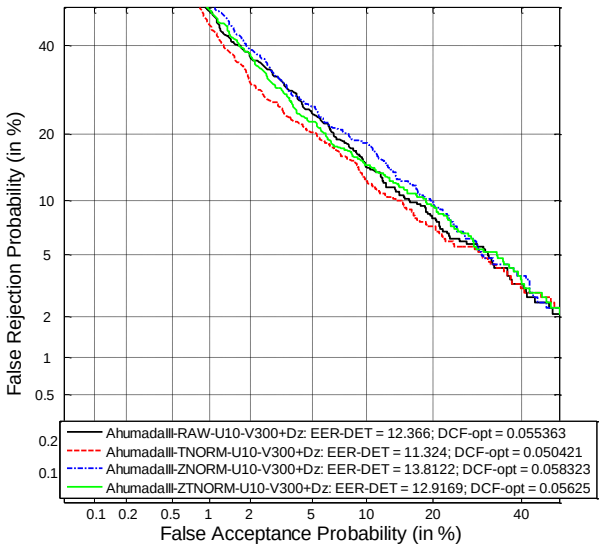


Figura 6-36. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 10 *eigenchannels*. Base de Datos Ahumada III.

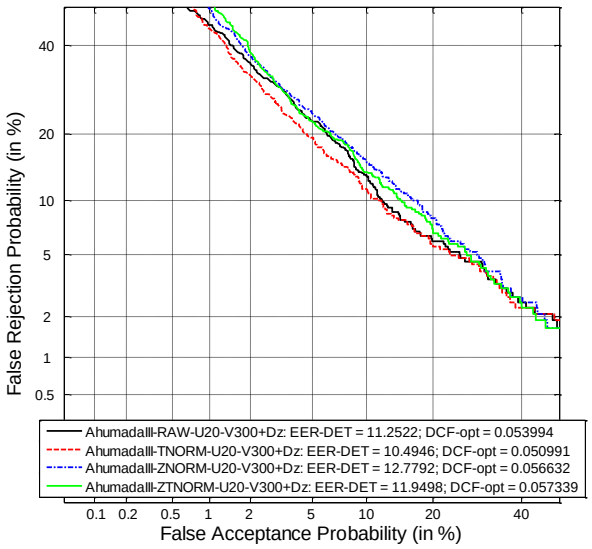


Figura 6-37. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 20 *eigenchannels*. Base de Datos Ahumada III.

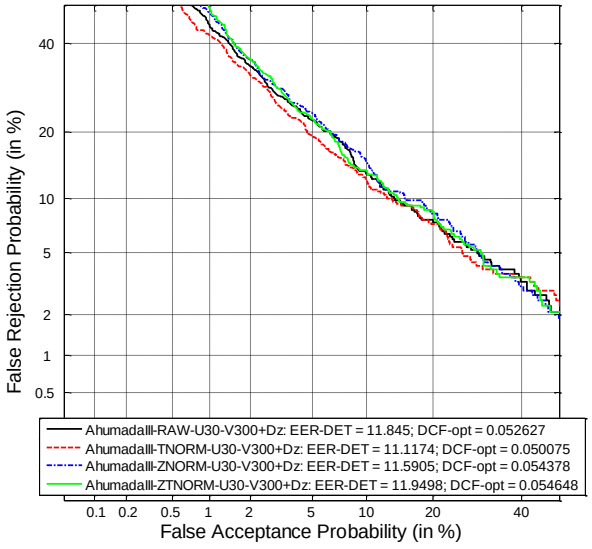


Figura 6-38. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 30 *eigenchannels*. Base de Datos Ahumada III.

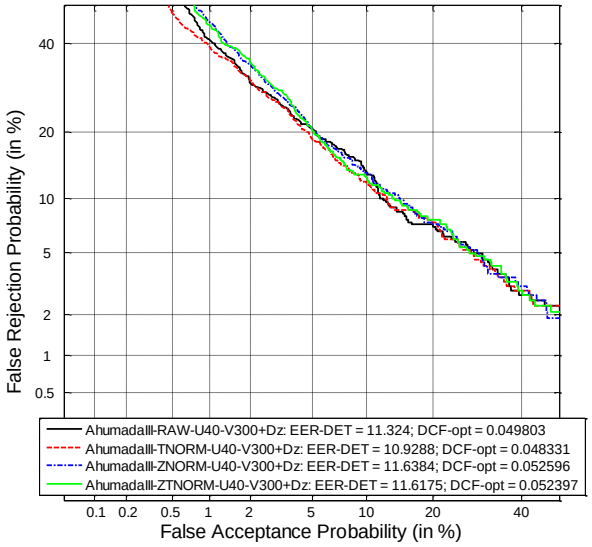


Figura 6-39. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 40 *eigenchannels*. Base de Datos Ahumada III.

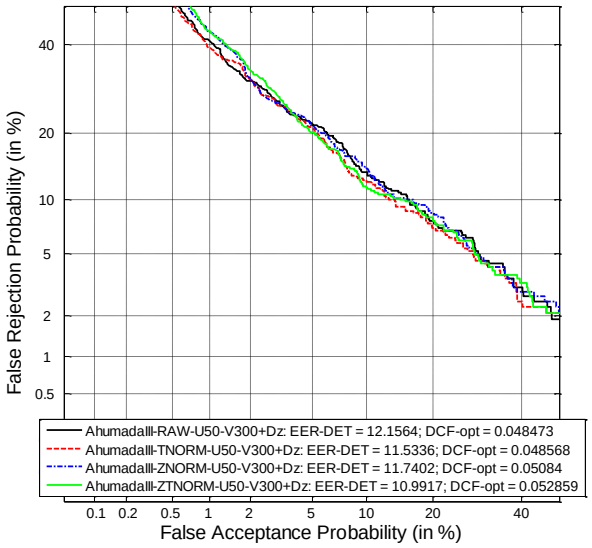


Figura 6-40. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 50 *eigenchannels*. Base de Datos Ahumada III.

Observando los resultados obtenidos podemos comprobar que únicamente en el caso de las puntuaciones con normalización en cero y test se consigue un resultado óptimo para 50 *eigenchannels*. En los casos RAW (sin normalizar) y TNORM, la puntuación más baja se obtiene para 20 *eigenchannels* mientras que en ZNORM es así para 30 *eigenchannels*. Estos hechos nos

llevan a concluir de nuevo que no es necesario entrenar la matriz con el máximo de direcciones de variabilidad puesto que no afectan significativamente a los resultados.

Además, como ocurría en apartados anteriores, la diferencia entre puntuaciones o, especialmente, entre funciones de coste, no se aprecia demasiado. El motivo puede ser el hecho de que las matrices de variabilidad están entrenadas con datos de unas bases de datos completamente diferentes a la base de datos principal utilizada para realizar los experimentos.

6.5.5. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando *eigenchannels* y *eigenvoices*

Por último, se realizará tanto modelado de la variabilidad de locutor como compensación de la variabilidad de canal variando tanto *eigenvoices* como *eigenchannels* para comprobar el efecto de la variación conjunta. Se utilizarán técnicas de JFA.

Utilizando una cantidad variable de *eigenvoices* (entre 50 y 300 incrementándose de 50 en 50) y 10 iteraciones EM entrenamos la matriz con la que modelaremos la variabilidad interlocutor en los modelos de entrenamiento y los de la cohorte de normalización en test.

Con un número variable de *eigenchannels* (entre 10 y 50 en incrementos de 10) y 10 iteraciones EM, entrenamos una matriz de variabilidad de canal con la que compensamos en canal los estadísticos de entrenamiento y test así como los de las cohortes de t-norm y z-norm.

Los resultados obtenidos en estas pruebas se resumen en las siguientes tablas, así como en las gráficas posteriores.

6.5.5.1. Rendimiento con 50 *eigenvoices* y *eigenchannels* variables

	50 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	13.53/0.060	12.22/0.054	13.82/0.060	12.99/0.060
20	11.53/0.055	10.60/0.052	13.38/0.056	11.98/0.058
30	11.95/0.053	10.91/0.048	12.78/0.054	11.12/0.056
40	11.53/0.049	10.70/0.048	12.04/0.054	10.91/0.057
50	12.57/0.049	11.16/0.049	11.95/0.052	11.29/0.055

Tabla 6-8. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 50 *eigenvoices*. Base de Datos Ahumada III.

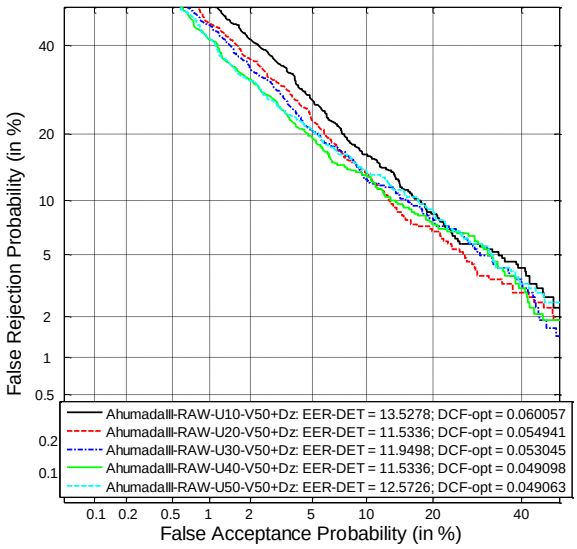


Figura 6-41. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 50 *eigenvoices* y *eigenchannels* variables. Base de Datos Ahumada III.

6.5.5.2. Rendimiento con 100 *eigenvoices* y *eigenchannels* variables

	100 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	12.78/0.058	11.57/0.052	13.82/0.058	12.73/0.057
20	11.10/0.055	10.44/0.051	13.61/0.058	12.37/0.060
30	12.16/0.055	10.91/0.050	12.99/0.057	11.32/0.057
40	11.06/0.049	10.99/0.048	11.95/0.053	10.70/0.056
50	12.13/0.048	11.13/0.049	12.62/0.052	10.91/0.055

Tabla 6-9. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 100 *eigenvoices*. Base de Datos Ahumada III.

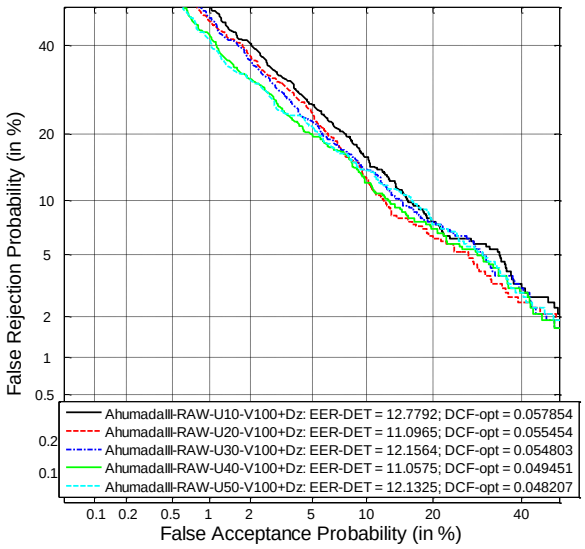


Figura 6-42. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 100 *eigenvoices* y *eigenchannels* variables. Base de Datos Ahumada III.

6.5.5.3. Rendimiento con 150 *eigenvoices* y *eigenchannels* variables

	150 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	13.13/0.057	11.41/0.052	14.23/0.059	12.50/0.056
20	11.32/0.054	10.66/0.050	12.99/0.057	11.95/0.058
30	11.82/0.054	11.17/0.050	12.71/0.057	11.59/0.055
40	11.24/0.050	11.26/0.048	11.69/0.053	11.08/0.055
50	11.87/0.049	11.53/0.050	12.37/0.052	11.52/0.054

Tabla 6-10. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 150 *eigenvoices*. Base de Datos Ahumada III.

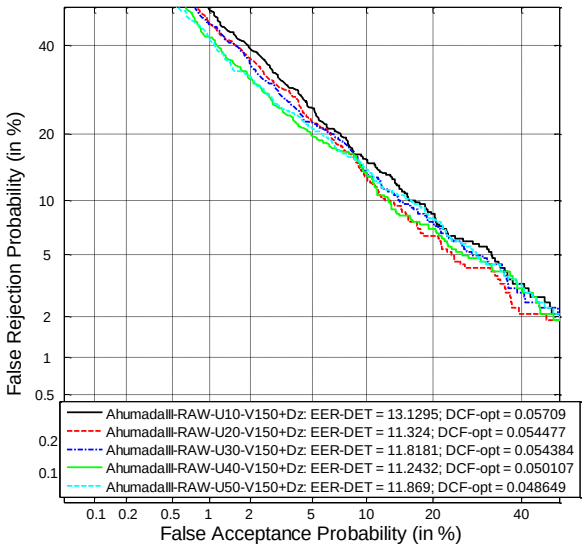


Figura 6-43. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 150 *eigenvoices* y *eigenchannels* variables. Base de Datos Ahumada III.

6.5.5.4. Rendimiento con 200 *eigenvoices* y *eigenchannels* variables

	200 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	12.45/0.056	11.45/0.051	14.03/0.058	12.57/0.056
20	11.22/0.054	10.49/0.051	12.78/0.057	12.16/0.058
30	12.16/0.054	11.61/0.049	12.02/0.056	11.88/0.055
40	11.36/0.050	11.32/0.048	11.61/0.053	11.53/0.055
50	12.16/0.049	11.32/0.049	12.24/0.051	11.56/0.054

Tabla 6-11. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 200 *eigenvoices*. Base de Datos Ahumada III.

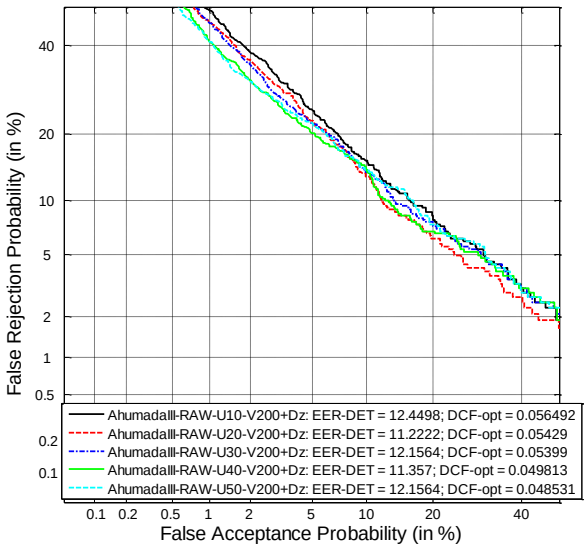


Figura 6-44. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 200 *eigenvoices* y *eigenchannels* variables. Base de Datos Ahumada III.

6.5.5.5. Rendimiento con 250 *eigenvoices* y *eigenchannels* variables

	250 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	12.50/0.056	11.56/0.051	13.82/0.058	12.81/0.056
20	11.00/0.054	10.70/0.052	12.99/0.057	12.33/0.058
30	12.08/0.053	10.91/0.051	12.00/0.054	11.69/0.055
40	11.53/0.050	11.00/0.049	11.74/0.053	11.63/0.054
50	12.16/0.048	11.32/0.050	11.95/0.051	11.37/0.053

Tabla 6-12. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 250 *eigenvoices*. Base de Datos Ahumada III.

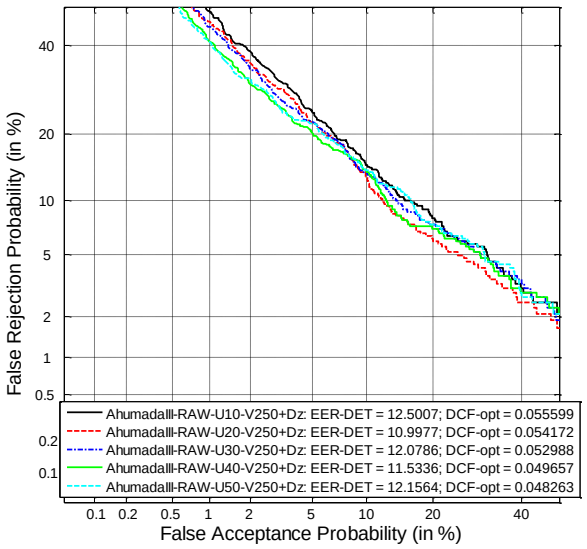


Figura 6-45. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 250 *eigenvoices* y *eigenchannels* variables. Base de Datos Ahumada III.

6.5.5.6. Rendimiento con 300 *eigenvoices* y *eigenchannels* variables

	300 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	12.37/0.055	11.32/0.050	13.81/0.058	12.92/0.056
20	11.25/0.054	10.49/0.051	12.78/0.057	11.95/0.057
30	11.85/0.053	11.12/0.050	11.59/0.054	11.95/0.055
40	11.32/0.050	10.93/0.048	11.64/0.053	11.62/0.052
50	12.16/0.048	11.53/0.049	11.74/0.051	10.99/0.053

Tabla 6-13. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 300 *eigenvoices*. Base de Datos Ahumada III.

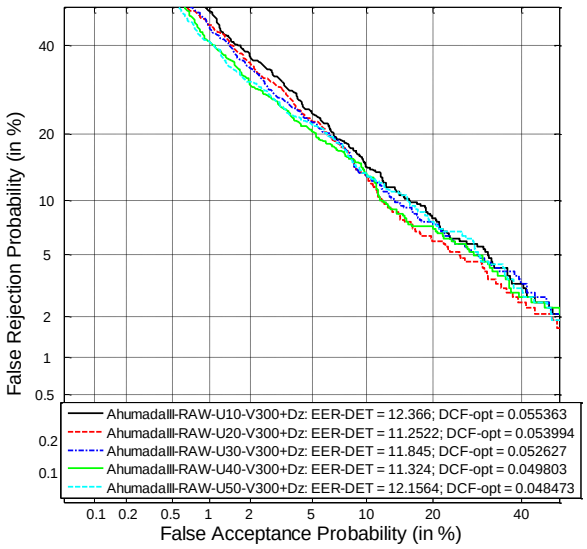


Figura 6-46. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 300 *eigenvoices* y *eigenchannels* variables. Base de Datos Ahumada III.

6.5.5.7. Comentarios

Tras estudiar los resultados obtenidos que se muestran en las tablas y gráficas anteriores, podemos determinar que la menor tasa de error obtenida es de 10.44, perteneciente al caso de 100 *eigenvoices* y 20 *eigenchannels* con normalización en test. Si tenemos en cuenta todos los casos, la menor tasa de error sin normalizar es de 11.00 para 250 *eigenvoices* y 20 *eigenchannels*, la menor en el caso de normalización en cero es 11.59 para 300 *eigenvoices* y 30 *eigenchannels* y, por último, en el caso de normalización en cero y test la menor tasa obtenida pertenece al caso de 100 *eigenvoices* y 40 *eigenchannels* y es de 10.70.

Teniendo en cuenta estos resultados queda reflejado el hecho de que no es necesario entrenar las matrices de variabilidad con el máximo número de direcciones de variabilidad posibles, al menos en el caso en el que no existe un ajuste entre los datos de entrenamiento y test, puesto que se sacrifica coste computacional y a cambio no se obtiene un mejor rendimiento.

Adicionalmente y como ya se ha comentado en apartados anteriores, los resultados obtenidos tras realizar los experimentos no varían demasiado entre sí, especialmente en términos de la DCF. Este hecho puede deberse a la diferencia de datos entre las matrices de variabilidad y los estadísticos utilizados para realizar las pruebas, perteneciendo las primeras a evaluaciones NIST SRE y los últimos a Ahumada III. Para comprobar este supuesto se han realizado los mismos experimentos utilizando para las pruebas archivos de NIST SRE 2008 de la tarea 1conv-1conv. Este estudio se verá en el siguiente apartado.

6.6. Efecto del Número de Direcciones de Máxima Variabilidad Utilizadas en Entornos Conocidos

6.6.1. Introducción

El objetivo de este apartado es comprobar la utilidad del número de direcciones de máxima variabilidad utilizadas para entrenar las matrices de variabilidad de canal y locutor. El motivo de realizar estos experimentos en entornos NIST es el hecho de que los mismos experimentos realizados en base de datos Ahumada III no ofrecían unos resultados muy significativos por lo que es deseable comprobar si esto es debido a la diferencia entre bases de datos utilizadas para las pruebas y para el entrenamiento de las matrices, o por el contrario, son otros los motivos que afectan a los resultados finales.

6.6.2. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando únicamente el número de *eigenvoices*

Los primeros experimentos que se han realizado se basan en variar el número de direcciones de máxima variabilidad de locutor. La técnica de compensación utilizada es *Joint Factor Analysis*.

Utilizando un número variable de *eigenvoices* (entre 50 y 300 en incrementos de 50) y 10 iteraciones EM entrenamos la matriz con la que modelaremos la variabilidad interlocutor en los modelos de entrenamiento y los de la cohorte de normalización en test.

Con 50 *eigenchannels* y 10 iteraciones EM, entrenamos una matriz de variabilidad de canal con la que compensamos en canal los estadísticos de entrenamiento y test así como los de las cohortes de t-norm y z-norm.

Los resultados obtenidos se ofrecen a continuación resumidos en una tabla y de forma gráfica.

NEIGENVOICES	RAW	TNORM	ZNORM	ZTNORM
50	7.65/0.046	7.84/0.040	7.58/0.043	7.57/0.041
100	7.80/0.044	7.34/0.039	7.52/0.041	7.48/0.040
150	7.41/0.044	7.16/0.039	7.33/0.041	7.35/0.039
200	7.48/0.044	7.3/0.038	7.33/0.041	7.27/0.039
250	7.41/0.044	7.31/0.038	7.48/0.040	7.37/0.039
300	7.37/0.044	7.36/0.038	7.57/0.040	7.25/0.039

Tabla 6-14. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

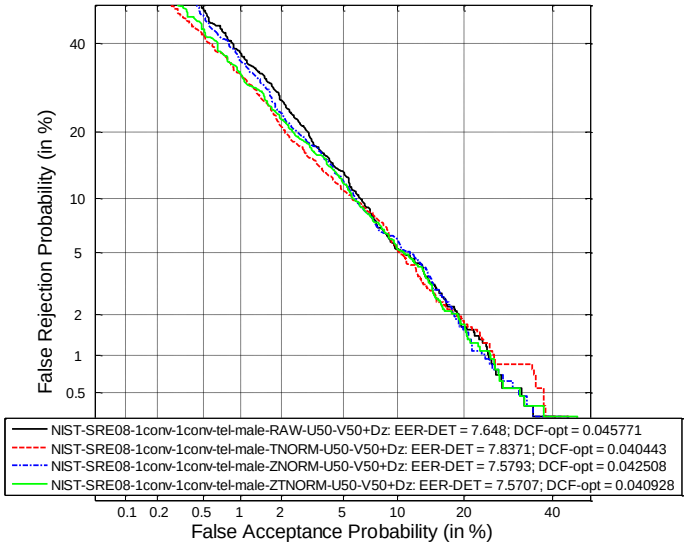


Figura 6-47. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 50 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

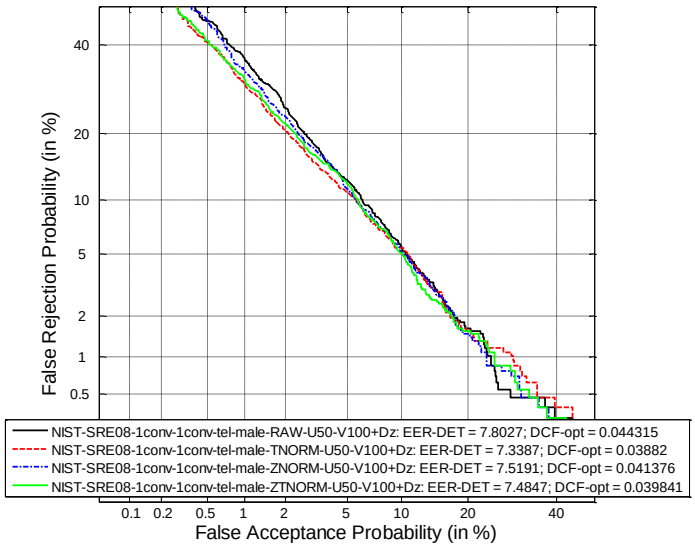


Figura 6-48. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 100 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

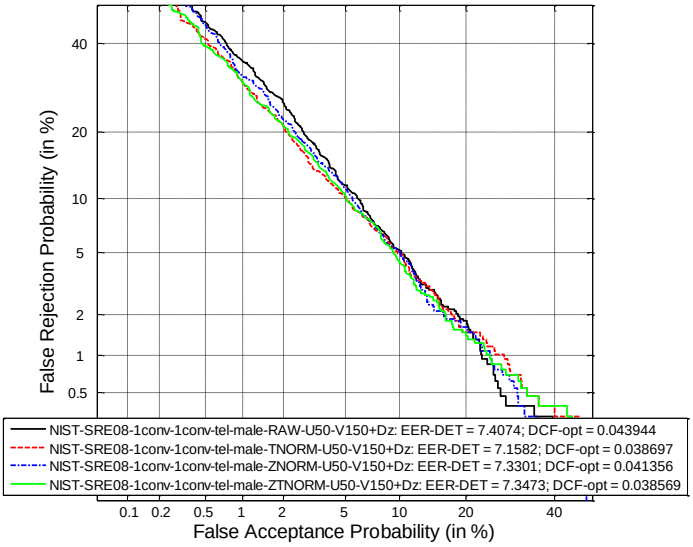


Figura 6-49. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 150 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

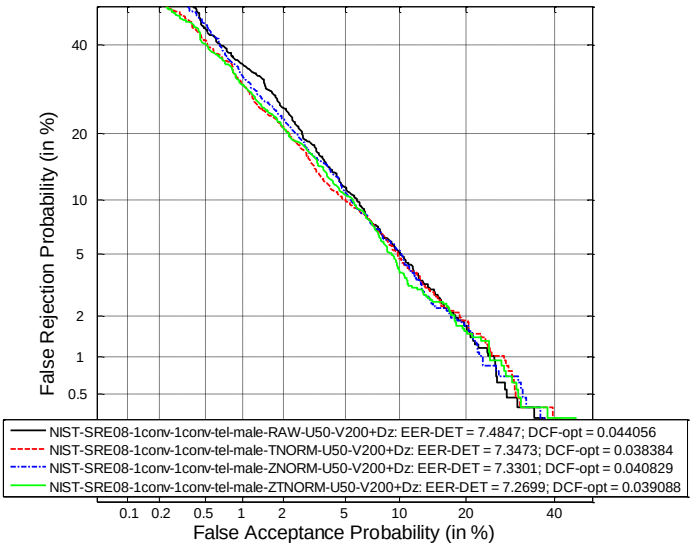


Figura 6-50. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 200 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

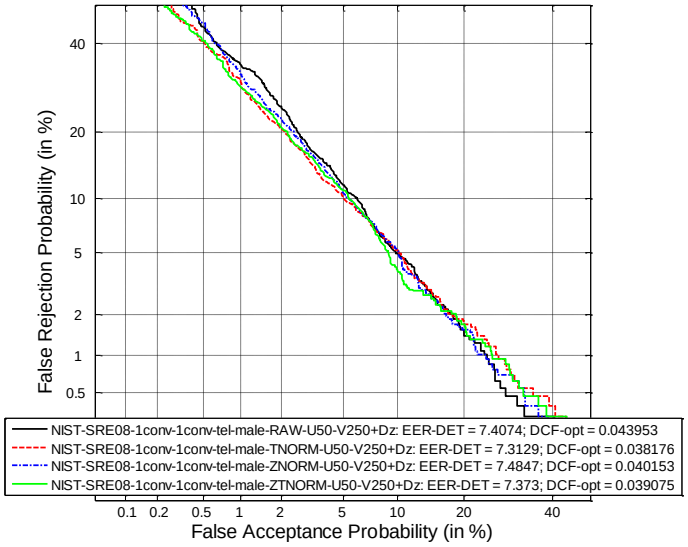


Figura 6-51. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 250 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

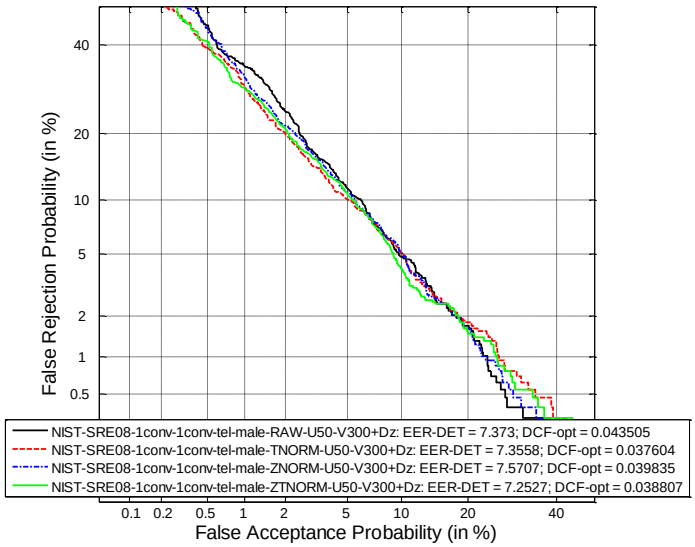


Figura 6-52. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con 300 *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

El resultado óptimo se obtiene para 150 *eigenvoices* con normalización TNORM. También para 150 *eigenvoices* tenemos el mejor resultado dentro de las puntuaciones normalizadas mediante ZNORM. Para RAW y ZTNORM la tasa de error más baja se produce para 300 *eigenvoices* pero no difiere demasiado de la conseguida en 150, por lo que podemos deducir que no es necesario entrenar la matriz con 300 *eigenvoices* sino que es suficiente con 150 para obtener resultados competitivos ahorrando coste computacional. En la siguiente gráfica vemos un ejemplo comparativo de la escasa diferencia que encontramos en los resultados variando el número de *eigenvoices* para el caso sin normalización de puntuaciones (RAW).

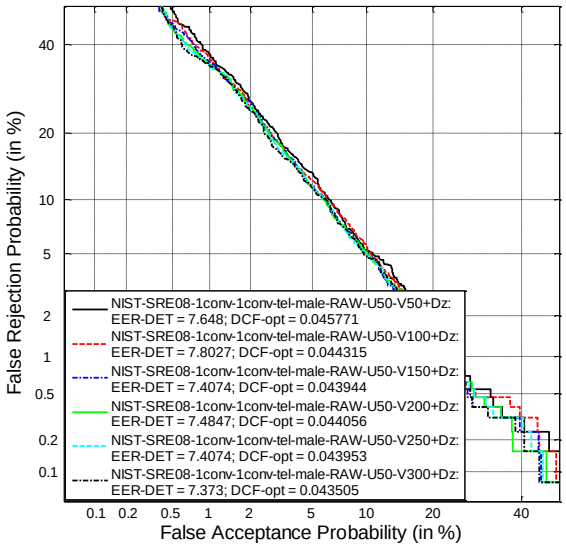


Figura 6-53. Rendimiento del Sistema tras modelar la Variabilidad de Locutor con número variable de *eigenvoices* y compensar la Variabilidad de Canal. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

6.6.3. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando únicamente el número de *eigenchannels*

De nuevo se realizará tanto modelado de la variabilidad de locutor como compensación de la variabilidad de canal. Para ello se utilizarán técnicas de JFA. El objetivo de este apartado es comprobar el efecto que tiene variar el número de direcciones utilizadas para entrenar la matriz de variabilidad en los resultados finales de las pruebas.

Utilizando 300 *eigenvoices* y 10 iteraciones EM entrenamos la matriz con la que modelaremos la variabilidad interlocutor en los modelos de entrenamiento y los de la cohorte de normalización en test.

Con un número variable de *eigenchannels* (entre 10 y 50 en incrementos de 10) y 10 iteraciones EM, entrenamos una matriz de variabilidad de canal con la que compensamos en canal los estadísticos de entrenamiento y test así como los de las cohortes de t-norm y z-norm.

Los resultados conseguidos tras estas pruebas se resumen en la siguiente tabla, así como en las gráficas a continuación de la misma.

NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	8.27/0.046	7.51/0.039	8.03/0.042	7.80/0.040
20	8.05/0.046	7.66/0.039	8.03/0.042	7.83/0.040
30	7.42/0.042	7.59/0.037	7.95/0.041	7.57/0.039
40	7.65/0.044	7.49/0.038	7.93/0.040	7.41/0.039
50	7.37/0.044	7.36/0.038	7.57/0.040	7.25/0.039

Tabla 6-15. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

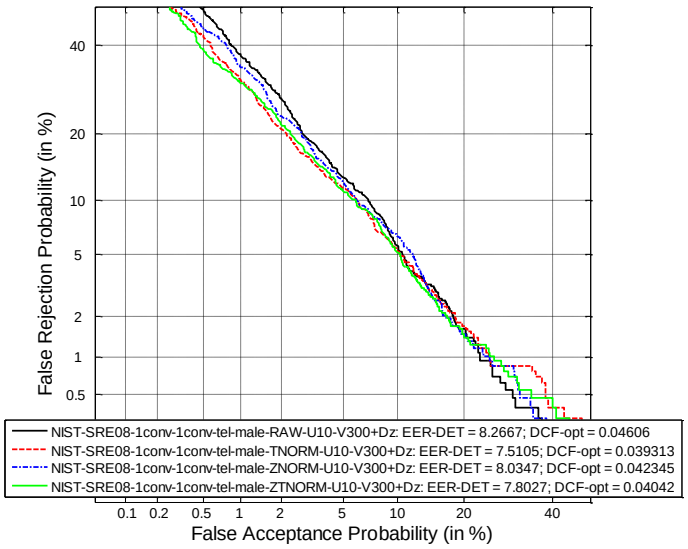


Figura 6-54. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 10 *eigenchannels*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

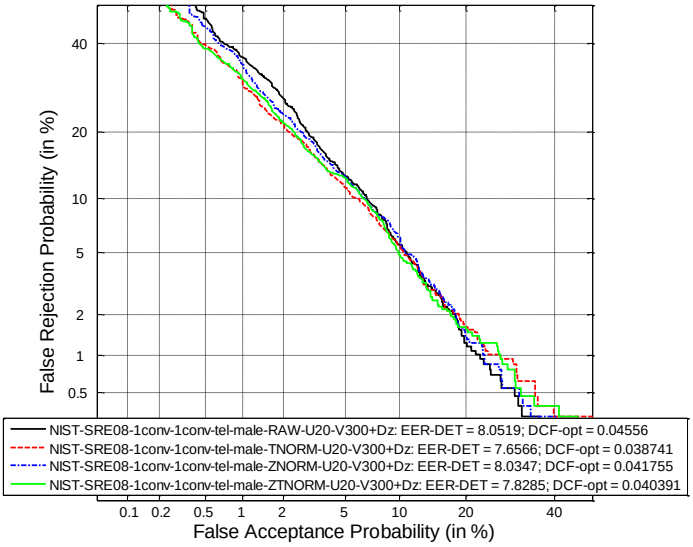


Figura 6-55. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 20 *eigenchannels*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

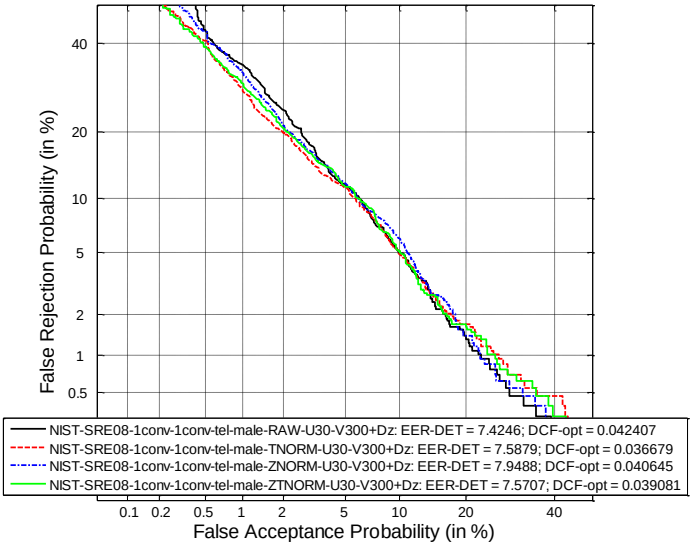


Figura 6-56. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 30 *eigenchannels*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

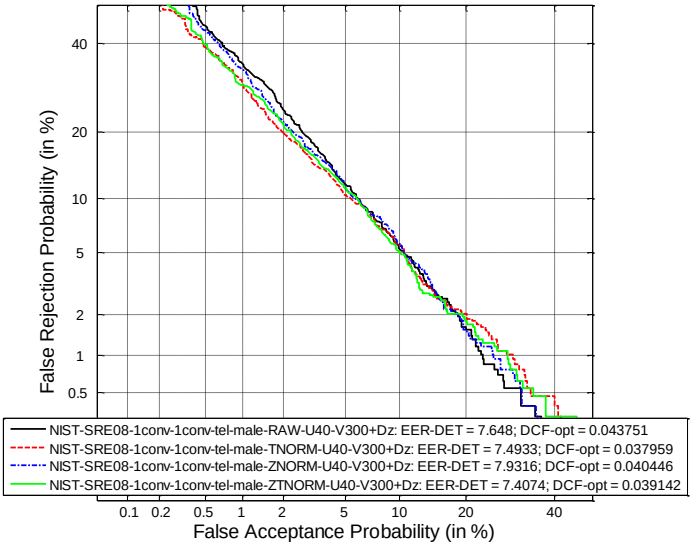


Figura 6-57. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 40 *eigenchannels*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

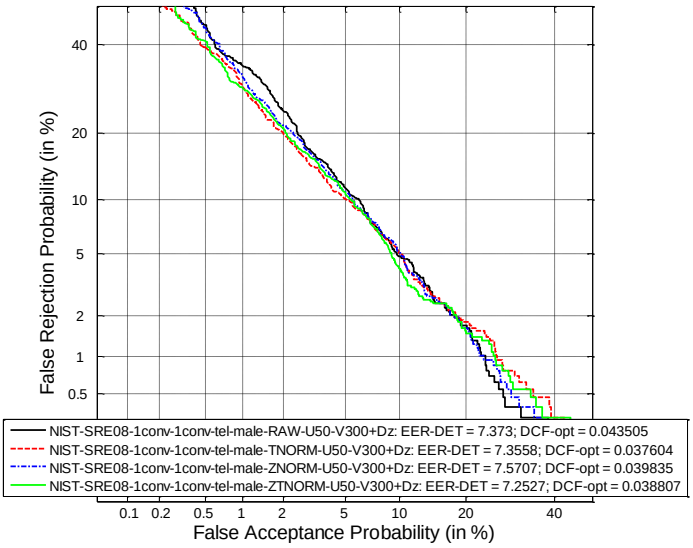


Figura 6-58. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal con 50 *eigenchannels*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

En este caso los mejores resultados se obtienen con 50 *eigenchannels* para cualquier tipo de normalización de puntuaciones. Los resultados para el caso RAW son de más de 1 punto de diferencia en la tasa de error con respecto al caso de 10 *eigenchannels* por lo que resulta útil

utilizar el máximo de direcciones de variabilidad para entrenar la matriz con la que compensaremos el canal, puesto que nos ofrece un mejor rendimiento general del sistema.

Podemos observar una tendencia a la baja de la tasa de error a medida que aumenta el número de *eigenchannels,* por lo que parecería lógico continuar incrementándolo, sin embargo, los experimentos se limitan al caso de 50 puesto que es en ese valor donde se produce la convergencia en la EER.

6.6.4. Rendimiento del Sistema tras Compensar la Variabilidad Intersesión e Interlocutor mediante JFA variando *eigenchannels* y *eigenvoices*

Como última prueba, se utilizarán las técnicas de JFA variando tanto el número de eigenvoices como de eigenchannels, comprobando así el efecto combinado de variar ambos parámetros.

Utilizando una cantidad variable de *eigenvoices* (entre 50 y 300 incrementándose de 50 en 50) y 10 iteraciones EM entrenamos la matriz con la que modelaremos la variabilidad interlocutor en los modelos de entrenamiento y los de la cohorte de normalización en test.

Con un número variable de *eigenchannels* (entre 10 y 50 en incrementos de 10) y 10 iteraciones EM, entrenamos una matriz de variabilidad de canal con la que compensamos en canal los estadísticos de entrenamiento y test así como los de las cohortes de t-norm y z-norm.

A continuación se ofrece un resumen de los resultados obtenidos.

6.6.4.1. Rendimiento con 50 *eigenvoices* y *eigenchannels* variables

	50 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	8.42/0.047	7.88/0.042	8.56/0.045	7.90/0.042
20	8.45/0.047	8.19/0.041	8.42/0.046	8.11/0.042
30	8.03/0.045	7.91/0.039	8.07/0.043	7.91/0.041
40	8.07/0.046	8.03/0.040	7.96/0.044	8.11/0.042
50	7.65/0.046	7.84/0.040	7.58/0.043	7.57/0.041

Tabla 6-16. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 50 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

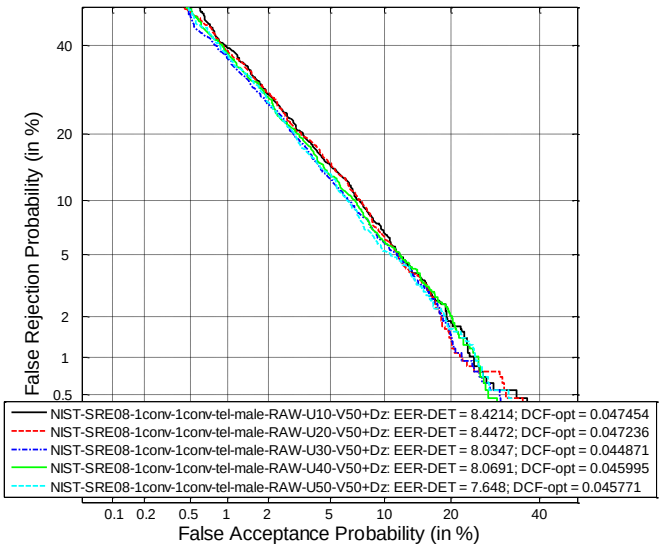


Figura 6-59. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 50 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

6.6.4.2. Rendimiento con 100 *eigenvoices* y *eigenchannels* variables

	100 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	8.32/0.046	7.74/0.040	8.25/0.044	8.03/0.041
20	8.27/0.046	7.90/0.040	8.31/0.044	7.88/0.040
30	7.80/0.043	7.64/0.038	7.96/0.042	7.80/0.040
40	8.11/0.044	7.80/0.039	8.02/0.042	7.80/0.041
50	7.80/0.044	7.34/0.039	7.52/0.041	7.48/0.040

Tabla 6-17. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 100 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

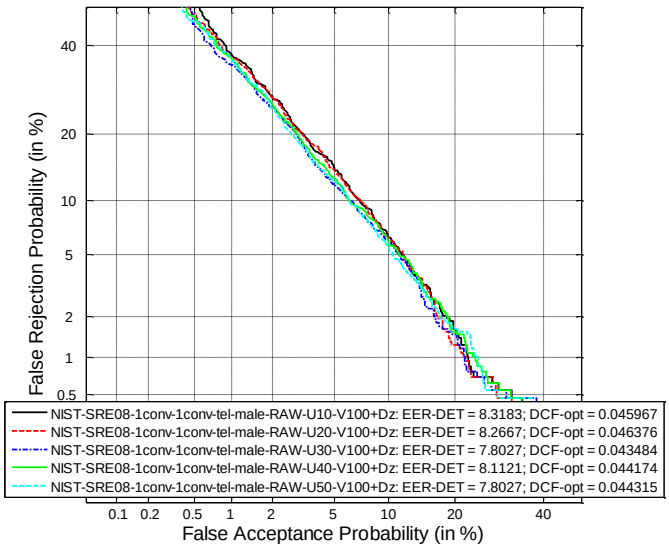


Figura 6-60. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 100 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

6.6.4.3. Rendimiento con 150 *eigenvoices* y *eigenchannels* variables

	150 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	8.33/0.047	7.49/0.040	8.03/0.043	7.48/0.041
20	8.11/0.046	7.69/0.040	8.16/0.040	7.46/0.041
30	7.72/0.043	7.41/0.037	7.76/0.041	7.33/0.038
40	7.93/0.044	7.49/0.040	7.80/0.041	7.33/0.040
50	7.41/0.044	7.16/0.039	7.33/0.041	7.35/0.039

Tabla 6-18. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 150 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

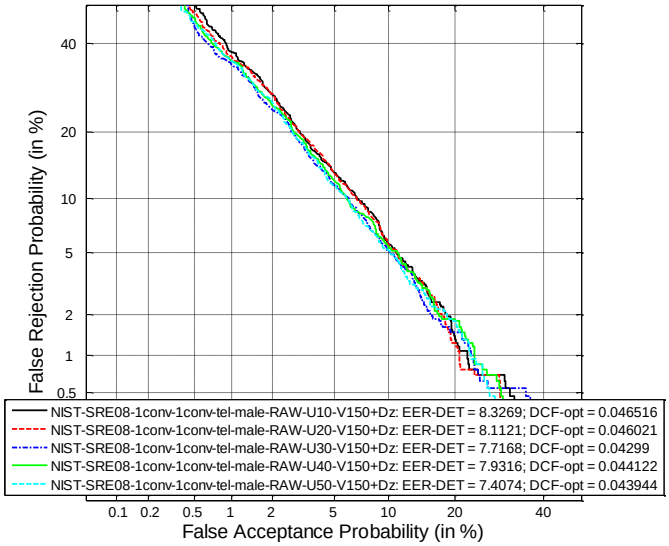


Figura 6-61. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 150 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

6.6.4.4. Rendimiento con 200 *eigenvoices* y *eigenchannels* variables

	200 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	8.35/0.046	7.61/0.040	7.98/0.043	7.73/0.041
20	8.10/0.046	7.65/0.040	8.16/0.042	7.65/0.041
30	7.48/0.043	7.64/0.038	7.73/0.041	7.44/0.039
40	7.61/0.044	7.73/0.039	7.59/0.041	7.26/0.040
50	7.48/0.044	7.35/0.038	7.33/0.041	7.27/0.039

Tabla 6-19. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 200 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

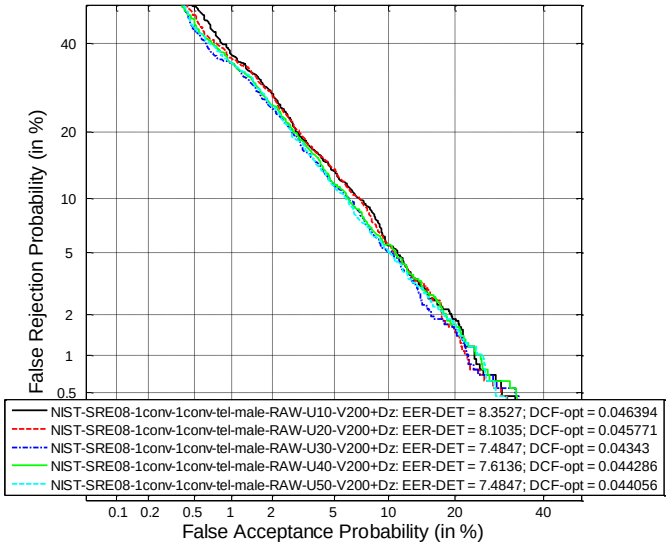


Figura 6-62. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 200 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

6.6.4.5. Rendimiento con 250 *eigenvoices* y *eigenchannels* variables

	200 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	8.25/0.046	7.48/0.040	7.90/0.043	7.72/0.041
20	8.12/0.046	7.65/0.039	8.06/0.043	7.73/0.040
30	7.67/0.043	7.76/0.037	7.85/0.041	7.52/0.039
40	7.65/0.044	7.61/0.038	7.73/0.041	7.36/0.040
50	7.41/0.044	7.31/0.038	7.48/0.040	7.37/0.039

Tabla 6-20. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 250 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

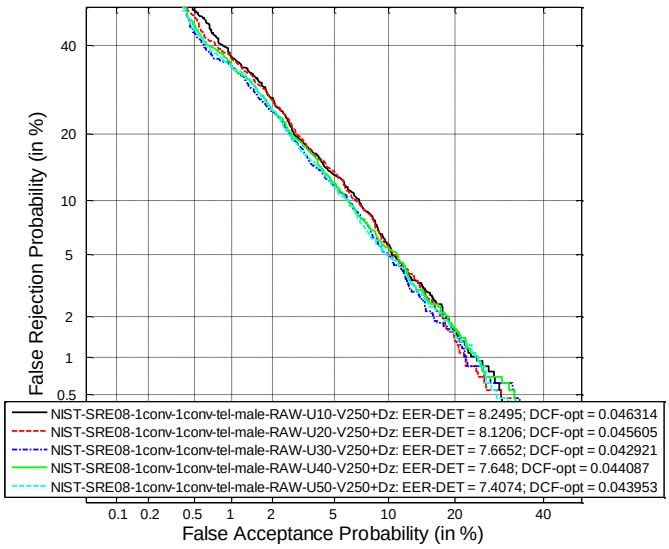


Figura 6-63. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 250 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

6.6.4.6. Rendimiento con 300 *eigenvoices* y *eigenchannels* variables

	300 EIGENVOICES			
NEIGENCHANNELS	RAW	TNORM	ZNORM	ZTNORM
10	8.27/0.046	7.51/0.039	8.03/0.042	7.80/0.040
20	8.05/0.046	7.66/0.039	8.03/0.042	7.83/0.040
30	7.42/0.042	7.59/0.037	7.95/0.041	7.57/0.039
40	7.65/0.044	7.49/0.038	7.93/0.040	7.41/0.039
50	7.37/0.044	7.36/0.038	7.57/0.040	7.25/0.039

Tabla 6-21. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 300 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

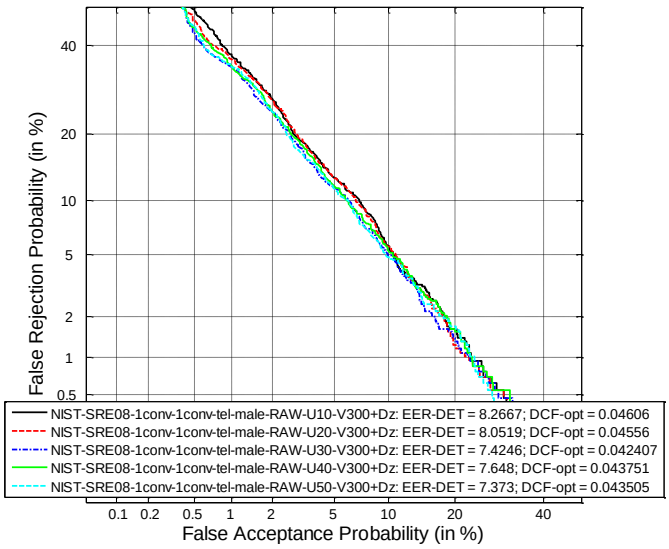


Figura 6-64. Rendimiento del Sistema tras modelar la Variabilidad de Locutor y compensar la Variabilidad de Canal variando el número de *eigenchannels* para 300 *eigenvoices*. Base de Datos NIST SRE 2008 tarea 1conv-1conv.

6.6.4.7. Comentarios

En vista de los resultados obtenidos, el caso óptimo es el realizado con 50 *eigenchannels* y 150 *eigenvoices* con normalización en test, que nos proporciona una tasa de error de 7.16. De forma general, la tasa más baja en el caso RAW es de 7.37 para 50 *eigenchannels* y 300 *eigenvoices*, en el caso ZNORM es de 7.33 para 50 *eigenchannels* y 150 *eigenvoices* y para ZTNORM tenemos una tasa mínima de 7.25 para 50 *eigenchannels* y 300 *eigenvoices*.

Observando este hecho, se puede concluir que el número de *eigenvoices* máximo utilizado para entrenar la matriz de variabilidad de locutor puede disminuirse hasta 150 sin que ello afecte significativamente a los resultados, puesto que son incluso mejores. Sin embargo, el número de *eigenchannels* sí tiene mayor peso en las tasas finales, por lo que es deseable que se mantenga en 50, que es el caso que nos proporciona los porcentajes mínimos de error.

7. Conclusiones y Trabajo Futuro

7.1. Conclusiones

En el presente proyecto fin de carrera se han realizado exhaustivos experimentos con el propósito de realizar el análisis de un sistema GMM de reconocimiento de locutor basado en *Factor Analysis* con el fin de estudiar tres aspectos fundamentales:

- Efecto de la compensación de variabilidad.
- Efecto del ajuste de las cohortes a las longitudes de la prueba.
- Efecto del número de direcciones de máxima variabilidad utilizadas.

7.1.1. Efecto de la compensación de variabilidad

De los experimentos realizados en este apartado utilizando técnicas de FA en entornos controlados y forenses se extraen las siguientes conclusiones:

- *Factor Analysis* demuestra ser una herramienta muy útil en la compensación de variabilidad intersesión en condiciones controladas (NIST SRE 2008) proporcionando reducciones de hasta el 42% en la tasa de error.
- También obtiene alto rendimiento, aunque en menor medida, en entornos forenses (Ahumada III), consiguiendo mejoras del 20% en EER.
- El éxito de *Factor Analysis* depende en gran medida del ajuste entre los datos de entrenamiento y los datos de test.
- Problema añadido: FA no es tan eficaz en duraciones cortas porque es más difícil estimar los parámetros.

7.1.2. Efecto del ajuste de las cohortes a las longitudes de la prueba

En este apartado únicamente se han realizado experimentos en entornos forenses (Ahumada III) y tras evaluar los resultados se puede concluir lo siguiente:

- Aumentar las cohortes de normalización con mayor cantidad de datos no produce mejora alguna, a menos que los datos utilizados para aumentarlas sean similares a los datos de operación.
- Ajustar las longitudes de los archivos utilizados para normalizar a las longitudes de la prueba no parece ser tampoco una medida eficaz, puesto que no reduce la variabilidad encontrada en los estadísticos.

- El ajuste de las bases de datos utilizadas en las matrices de variabilidad de canal y locutor a las duraciones de la prueba ofrece leves mejoras en el rendimiento del sistema, y es mejor que ajustar las cohortes de normalización.
- En general, el ajuste de las bases de datos a una duración menor nos proporciona resultados muy similares o peores a los obtenidos en los casos sin ajustar.

7.1.3. Efecto del número de direcciones de máxima variabilidad utilizadas

Este bloque de experimentos se centra en la variación separada y conjunta del número de *eigenvoices* y *eigenchannels* utilizados. Se han realizado experimentos en entornos forenses y en entornos controlados. Tras realizar un análisis de los resultados obtenidos se extraen las siguientes conclusiones:

- En entornos forenses, haciendo uso de datos que no son similares a los datos operacionales, el número de direcciones de máxima variabilidad tanto de locutor como de canal no afecta significativamente a los resultados obtenidos. No es útil utilizar el máximo de las mismas puesto que los resultados convergen antes.
- En entornos conocidos, el número de *eigenvoices* produce convergencia aproximadamente en 200, sin embargo, en los *eigenchannels* se observa una tendencia a la baja de la tasa de error a medida que aumentamos el número de los mismos.
- Por tanto, la variación del número de *eigenvoices* o *eigenchannels* está íntimamente ligada a la naturaleza de los datos. Mientras en entornos controlados (NIST SRE 2008) obtener más *eigenvoices*/*eigenchannels* mejora, no ocurre así en el caso de entornos forenses (Ahumada III). Éste hecho puede deberse a que cuando tratamos de ser más precisos a la hora de elegir las direcciones de máxima variabilidad funciona mejor en condiciones en las que disponemos de suficientes datos.

7.2. Trabajo Futuro

Dentro de las técnicas basadas en *Factor Analysis* en las que se centra el presente proyecto, existen dos nuevas vertientes que pueden resultar interesantes y ofrecer mejores resultados. Dichas técnicas son *Total Variability*, en la que se entrena una única matriz de variabilidad con contenido tanto de canal como de locutor, y *Variational Bayes*, que realiza una estimación Bayesiana completa de cada parámetro y variable latente.

Adicionalmente, y continuando en la línea de este proyecto fin de carrera, sería interesante comprobar hasta qué punto afecta el desajuste entre los datos de entrenamiento y test, entrenando matrices de variabilidad con datos forenses y comprobando de nuevo el efecto del número de direcciones de máxima variabilidad.

Finalmente, en este proyecto se han utilizado únicamente datos de habla pertenecientes a locutores masculinos. Supone un reto realizar los mismos experimentos utilizando habla de locutores femeninos e, incluso, bases de datos que contengan archivos de ambos sexos.

Referencias

[1] F. Bimbot, J-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz and D. A. Reynolds : *A Tutorial on Text-Independent Speaker Verification*.

[2] D. A. Reynold and, W. M. Campbell: *Text-Independent Speaker Recognition*. Springer Handbook of Speech Processing, pp. 763-781. Springer 2008.

[3] T. Kinnunen and H. Li: *An overview of text-independent speaker recognition: From features to supervectors*. Speech Communications 52 (2010) 12-40.

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

[5] D. Ramos-Castro, J. González-Rodriguez, J. González-Dominguez, and J. J. Lucena-Molina: *Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish*. ATVS – Biometric Recognition Group (EPS – UAM) & Acoustics and Image Processing Department, Criminalistic Service (D. Gral. Policía y Guardia Civil, Ministerio del Interior).

[6] P. Kenny, G. Boulianne and P. Dumouchel: *Eigenvoice Modeling With Sparse Training Data*, IEEE Transaction Speech and Audio Processing, vol. 13, pp. 345-354. 2005.

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

[8] J. A. Sigüenza Pizarro and M. Tapiador Mateos: Introducción a la Biometría. *Tecnologías biométricas aplicadas a la seguridad*, Ra-Ma 2005, pp. 3-20.

[9] V. Ruiz: *Reconocimiento biométrico de iris basado en características SIFT*. Proyecto Fin de Carrera, Septiembre 2010.

[10] S. Pérez: *Análisis y compensación de variabilidad de la señal de voz en sistemas automáticos de verificación de locutor utilizando información de duración y calidad*. Proyecto Fin de Carrera, Julio 2010.

[11] Proceedings of the National Academy of Sciences of the USA: *Simulation of talking faces in the human brain improves auditory speech recognition***.** Available at http://www.pnas.org/content/105/18/6747/F2.expansion.html, 2008.

[12] Evaluseek Publishing 2005: *Biometrics History – Looking at Biometric Technologies from Past to Present*. Available at http://www.compute-rs.com/en/advice-91803.htm.

[13] Biometría Argentina: *Historia de la Biometría*. Disponible en http://www.biometria.gov.ar.

[14] W. Campbell, D. Sturim and D. Reynolds: *Support vector machines using GMM supervectores for speaker verification.* IEEE Signal Process. Lett. 13 (5), 308-311. 2006.

[15] International Biometric Group: *Biometrics Market and Industry Report 2009-2014* (December 2009). Available at http://www.biometricgroup.com/reports/public/market_report.php.

[16] J. L. Wayman, A. K. Jain, D. Maltoni and D. Maio: *Biometric Systems: Technology,*

Design and Performance Evaluation. Springer 2006.

[17] M. Martinez: *Vulnerabilidades en sistemas de reconocimiento de huella dactilar: ataques hill-climbing*. Proyecto Fin de Carrera, Septiembre 2006.

[18] D. J. Hurley, B. Arbab-Zavar and M. S. Nixon: The Ear as a Biometric. *Handbook of Biometrics*. Springer 2008, pp. 131-150.

[19] D. Ramos-Castro: *Forensic evaluation of the evidence using automatic speaker recognition systems*. Tésis Doctoral, Noviembre 2007.

[20] M. Puertas: *Cálculo del peso de la evidencia forense utilizando sistemas biométricos*. Proyecto Fin de Carrera, Febrero 2010.

[21] C. Aitken and F. Taroni: Uncertainty in forensic science. *Statistics and the Evaluation of Evidence for Forensic Scientists.* Wiley 2004, pp. 1-34.

[22] A. J. Mansfield and J. L. Wayman: *Best Practices in Testing and Reporting Performance of Biometric Devices*. Centre for Mathematics and Scientific Computing, 2002.

[23] V. Espinosa Duró: Huella Dactilar. *Tecnologías biométricas aplicadas a la seguridad*, Ra-Ma 2005, pp. 83-112.

[24] C. Sánchez Ávila: Iris y Retina. *Tecnologías biométricas aplicadas a la seguridad*, Ra-Ma 2005, pp. 113-142.

[25] R. Sánchez Reíllo: Geometría de la Mano. *Tecnologías biométricas aplicadas a la seguridad*, Ra-Ma 2005, pp. 143-164.

[26] M. Tapiador Mateos and J. A. Sigüenza Pizarro: Escritura Manuscrita. *Tecnologías biométricas aplicadas a la seguridad*, Ra-Ma 2005, pp. 223-246.

[27] M. Tapiador Mateos and J. A. Sigüenza Pizarro: Dinámica de Tecleo. *Tecnologías biométricas aplicadas a la seguridad*, Ra-Ma 2005, pp. 247-266.

[28] H. Chen and A. K. Jain: Automatic Forensic Dental Identification. *Handbook of Biometrics*. Springer 2008, pp. 231-252.

[29] D. Dessimo and, C. Champod: Linkages between Biometrics and Forensic Science. *Handbook of Biometrics.* Springer 2008, pp. 425-460.

[30] C. E. Vivaracho Pascual: Evaluación de Sistemas Biométricos. *Tecnologías biométricas aplicadas a la seguridad,* Ra-Ma 2005, pp. 51-80.

[31] A. E. Rosenberg, F. Bimbot and S. Parthasarathy: Overview of Speaker Recognition. *Springer Handbook of Speech Processing*, Springer 2008, pp. 725-741.

[32] J. González-Rodríguez, D. Torre Toledano and J. Ortega-García: Voice Biometrics. *Handbook of Biometrics*. Springer 2008, pp. 151-170.

[33] M. Hébert: Text-Dependent Speaker Recognition. *Springer Handbook of Speech Processing*, Springer 2008, pp. 743-760.

[34] S. Young: HMMs and Related Speech Recognition Technologies. *Springer Handbook of Speech Processing*, Springer 2008, pp. 539-557.

[35] J. Ortega-Garcia, J. Gonzalez-Rodriguez and V. Marrero-Aguilar. *AHUMADA: A large speech corpus in Spanish for speaker characterization and identification*, Speech Communication, 31 (2-3), 2000.

[36] J.P. Campbell, H. Nkasone, C. Cieri, D. Miller, K. Walker, A. F. Martin and M. A. Przyboki. *The MMSR bilingual and crosschannel corpora for speker recognition research and evaluation*. Proc. Of Oddyssey, pp. 29-32, 2004.

[37] NIST SRE. Descripción de todas las evaluaciones NIST de reconocimiento de locutor. http://www.itl.nist.gov/iad/mig//tests/spk/

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

[39] *The NIST Year 2006 Speaker Recognition Evaluation Plan.* http://www.itl.nist.gov/iad/mig/tests/sre/2006/sre-06_evalplan-v9.pdf

[40] *The NIST Year 2008 Speaker Recognition Evaluation Plan.* http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf

[41] M. A. Turk and A. P. Pentland. *Eigenfaces for recognition.* Journal of Cognitive Neuroscience, 3 (1), pp. 71-86, 2000.

[42] S. Watanabe. *Karhunen-loeve expansion and factor analysis theoretical remarks and applications*. Transactions of the Fourth Prague Conference, pp. 635-660, 1965.

[43] M. Kirby and L. Sirovich. *Application of the karhunen-loeve procedure for the characterization of human faces*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, pp. 103-108, 1990. ISSN 0162-8828.

[44] I. Jolliffe. *Principal Component Analysis*. Springer-Verlag New York Inc, 175 Fifth Ave, New York City, New York, USA, 1986.

[45] R. Kuhn, J. C. Junqua, P. Nguyen and N. Niedzielski. *Rapid speaker adaptation in eigenvoice space*. IEEE Transactions on Speech and Audio Processing, 8(6), pp. 695-707, 2000. URL http://dx.doi.org/10.1109/89.876308

[46] O. Thyes, R. Kuhn, P. Nguyen and J. Junqua. *Speaker identification and verification using eigenvoices*. International Conference on Spoken Language Processing, volume 2, pp. 242-245, 2000.

[47] P. Kenny, G. Boulianne and P. Dumouchel. *Eigenvoice Modeling With Sparse Training Data*. IEEE Trans. on Speech and Audio Processing, 13(3), pp. 345-354, 2005.

[48] P. Kenny, M. Mihoubi, and P. Dumouchel. *New map estimators for speaker recognition*. In INTERSPEECH, 2003.

[49] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel. *Speaker and Session Variability in GMM-Based Speaker Verification.* IEEE Trans. Audio Speech and Language Processing.

[50] W. Wan and W. Campbell*. Support vector machines for speaker verification and identification*. Proc. of IEEE International Workshop on Neural Networks for Signal Processing, 2000, pp. 775-784.

[51] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer and P. A. Torres-Carrasquillo. Support *vector machines for speaker and language recognition*. Computer Speech and Language, 2006, Vol. 20, pp. 210-229.

[52] S. Furui. *Cepstral analysis technique for Automatic Speaker verification*. IEEE Trans. Acoust. Speech, Signal Processing, 1981, Vol. 29, pp. 254-272.

[53] H. Hermansky and N. Morgan. *Rasta Processing of Speech*. IEEE Transactions on Speech and Audio Processing, special issue on Robust Speech Recognition, October 1994, Vol. 2, pp. 578-589.

[54] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. *A study of interspeaker variability in speaker verification*. IEEE Trans. on Audio, Speech and Language Processing, 2008, Vol. 16, pp. 980–988.

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

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

[57] R. Vogt, B. Baker y S. Sridharan. *Modelling session variability in text-independent speaker verification*. Proc. Interspeech 2005, Lisboa, Portugal, Septiembre 2005, pp. 3117-3120.

[58] Yingzi Du and Chein-I Chang : *Rethinking the effective assessment of biometric systems.* Available at http://spie.org/x17545.xml?pf=true&ArticleID=x17545

[59] C. Barras, J.-L. Gauvain: *Speaker Verification of Cellular Data*. Available at http://archives.limsi.fr/RS2003GB/CHM2003GB/TLP2003/TLP9/modelechmgb.html

[60] 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

[61] M. Gasem: *Vector Quantization.* Available at http://www.mqasem.net/vectorquantization/vq.html

[62] B. Üstün, W.J. Melsssen and L.M.C. Buydens: *Visualisation and interpretation of Support Vector Regression model.,* Analytica chemical acta 595 (2007) pp. 299-309.

[63] R. Kaundal, A.S Kapoor, G.P.S. Raghava: *A SVM-based server for rice blast prediction dedicated to the farming community*. Available at http://www.imtech.res.in/raghava/rbpred/svm.jpg

[64] S. Tribedi: *Face Recognition using Eigenfaces and Distance Classifiers: A Tutorial*. Available at http://onionesquereality.wordpress.com/2009/02/11/face-recognition-using-eigenfaces-and-distance-classifiers-a-tutorial/

[65] J. Gonzalez-Dominguez, I. Lopez-Moreno, J. Franco-Pedroso, D. Ramos, D. T. Toledano and J. Gonzalez-Rodriguez: *Multilevel and Session Variability Compensated Language Recognition: ATVS-UAM Systems at NIST LRE 2009*, *IEEE Journal on Selected Topics in Signal Processing*, IEEE, 2010.

[66] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano and J. Ortega-Garcia: *Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition*, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, n. 7, pp. 2104-2115, September 2007

[67] J. Gonzalez-Dominguez, B. Baker, R. Vogt, J. Gonzalez-Rodriguez and S. Sridharan : *On the Use of Factor Analysis with Restricted Target Data in Speaker Verification*, in *Odyssey-10: The Speaker and Language Recognition Workshop, Brno, Czech Republic*, June 2010.

[68] *Probability Distributions Used for Multivariate Modeling.* The MathWorks, Inc. 2011. Available at http://www.mathworks.de/help/toolbox/stats/brklrj3.html.

Glosario

ATVS (Área de Tratamiento de Voz y Señales

CMN (Cepstral Mean Normalization)

DCF (Detection Cost Function)

DCT (Discrete Cosine Transform)

DET (Detection Error Tradeoff)

EER (Equal Error Rate)

EM (Expectation Maximization)

FA (Falsa Aceptación)

FFT (Fast Fourier Transform)

FM (Feature Mapping)

FR (Falso Rechazo)

FW (Feature Warping)

GMM (Gaussian Mixture Model)

HMM (Hidden Markov Model)

JFA (Joint Factor Analysis)

LPCC (Linear Predictive Cepstral Coefficients)

LVM (Latent Variable Model)

MAP (Maximum A Posteriori)

MFCC (Mel Frequency Cepstral Coefficients)

MLLR (Maximum Likelihood Linear Regression)

NIST (National Institute of Standards and Technology)

PCA (Principal Component Analysis)

RASTA (RelAtive SpecTrAl)

ROC (Receiver Operating Characteristics)

SRE (Speaker Recognition Evaluation)

SVM (Super Vector Model)

TNORM (Test-Normalization)

UBM (Universal Background Model)

VQ (Vector Quantization)

ZNORM (Zero-Normalization)

ZTNORM (Zero and Test Normalization)

Presupuesto

Ejecución Material

- Compra de ordenador personal (Software incluido)..... ..... 2.000 €
- Alquiler de impresora láser durante 18 meses ..... 150 €
- Material de oficina ..... 150 €
- Total de ejecución material..... 2.300 €

Gastos generales

16 % sobre Ejecución Material ..... 368 €

Beneficio Industrial

6 % sobre Ejecución Material ..... 138 €

Honorarios Proyecto

1440 horas a 15 € / hora ..... 21600 €

Material fungible

- Gastos de impresión..... 60 €
- Encuadernación..... 200 €

Subtotal del presupuesto

Subtotal Presupuesto ..... 24160 €

I.V.A. aplicable

18% Subtotal Presupuesto ..... 4348,8 €

Total presupuesto

Total Presupuesto ..... 28508,8 €

Madrid, Julio de 2011

El Ingeniero Jefe de Proyecto

Fdo.: Eugenio Arévalo González
Ingeniero Superior de Telecomunicación

Pliego de Condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un Reconocedor Automático de Locutor en Entornos Forenses basado en Técnicas de *Factor Analysis* aplicadas a nivel acústico. 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.