UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

PROYECTO FIN DE CARRERA

ANÁLISIS Y COMPENSACIÓN DE VARIABILIDAD DE LA SEÑAL DE VOZ EN SISTEMAS AUTOMÁTICOS DE VERIFICACIÓN DE LOCUTOR UTILIZANDO INFORMACIÓN DE DURACIÓN Y CALIDAD

SERGIO PÉREZ GÓMEZ

JULIO 2010

ANÁLISIS Y COMPENSACIÓN DE VARIABILIDAD DE LA SEÑAL DE VOZ EN SISTEMAS AUTOMÁTICOS DE VERIFICACIÓN DE LOCUTOR UTILIZANDO INFORMACIÓN DE DURACIÓN Y CALIDAD

AUTOR: Sergio Pérez Gómez

TUTOR: Daniel Ramos Castro

Área de Tratamiento de Voz y Señales (ATVS)

Dpto. de Ingeniería Informática

Escuela Politécnica Superior

Universidad Autónoma de Madrid

Julio 2010

Este trabajo ha sido financiado tanto por el Ministerio de Ciencia e Innovación (TEC2009-14719-C02-01) y la cátedra UAM-Telefónica, como por una beca de colaboración adquirida a través del Ministerio de Educación de forma competitiva y evaluada con máxima nota por el consejo de departamento de la Escuela Politécnica Superior de la Universidad Autónoma de Madrid.

AGRADECIMIENTOS

Quiero dedicar este proyecto a todas aquellas personas que con su consejo y ánimo me han apoyado en esta etapa de la vida que con añoranza ya concluye.

En lo profesional, quiero agradecer a Joaquín González su confianza, por brindarme la oportunidad de formar parte de un grupo de investigación de reconocido prestigio internacional y por designarme como tutor a Daniel Ramos Castro, un magnifico profesional y una excelente persona sin cuyo apoyo ni ímpetu, no habría podido embarcarme en un trabajo de tan amplias dimensiones. En especial, quería dedicarle a mi tutor y amigo las dos publicaciones derivadas de este trabajo que con tanto empeño ha luchado junto a mí para sacar adelante.

Tampoco quería olvidarme de todos los compañeros que he tenido durante estos tres años en el grupo ATVS, en especial, de aquellos que más me ayudaron en los primeros meses a salir adelante: Javier Franco, Víctor González, Alejandro Abejón, Javier González, Ismael Mateos e Ignacio López.

Por supuesto quería mencionar el compañerismo y ánimo que me han aportado el resto de compañeros que han conseguido arrancar de mí una sonrisa cuando las cosas se complicaban, pero especialmente a Sergio Lucas, Alberto Harriero, Carlos Ortego, Jesús Marcos y Eugenio Arévalo.

De todos ellos quería destacar a Alberto Harriero, al que también quería darle las gracias por facilitarme la continuación de su trabajo aun habiendo defendido ya su proyecto fin de carrera y por supuesto por su amistad y consejo.

Y como no, quería dedicar también este trabajo a mis compañeros del día a día, con los que he vivido y revivido grandes momentos durante estos apasionantes 5 años que pronto echaré de menos y no será más que por la falta de discusiones, jornadas enteras en la escuela, largas comidas en la cafetería y noches en vela haciendo prácticas.

Por último, en lo personal, quería dar las gracias especialmente **a mis padres, Emiliano y Valeria, a mi hermano Alberto, a mi novia Lucía y a Nela**. Sin vosotros no hubiera sido capaz de levantarme una y otra vez y sacar fuerzas de donde ya no quedaban para luchar día tras día para sacar esta carrera adelante. Tampoco quería terminar estos agradecimientos sin pedir perdón por desahogar con vosotros todas mis frustraciones, porque al fin y al cabo sois la gente que más me importa.

POR TODOS ESTOS AÑOS Y A TODOS VOSOTROS, MUCHAS GRACIAS…

La suerte no es más que la ilusión del trabajo bien hecho.

Sergio Pérez Gómez

A MI FAMILIA Y AMIGOS

PALABRAS CLAVE

Biometría, sistema de reconocimiento automático de locutor, verificación, verosimilitud, variabilidad, compensación, normalización, desalineamiento, *score*, duración, longitud, modelo de entrenamiento, fichero de test, calidad, degradación, KLPC, KCEP, UBML, SNR, modelado gaussiano, regresión logística.

RESUMEN

En este proyecto se estudia el impacto de la variabilidad en el rendimiento de los sistemas automáticos de reconocimiento de locutor. Para ello, se analizarán distintas condiciones de la señal de voz como son la duración o la calidad, así como once métodos diferentes para su compensación y mejora del rendimiento de los sistemas biométricos. Estos métodos se basan en técnicas de modelado gaussiano y regresión logística para compensar el desalineamiento de las puntuaciones finales del sistema de identificación en condiciones de variabilidad. Algunos de ellos han sido definidos previamente en la literatura y otros constituyen una contribución misma de este proyecto, como se manifiesta en las dos publicaciones derivadas de este trabajo (1) (2), una de ellas de carácter internacional. Para medir la eficacia de los algoritmos implementados se ha utilizado el sistema de identificación presentado por el grupo biométrico ATVS a las evaluaciones de NIST SRE de 2008 (3) y las bases de datos de NIST SRE 2006 (4) y 2008, cuyo uso se encuentra ampliamente extendido en la actualidad.

ABSTRACT

In this project impact of variability in performance of automatic speaker recognition systems is studied. For this purpose, different conditions of speech signal like duration or quality will be analyzed, as well as eleven different methods to compensate and increasing its performance in biometric systems. These methods are based on Gaussian models and logistic regression to compensate final scores misalignment of the identification system in variability conditions. Some of them have been proposed previously in the literature and others constitute a contribution of this project as manifested in two publications resulting from this work (1) (2), one of them internationally. Measuring performance of these algorithms the identification system presented by ATVS biometric group in NIST SRE 2008 (3) and NIST SRE 2006 (4) and 2008 databases are used, whose use is really widespread.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

Índice de figuras

ÍNDICE DE TABLAS

1 INTRODUCCIÓN

1.1 MOTIVACIÓN

Actualmente el uso de técnicas de reconocimiento y autenticación mediante voz está cobrando gran relevancia en todos los ámbitos, ya que supone una forma más segura de identificación de personas que mediante tarjetas o claves que pueden ser extraviadas u olvidadas. Algunos de los ejemplos más comunes donde se engloban estos sistemas se enumeran a continuación: controles de acceso, aplicaciones telefónicas, comercio electrónico, aplicaciones forenses, domótica y accesibilidad.

Un sistema de reconocimiento automático de locutor, en adelante SRAL, es aquél que pretende identificar la pertenencia de un archivo de voz desconocido a partir de la medida de similitud o parecido frente a un modelo estadístico de locutor que representa al locutor bajo estudio. Este parecido entre las muestras de voz se denominada puntuación o *score*, y estará directamente relacionado con la tasa de error de identificación del sistema (rendimiento). Las principales ventajas de los SRAL radican en que son difícilmente falseables y poco intrusivos, es decir, el usuario no debe realizar un esfuerzo adicional que resulte incómodo para que un sistema, por ejemplo de acceso, analice su voz a costa de ganar en seguridad.

Debido a la gran diversidad de aplicaciones en las que puede implementarse un SRAL, el escenario y las condiciones de la locución del habla pueden presentar extensas diferencias, como por ejemplo en contextos en los cuales la duración del fichero de voz a identificar está limitada, como es el caso de las aplicaciones forenses donde la muestra de voz del usuario no identificado puede presentar cualquier longitud. Del mismo modo ocurre cuando la calidad de los archivos de audio presenta variabilidad de un fichero a otro, debido en parte a las condiciones de captura de la muestra (ruido de fondo, conversaciones simultáneas, fidelidad del dispositivo de adquisición) y al estado emocional del hablante en cada caso.

Trabajos anteriores demuestran que para implementar un sistema robusto frente a la existente variabilidad en cuanto a la duración de un fragmento de voz (6) o en cuanto a su calidad (7), es necesario compensar de algún modo dicha variabilidad. Como se describirá en los capítulos 6 y 7, este fenómeno puede provocar un descenso en el rendimiento fruto de un desalineamiento de las distribuciones *target* (probabilidad de que la puntuación obtenida del sistema cumpla la hipótesis de que modelo y fichero son del mismo locutor) y *non target* (probabilidad de que la puntuación cumpla la hipótesis de que modelo y

fichero son de diferentes usuarios), cuyo rango de *scores* depende de las condiciones de las muestras de voz (2). El efecto de este desalineamiento se muestra en la Figura 1.1, donde se puede apreciar la diferencia del rango de puntuaciones existente entre muestras de diferente longitud, extraídas de un mismo locutor, cuando se comparan con los mismos ficheros a reconocer. En consecuencia, este efecto supondrá un problema importante a la hora de establecer un único umbral de decisión que ayudará a resolver si la muestra de voz corresponde al locutor o no (siendo la puntuación mayor o menor que el umbral, respectivamente), o cuando se desea combinar diferentes sistemas (apartado 3.3). De hecho, la compensación de la variabilidad (en términos generales) que presenta la voz, ha sido durante la última década estudio de grandes evaluaciones de carácter competitivo como NIST SRE (*National Institute of Standards and Technology Speaker Recognition Evaluation*) (8), que tratan de evaluar de forma objetiva el rendimiento de la multitud de sistemas existentes en la actualidad implementados por la comunidad científica bajo las mismas condiciones de ensayo. Si bien existen técnicas de normalización y compensación de la variabilidad intrasesión (variaciones durante la sesión de captura) propuestas en la literatura como T-Norm (4) o *Factor Analysis* (9), el estudio llevado a cabo como parte de este trabajo demuestra que no son determinantes, por lo que abre la puerta a una investigación exhaustiva en este sentido. Sobre el análisis y la compensación de esta variabilidad versará este proyecto.

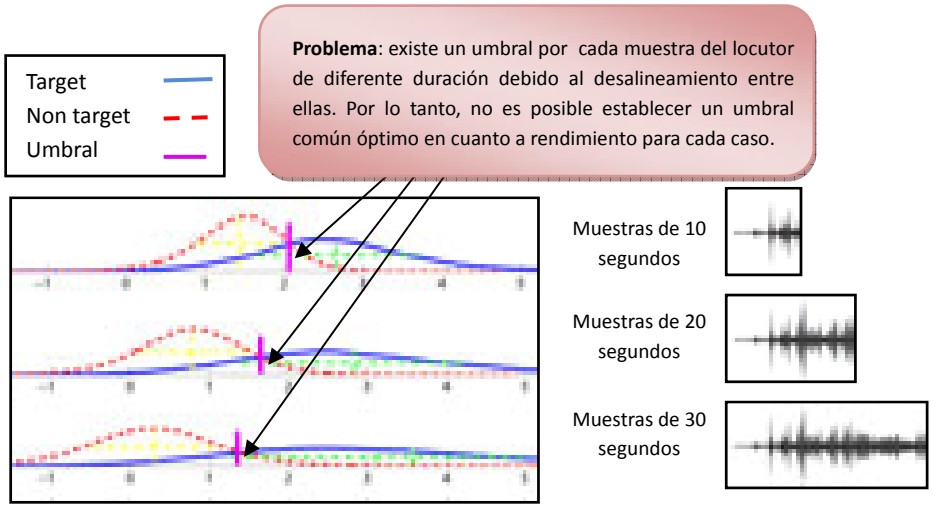


Figura 1.1. **Efecto del desalineamiento: distribución de las puntuaciones en función de la duración de las muestras de voz.**

1.2 OBJETIVOS Y METODOLOGÍA

Los objetivos que este trabajo persigue y la metodología seguida para su consecución se detallan a continuación:

1. Caracterizar de forma cuantitativa el impacto de la variabilidad en la duración y otras medidas de calidad de la voz en el rendimiento de los sistemas automáticos de reconocimiento de locutores en el estado del arte. Esta variabilidad en la calidad o en la duración de voz afectará tanto al *modelo de entrenamiento*, el cual representará al individuo de forma estadística, como a la muestra de voz o *fichero de test* cuya pertenencia se desea conocer.

2. Dado que en la actualidad no existen bases de datos de carácter abierto con suficiente variabilidad en cuanto a duración, el primer paso para acometer el estudio del impacto de la variabilidad en la longitud de las muestras, será adaptar las bases de datos de NIST SRE 2006 (9) y NIST SRE 2008 (3): para ello se crearán dos bases de datos nuevas (apartado 5.1) con ficheros de audio entre 3 y 150 segundos.

3. Una vez estudiado el efecto de la duración en el sistema, el siguiente paso es analizar de forma cuantitativa la dependencia de distintos indicadores de calidad propuestos en trabajos anteriores (8) con el rendimiento del sistema. Para ello se hará uso de las bases de datos de NIST SRE 2006 (constituida por muestras de voz telefónicas) y NIST SRE 2008 (formada por muestras telefónicas y microfónicas), lo cual permitirá medir la eficacia de los algoritmos en condiciones de mayor variabilidad.

4. El siguiente paso será implementar 11 métodos diferentes para compensar el desalineamiento fruto de la variabilidad en las condiciones del habla, algunos de ellos basados en trabajos anteriores (7) (10) (11) y otros como contribución original de este trabajo (1) (2).

5. Por último, se medirá la eficacia de estos algoritmos en cuanto a compensación del rendimiento del sistema se refiere.

1.3 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 en sistemas automáticos de reconocimiento de locutor.
- ✓ Creación de dos bases de datos con variabilidad en cuanto a duración de las muestras de voz basadas en evaluaciones actuales de carácter abierto: *DurTelSRE06 y DurTelSRE08* (apartado 5.1).
- ✓ Análisis del impacto de la variabilidad en cuanto a duración de los archivos de voz de entrenamiento y los ficheros de test (2) (capítulo 6).
- ✓ Análisis del impacto de la variabilidad en cuanto a distintas medidas de calidad de los archivos de voz de entrenamiento y los ficheros de test (1) presentadas a continuación: KLPC (*Kurtosis of Linear Prediction Coefficients*), KCEP (*Kurtosis Cepstral*), UBML (*Universal Background Model Likelihood*), SNR (*Signal to Noise Ratio*) y la recomendación P.563 de la ITU (*International Telecommunication Union*) (ver capítulo 7).
- ✓ Implementación de 11 técnicas de compensación de variabilidad en las condiciones del habla (1) (2), algunas de ellas basadas en trabajos previos encontrados en la literatura y otras de carácter novedoso como contribución misma de este proyecto.
- ✓ Estudio del rendimiento de las técnicas de compensación bajo las condiciones de NIST SRE 2006 y 2008, haciendo uso de datos telefónicos para compensar la duración y datos telefónicos y microfónicos para compensar la variabilidad en la calidad de las muestras de voz.

Remarcar que tanto el análisis del impacto de la variabilidad en las condiciones del habla como los métodos implementados han sido aceptados como artículos de congreso con revisión científica, tanto de carácter internacional como nacional (1) (2).

1.4 ORGANIZACIÓN DE LA MEMORIA

El presente documento está organizado como sigue:

El capítulo 2 describe los fundamentos y características de los rasgos biométricos haciendo hincapié en la voz, en los aspectos que la hacen válida como característica biométrica y sus ventajas como identificador. Seguidamente, los capítulos 3 y 4 abordan el estado del arte de los sistemas de reconocimiento biométrico y los sistemas de reconocimiento de locutor, describiendo para estos últimos las técnicas existentes en la actualidad de clasificación y compensación. Después, en el capítulo 5, se describe de forma detalla tanto el sistema como las bases de datos utilizadas para el desarrollo del análisis y la elaboración de los experimentos. El bloque central de esta memoria lo componen los apartados 6 y 7, donde se analiza en profundidad el efecto de la variabilidad en cuanto a duración de los archivos de voz y su calidad. A continuación se definen los métodos a utilizar para compensar su impacto (capítulo 8). Será en el apartado 9 cuando se evalúe la eficacia de los algoritmos propuesto cuyas conclusiones se comentarán en el 10, terminando la memoria con futuras líneas de investigación a seguir derivadas de este trabajo.

Al final del trabajo se presentan las referencias consultadas para la elaboración del estudio, las publicaciones derivadas del mismo y el presupuesto estimado para la elaboración de este proyecto.

2 RASGOS BIOMÉTRICOS

2.1 DEFINICIÓN Y CARACTERÍSTICAS GENERALES

Hace cuatro décadas la empresa IBM (*International Business Machines*) sugirió que el acceso de un empleado a la oficina y a su ordenador personal podía ser autenticado mediante (10):

- **Algo que conocía y memorizaba el usuario**: una clave, que puede ser olvidada o extraviada, por no mencionar que es transferible.
- **Algo que portaba consigo**: una tarjeta de identificación, que de nuevo puede ser olvidada o perdida por no olvidar que puede ser sustraída.
- **Una característica física o conductual de la persona**: un rasgo biométrico como por ejemplo la voz, intransferible, personal y que siempre se *lleva encima*, aunque poco estudiada para su implantación en sistemas de acceso automáticos e impracticable con los avances tecnológicos de entonces.

Hoy en día el notable avance en el mundo tecnológico ha hecho posible la integración de los rasgos biométricos como elementos de autenticación personal que, en la actualidad, están presentes en aplicaciones relacionadas con la seguridad, las gestiones telefónicas, el ámbito forense o la ayuda a personas dependientes. Algunos ejemplos típicos de estas aplicaciones son la realización mediante *firma on-line* de transferencias bancarias, la *huella dactilar* como medida de apoyo en la autenticación de criminales o *la voz* como herramienta de ayuda a personas con sufren algún tipo de discapacidad motriz.

Si bien la *biometría*[1] *es* como la ciencia que estudia la identificación de personas mediante atributos físicos o conductuales definidos por una serie de propiedades (11), el problema de la autenticación mediante rasgos biométricos está lejos de ser solucionado. Y es por eso que en la actualidad en muchos sistemas de autenticación se combinan las tres apreciaciones anteriores a costa de ganar en seguridad.

[1] El término se deriva de las palabras griegas "bios" de vida y "metron" de medida.

Según (11) y (12), para que un rasgo personal sea considerado biométrico deseablemente debe cumplir los siguientes criterios:

- **Universalidad**: el rasgo deben poseerlo todas las personas.
- **Unicidad**: una muestra debe identificar de manera unívoca a un usuario.
- **Permanencia**: debe permanecer invariante a medio y largo plazo.
- **Mensurabilidad**: la característica debe ser medible de forma cuantitativa.
- **Rendimiento**: debe ser preciso y no afectarle el entorno en el proceso de identificación.
- **Aceptabilidad**: los usuarios deben estar dispuestos a la captura de ese rasgo.
- **Elusión**: el rasgo debe ser capaz de eludir fraudes, por ejemplo haciendo uso de detección de vida (*liveness detection*).

2.2 RASGOS BIOMÉTRICOS TÍPICOS Y SUS APLICACIONES.

Los rasgos biométricos se pueden clasificar por orden de importancia, de mayor a menor, conforme a la siguiente división:

- **Físicos**: huella dactilar, rostro, geometría de la mano, iris, voz, retina, venas de la mano, termografía facial, geometría de la oreja, ADN (ácido desoxirribonucleico), etc.
- **Conductuales**: voz, firma, escritura, dinámica de tecleo, forma de caminar, etc.

La siguiente tabla muestra la comparación de estos atributos biométricos en cuanto al grado de posesión de los criterios citados en el apartado anterior, correspondiéndose con ↑≈↓ a un grado alto, medio y bajo respectivamente (12).

Característica / Rasgo biométrico	Universalidad	Unicidad	Permanencia	Mensurabilidad	Rendimiento	Aceptabilidad	Elusión
Huella dactilar	≈	↑	↑	≈	↑	≈	≈
Rostro	↑	↓	≈	↑	↓	↑	↑
Geometría de la mano	≈	≈	≈	↑	≈	≈	≈
Iris	↑	↑	↑	≈	↑	↓	↓
Voz	≈	↓	↓	≈	↓	↑	↑
Venas de la mano	≈	≈	≈	≈	≈	≈	↓
Retina	↑	↑	≈	↓	↑	↓	↓
Termografía facial	↑	↑	↓	↑	≈	↑	↓
Oreja	≈	≈	↑	≈	≈	↑	≈
Firma manuscrita	↓	↓	↓	↑	↓	↑	↑
Escritura	↓	↓	↓	↑	↓	≈	↑
Dinámica de tecleo	↓	↓	↓	≈	↓	≈	≈
Forma de caminar	≈	↓	↓	↑	↓	↑	≈
ADN	↑	↑	↑	↓	↑	↓	↓

Tabla 2.1. Comparación de los rasgos biométricos en cuanto a nivel de criterios (12).

Observando detalladamente la **Tabla** 2.1 se puede comprobar de forma cualitativa la viabilidad de estudiar un rasgo biométrico y su aplicabilidad en la vida real. Acerca de la voz, el rasgo biométrico que se estudia en este proyecto, se puede destacar que es una característica difícilmente falsificable ya que depende de las características fisiológicas de los órganos que intervienen en su producción (apartado 2.3) así como de factores socioculturales y por lo tanto, de alta aceptación en la sociedad como rasgo identificador. Por contra, es una característica no muy distintiva cuando se trabaja con un conjunto elevado de muestras (millones de usuarios), ya que depende del estado emocional y su tasa de error se ve afectada por las condiciones del entorno en las que se captura. También, a diferencia de otros rasgos como la huella, permite una identificación a distancia, sin necesidad de que el usuario esté presente. Una gran ventaja de este rasgo es que es un atributo fuertemente estudiado y caracterizado para el que existen hoy en día grandes avances en el campo de la biometría, como son sistemas para el reconocimiento de locutores, el reconocimiento de idioma y de texto.

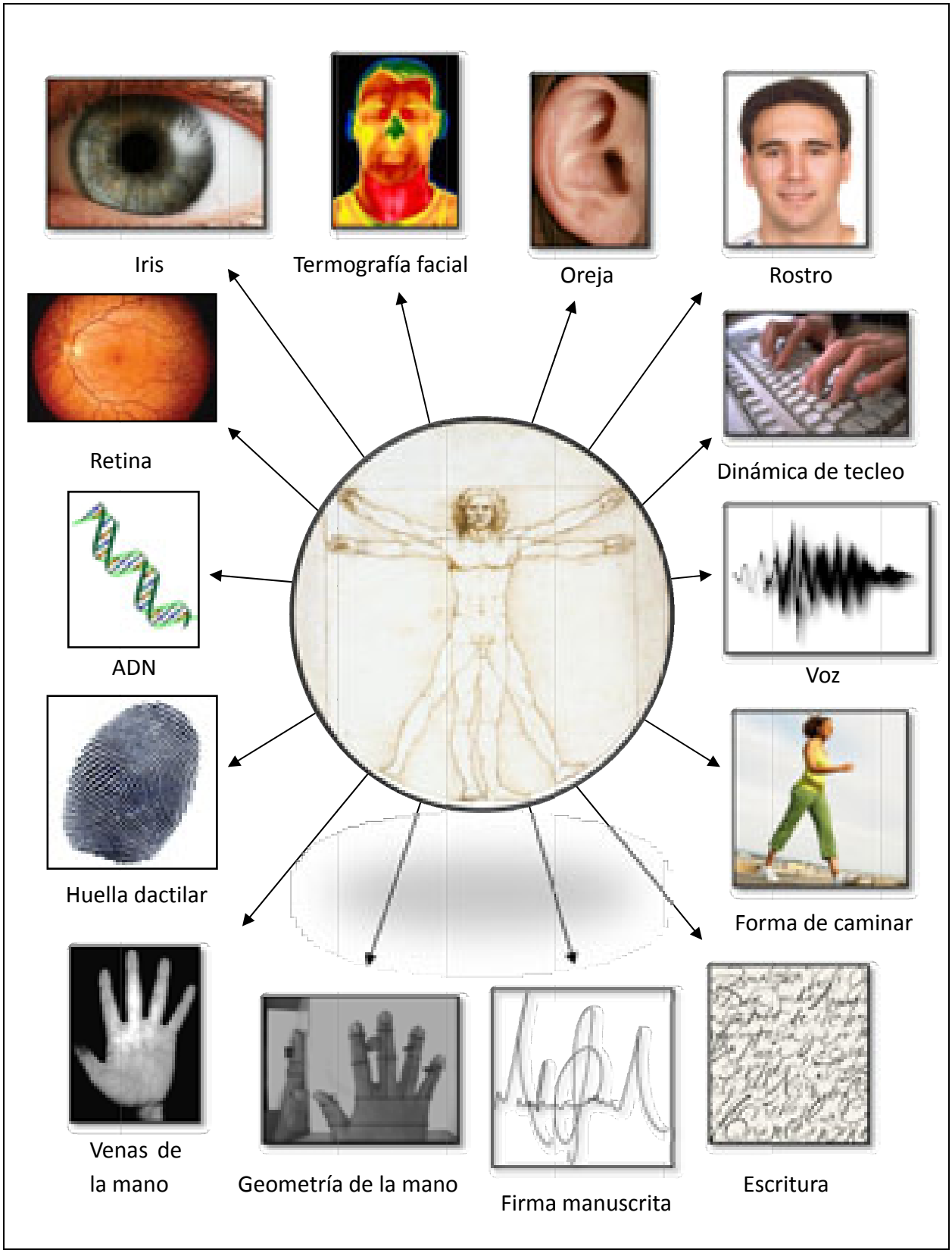


Figura 2.1. Representación de los rasgos biométricos más destacados. Figura adaptada de (11).

A continuación se describen brevemente los tipos de rasgos enunciados en la **Tabla** 2.1 y en la **Figura** 2.1.

HUELLA DACTILAR

Es el rasgo biométrico de mayor utilización debido a su gran poder discriminativo. De hecho, hoy en día constituye aproximadamente el 40% del mercado de reconocimiento biométrico. En cuanto a su caracterización, existen varias técnicas para la extracción de patrones (13), siendo la más popular la caracterización de minucias o puntos relevantes de las crestas de la huella. Algunas ventajas sobre otros rasgos biométricos son la invariabilidad con la edad, la regeneración y la facilidad de adquisición. Si bien es relativamente sencilla su falsificación se combina con técnicas detección de vida como el reconocimiento de olor, el sudor corporal o la impedancia de la piel (14).

ROSTRO

Es probablemente el método más extendido que tienen los humanos para reconocerse. Su principal ventaja es que es un rasgo de carácter no intrusivo. La aproximación clásica más popular al problema es reconocer la cara mediante la localización de formas y atributos faciales como los ojos o los labios (15). Algunas de las líneas de investigación actuales en relación al reconocimiento facial son la identificación desde distintas perspectivas, la variación en las condiciones de iluminación o la eliminación de artefactos (gafas, gorras, etc.) y cabello.

GEOMETRÍA DE LA MANO

Se basa en la realización de distintas medidas características de la mano, incluyendo forma, tamaño, la longitud y anchura de los dedos (13). Una ventaja muy importante de este rasgo es que los sistemas de adquisición son económicos, permitiendo su incorporación al mercado dado su también su carácter poco intrusivo. Algunas de sus limitaciones son: i) es un rasgo poco discriminativo, por lo que su uso está sólo recomendado para sistemas de acceso privados en el que el conjunto de usuarios es limitado. ii) requiere de un sistema de captura grande por lo que la hace inapropiado para incorporarlo a ordenadores personales o dispositivos móviles. Actualmente algunas investigaciones enfocan el hecho de mejorar la poca discriminación que ofrece la geometría de la mano mediante reconocimiento palmar, es decir, la caracterización de las líneas de la palma mediante métodos similares a los usados para la huella dactilar.

Iris

Al ser un rasgo genético queda definido por su gran poder discriminante. Prácticamente no varía desde la niñez aunque se despigmenta ligeramente con la vejez. Su principal problema radica en que la adquisición es intrusiva y requiere de la cooperación del sujeto por no hablar de la inversión económica necesaria en el sistema (16).

Voz

Constituye un rasgo ampliamente aceptado por la sociedad por ser el principal instrumento de comunicación entre las personas. Es un rasgo tanto físico como conductual (18), y por ello variable en función del estado de ánimo del sujeto, lo que lo convierte en un rasgo difícilmente falsificable. Este es el rasgo biométrico sobre el cual versará este proyecto, el cual será explicado en detalle en el apartado 2.3.

Venas de la mano

Consiste en la captura del patrón de venas de la palma y/o de los dedos para su posterior caracterización. La imagen del patrón se realiza mediante emisión de luz de alta frecuencia. Se utiliza en sistemas de alta seguridad ya que destaca por su buen rendimiento (18).

Retina

Utiliza un escáner de infrarrojos que ilumina la retina a través de la pupila para capturar los patrones de la red vascular alrededor del nervio óptico. El reconocimiento de retina es una técnica que dota a los sistemas de mucha seguridad pero es a la vez muy invasiva, por lo que su uso se encuentra poco extendido.

Termografía facial

Mide el patrón infrarrojo de la emisión de calor de la cara causado por el flujo de la sangre bajo la piel. Es una técnica de reconocimiento no invasiva que no requiere contacto físico y resuelve los problemas derivados de los artefactos así como el peinado como ocurría con el reconocimiento facial.

Oreja

Es una modalidad emergente de gran acogida ya que mediante técnicas de carácter térmico puede *esquivarse* el pelo. Es invariante y no cambia cuando se habla o se gesticula, es decir, no depende del estado de ánimo (18). Su análisis se combina con la identificación de rostro para dotar al sistema de mayor robustez.

FIRMA MANUSCRITA

Existen dos tipos de reconocimiento de firma: *on-line* y *off-line* (13). El más común como rasgo biométrico es el segundo, en el cual la firma debe realizarse sobre una superficie especial que es capaz de medir la variación de las coordenadas cartesianas provocadas por el trazo de la firma en función del tiempo. Algunos sistemas proveen también de información de inclinación y presión del elemento con el que se firma, lo que los dota de un carácter más robusto. En la actualidad, es un rasgo bastante estudiado debido tanto a su alta aceptación social como a la aparición de dispositivos táctiles que permiten su implantación inmediata en el mercado.

ESCRITURA

Es similar al reconocimiento de firma off-line. Algunos parámetros que se extraen para su reconocimiento son la dirección del texto, la inclinación, el grosor de las líneas, etc. Este tipo de identificación exige previamente la utilización de algoritmos OCR o lo que es lo mismo de reconocimiento óptico de caracteres. Sin embargo, un sistema completo de reconocimiento de escritura también maneja el formato, realiza la segmentación correcta en caracteres y encuentra las palabras más plausibles.

DINÁMICA DE TECLEO

Es un rasgo biométrico de tipo conductual. Para su captura basta con emplear secuencias de tecleo del usuario por lo que es muy poco intrusivo (13). Como rasgo biométrico es poco seguro pero puede utilizarse como identificación en entornos que no requieren de mucha privacidad.

FORMA DE CAMINAR

La forma de andar es muy distintiva de cada persona. Sin embargo los sistemas de identificación basados en este rasgo presentan un rendimiento muy por debajo de otros sistemas como usan la huella dactilar o la voz. Este rendimiento se ve afectado por factores como la vestimenta o el peso que pueda portar el sujeto (15). No obstante es un rasgo muy útil cuando sólo se tiene una grabación de la persona a identificar como ocurre en los sistemas de video-vigilancia. Este rasgo continúa en las primeras etapas de desarrollo por lo que aún se encuentra poco extendido.

ADN

El ADN (ácido desoxirribonucleico) está presente en toda célula viva y es único para cada individuo excepto para gemelos monocigóticos, lo que le otorga un alto poder discriminante. De hecho, es el método más utilizado en aplicaciones de reconocimiento en ámbitos forenses. Sin embargo, su principal desventaja es que es fácilmente substráible (un pelo bastaría) y presenta un tiempo de procesado excesivamente alto, lo que le descarta para aplicaciones que requieren de reconocimiento en tiempo real. Otra principal desventaja es que es poco aceptado ya que puede revelar aspectos genéticos del usuario tales como enfermedades (no respeta la privacidad del usuario). Pese a su alto poder discriminante, muchos científicos no lo consideran como rasgo biométrico ya que el proceso de identificación mediante ADN está lejos de ser automático.

OTRAS MODALIDADES

En función de la aplicación a implementar conviene estudiar un rasgo biométrico en particular que puede proporcionar mayor información que el resto aunque en condiciones normales su rendimiento sea peor. Algunos rasgos que se están estudiando en los últimos años son: las uñas, las arrugas de los dedos, las crestas de los nudillos, la impedancia de la piel, el olor, etc.

No obstante existen una serie de parámetros que no pueden ser considerados como biométricos por no cumplir las características anteriormente citadas pero constituyen un complemento que puede aportar fiabilidad al reconocimiento de personas. Estos parámetros, que se conocen como *soft biometrics*, pueden ser: la altura del sujeto, el color de los ojos, el peso, el sexo o la raza entre otros (11).

A continuación se representa, de manera cuantitativa, la presencia en el mercado en 2006 de los distintos rasgos biométricos definidos.

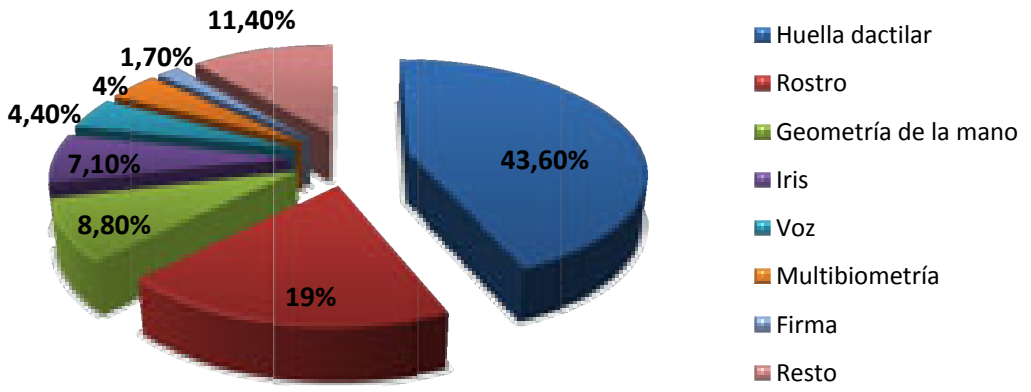


Figura 2.2. Presencia en el mercado de los diferentes rasgos biométricos (13).

Por último, destacar que el rendimiento de multitud de estos algoritmos se pone a prueba en distintas evaluaciones abiertas de carácter internacional en el que el organismo organizador define las bases de datos y protocolos a utilizar, incentivando así la investigación de empresas privadas y entidades públicas en este sentido. Un ejemplo de evaluaciones en voz son las organizadas por NIST SRE (8), que definen el marco experimental de este proyecto.

2.3 La voz como rasgo biométrico

2.3.1 Características fisiológicas y análisis espectral

La voz es una onda de presión que se genera cuando se expele el aire de los pulmones a través de la tráquea. Es también entendida como una señal que codifica mediante sonidos el lenguaje hablado. Estos sonidos vienen producidos por la relajación o tensión de las cuerdas vocales, situadas en el tracto vocal (**Figura** 2.3).

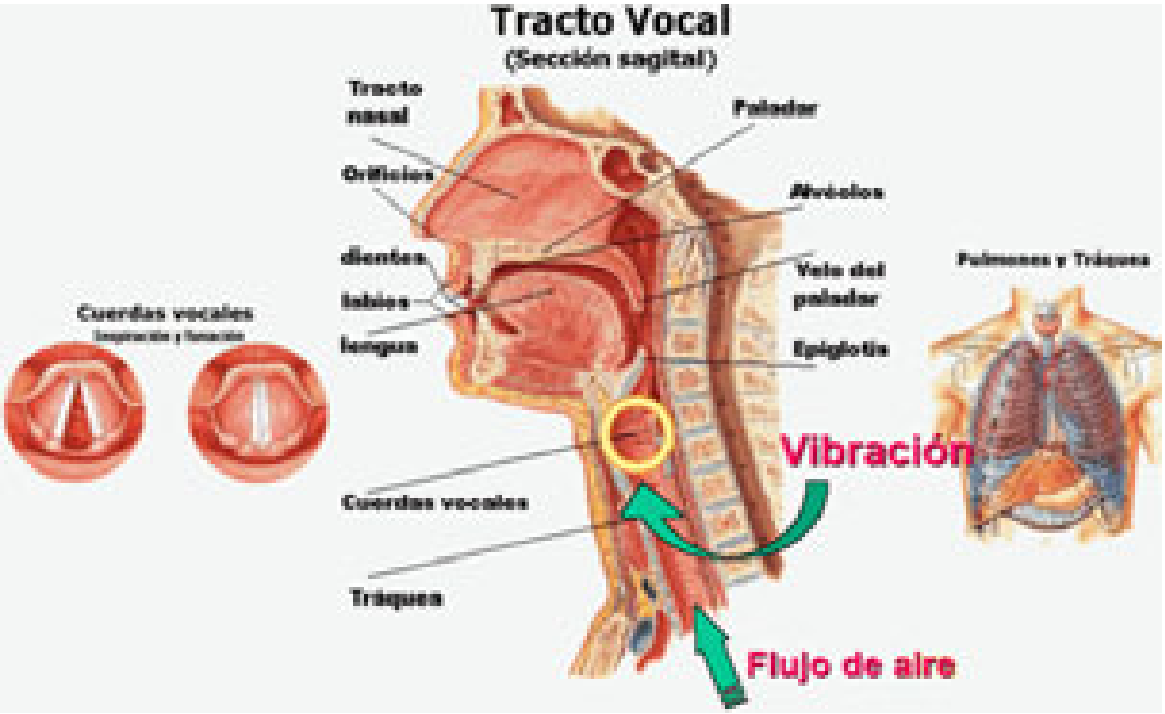


Figura 2.3. El tracto vocal y los órganos que intervienen en la generación de los sonidos.

El tracto vocal está formado por tres cavidades acústicas:

- La **cavidad faríngea**, situada inmediatamente después de la laringe.
- La **cavidad oral**, formada por el paladar, la lengua, los dientes y los labios.
- La **cavidad nasal**, que se encuentre entre el velo del paladar y los orificios nasales.

En la producción de sonidos hablados la laringe excita estas cavidades produciendo distintas frecuencias de resonancia denominadas *formantes*, que desempeñan un papel fundamental en la diferenciación de los sonidos, la forma con la que se pronuncian y por lo tanto la identificación del individuo. La detección de estas frecuencias de resonancia se realiza en la envolvente espectral de la señal, siendo estos formantes los máximos relativos de dicha envolvente (**Figura** 2.4).

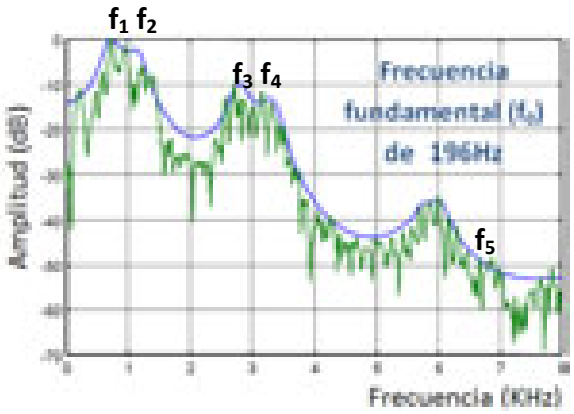


Figura 2.4. Espectro de la señal de voz. Formantes y estructura fina.

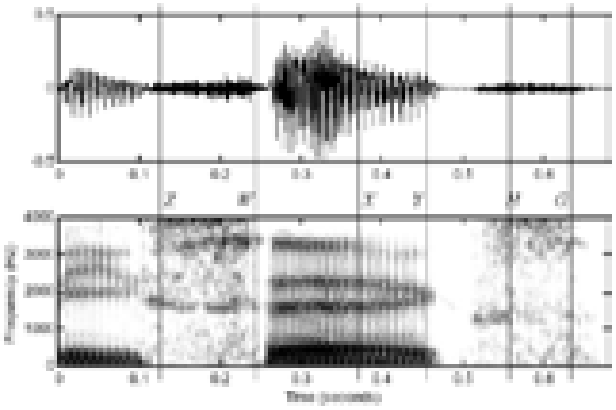


Figura 2.5. Espectrograma de un fragmento de voz.

La **Figura** 2.5 muestra el espectrograma (representación de la evolución de los formantes en función del tiempo) de otro fragmento de voz en el que se aprecian los formantes y su energía en un color más oscuro.

Para su reconocimiento biométrico se analiza a corto plazo (milisegundos), donde la señal presenta un carácter pseudoestacionario y se distingue entre dos tipos de sonidos:

1. **Sonidos sonoros**: que presentan un carácter pseudoperiódico. Son producidos por la vibración de las curvas vocales en tensión (**Figura** 2.6).

Figura 2.6. Forma de onda de una vocal de duración 80ms.

2. **Sonidos sordos**: que son de apariencia ruidosa y se producen por el paso libre del flujo de aire a través de las cuerdas vocales en estado de relajación. Se forman con la generación de diferentes sonidos en función de la articulación empleada, como por ejemplo los sonidos fricativos (/f/, /s/, /z/).

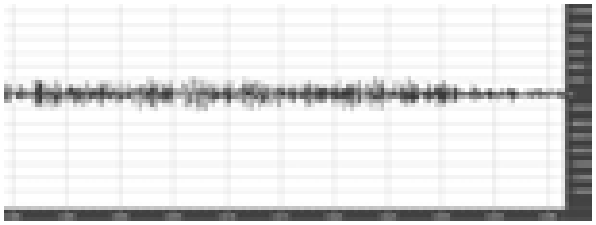


Figura 2.7. Forma de onda de un sonido sordo.

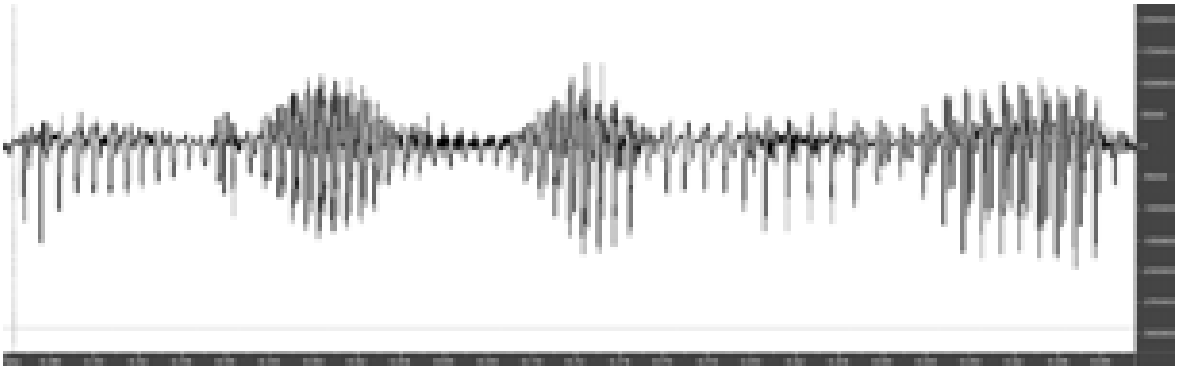


Figura 2.8. Señal de voz compuesta por sonidos sonoros y sordos.

2.3.2 LIMITACIONES DE LOS RASGOS BIOMÉTRICOS

Aunque la autenticación biométrica presenta claras ventajas frente a la autenticación por posesión o conocimiento como ya se explicó al inicio de este capítulo, es necesario considerar las limitaciones de estos sistemas que se agravan cuando se trabaja con millones de usuarios. Algunas de estas limitaciones se presentan a continuación:

RUIDO Y DISTORSIÓN

El ruido es una variable que introduce variabilidad en la muestra, haciendo que el rendimiento disminuya con la calidad de ésta. Tratándose de la voz, se puede entender como el entorno que envuelve la información que se desea transmitir: voz de otros usuarios en la grabación, el ruido de la red telefónica o el error cometido al digitalizar dicha señal mediante un conversor analógico digital (distorsión).

Si bien es cierto que existen técnicas como *factor analysis* que reducen de forma significativa la variabilidad introducida por el canal (9), estas técnicas dependen en gran medida de la disponibilidad de un corpus apropiado, deseablemente con las mismas condiciones de la voz a reconocer. Pero este hecho no es frecuente en aplicaciones reales, por lo que motiva la definición de nuevos algoritmos para compensar esta variabilidad como se presentan en este trabajo (ver capítulo 8). Para compensar la influencia del ruido y otros efectos de perturbadores también existen otras técnicas como la normalización por media *cesptral* (CMN), el filtrado *RASTA* o *Feature Warping* y *Feature Mapping* que serán explicadas en el apartado 4.5.

Otros estudios enfocan la eliminación del ruido realizando la combinación de sistemas de reconocimiento basados en las características espectrales de la voz, sujetas a variabilidad, con información de alto nivel como características dialectales o secuencias típicas de palabras que son robustas al ruido con resultados prometedores pero lejos de una solución definitiva.

VARIABILIDAD INTER-SESIÓN

La variabilidad inter-sesión es típicamente causada por una interacción incorrecta con el sensor. Algunos ejemplos comunes en voz son: la grabación de ésta mediante un micrófono a diferentes distancias, hablar mediante un teléfono distinto con el que se han grabado las conversaciones anteriores, el cambio de voz con la edad, la duración de las muestras capturadas o el ruido y distorsión introducida en la señal de voz al capturarla y almacenarla. Típicamente, para solucionar este problema la mejor aproximación es disponer de numerosas muestras que contemplen toda la variabilidad posible y clasificarlas en función del archivo a enfrentar. Una vez realizada dicha clasificación se puede recurrir a técnicas de normalización que eliminen esta variabilidad, como es el objetivo de este trabajo y cuyos métodos propuestos se describirán en el apartado 8.

VARIABILIDAD INTRA-SESIÓN

Es la variabilidad producida en las muestras por el mismo usuario durante la sesión de captura. Típicamente se puede ejemplificar como que la calidad y el espacio de características de la voz se ven afectados por un cambio en el estado emocional del locutor.

NO UNIVERSALIDAD

El sistema puede no estar capacitado para captar el rasgo biométrico de un conjunto de individuos, como por ejemplo ocurre en voz con las personas que sufren de algún tipo de discapacidad para producir discurso (mudez).

INTEROPERABILIDAD

El rendimiento de un sistema biométrico viene marcado por el tipo de sensor con el que se ha capturado la muestra, ya que cada sensor presenta una sensibilidad distinta y una resolución y error diferente. Este problema es estudiado en este proyecto, cuyos experimentos relacionados con la calidad de la voz integran ficheros de audio de fuentes telefónicas y microfónicas.

ATAQUES AL SISTEMA

Existen multitud de ataques diferentes que pueden burlar a un sistema de reconocimiento biométrico (14). Algunos de ellos se presentan a continuación:

- **Ataques falsos (*Spoof Atacks*)**: implican la manipulación deliberada de algún rasgo biométrico para evitar el reconocimiento o para burlar el sistema. La voz, al ser un rasgo conductual está fuertemente vinculado a este tipo de evasiones o ataques. También incluyen la creación de artefactos para imitar la identidad de otra persona, combatidos con la detección de vida o combinación con el reconocimiento de otros rasgos biométricos.

- **Ataques *Hill-Climbing***: es un tipo de ataque que, cuando se tiene acceso al sistema de reconocimiento, consiste en generar parámetros de acuerdo con el resultado extraído, y en función de este resultado optimizar los parámetros de forma recursiva para burlar al sistema (14).

3 SISTEMAS DE RECONOCIMIENTO BIOMÉTRICO

Un sistema biométrico es esencialmente un sistema de reconocimiento de patrones. Puede definirse de forma aproximada como cuatro bloques o módulos caracterizados de la siguiente forma:

1. **Módulo de captura**: que adquiere muestras de un individuo a través de un sensor.
2. **Módulo de procesado**: que extrae las características relevantes del rasgo biométrico. Algunos sistemas incorporan un control de calidad que obliga a recapturar la muestra biométrica si dicha calidad es insuficiente para la posterior identificación.
3. **Módulo de cálculo de similitud (*matching*)**: que compara las características extraídas con el conjunto de características de otro usuario y calcula una medida de dicha similitud (puntuación o *score*).
4. **Base de datos**: contiene el conjunto de características de todos los usuarios registrados en el sistema. Este registro puede ser supervisado o no por una persona o por un sistema de detección de calidad.

3.1 MODOS DE FUNCIONAMIENTO DE UN SISTEMA BIOMÉTRICO

A continuación se describen los 3 modos básicos de operación de un sistema de reconocimiento biométrico (25).

MODO REGISTRO

Previamente a la autenticación de una persona, el usuario que se desea identificar debe ser inscrito mediante una ficha personal que contenga su identificador (ID) y el rasgo biométrico a evaluar. Para ello se captura una muestra de este rasgo cuya calidad será comprobada y en el caso de ser apta se le extraerá su patrón característico para su posterior incorporación a la base de datos del sistema.

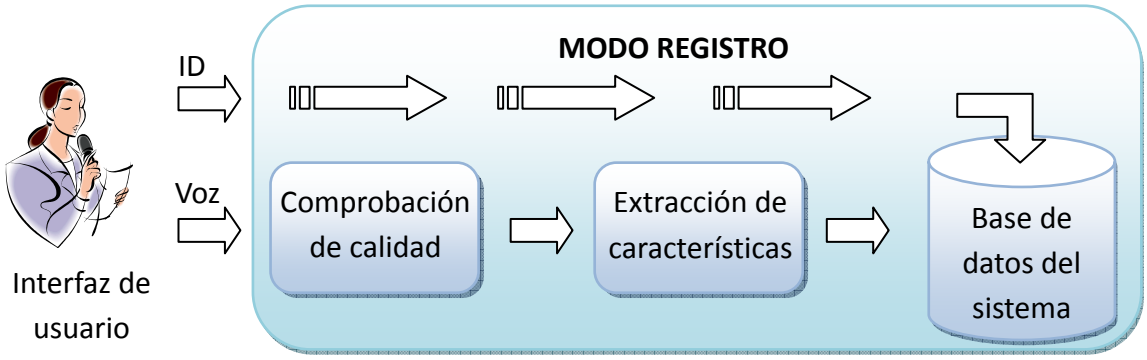


Figura 3.1. Modo registro. Figura adaptada de (25).

Dependiendo del contexto de la aplicación, un sistema biométrico puede operar de dos formas diferentes: modo verificación y modo identificación.

MODO VERIFICACIÓN

En el modo verificación el usuario introduce su ID en el sistema y éste compara los parámetros extraídos de la muestra capturada con la ya existente en la base de datos, es decir, se realiza una comparación uno a uno. Finalmente el sistema de reconocimiento acepta al usuario si la puntuación obtenida es mayor que un umbral o lo rechaza en caso contrario. Sobre este tipo de sistemas será en los que se base este proyecto, por lo que serán estudiados con posterioridad.

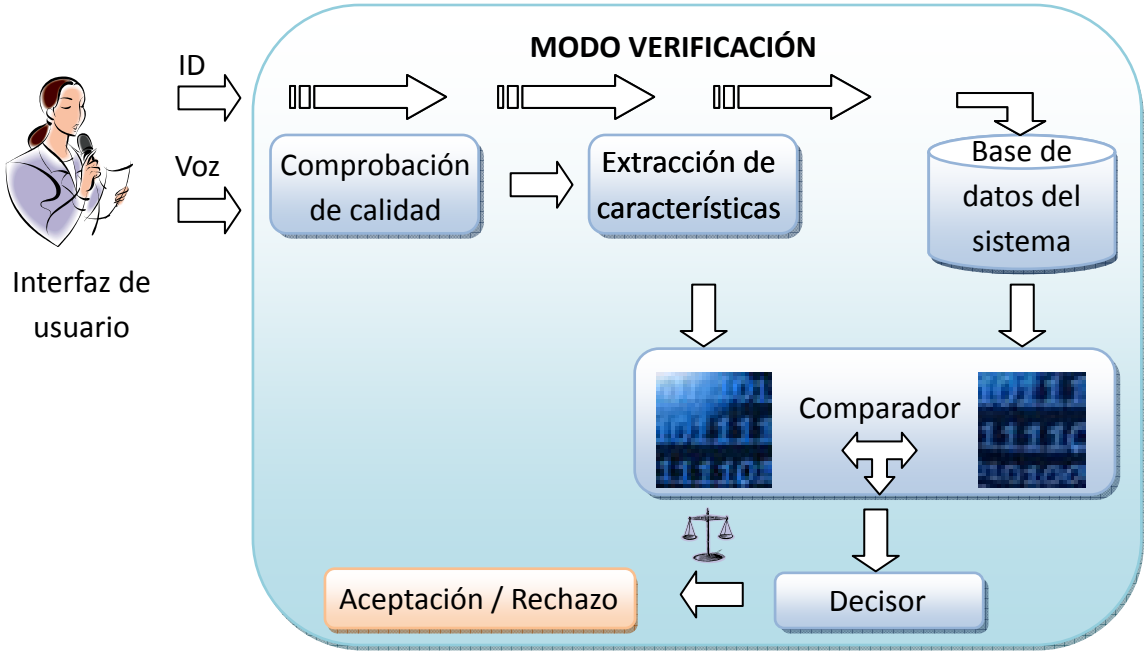


Figura 3.2. Modo verificación. Figura adaptada de (25).

MODO IDENTIFICACIÓN

En el modo identificación se desconoce la identidad del usuario, por lo que se realizan tantas comparaciones como usuarios existen en la base de datos y se devuelve una lista ordenada de candidatos de mayor a menor probabilidad de ser el usuario a identificar sobre el conjunto cerrado.

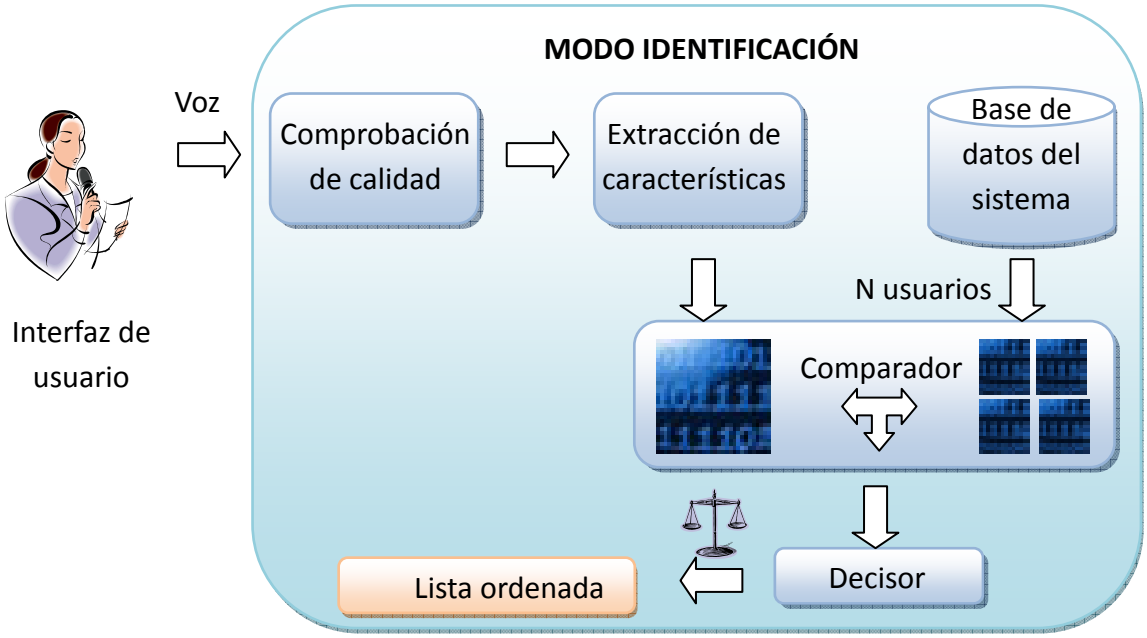


Figura 3.3. Modo identificación. Figura adaptada de (25).

El principal problema del modo identificación se encuentra cuando el reconocimiento es negativo, es decir, cuando el usuario a identificar no se encuentra en la base de datos obteniéndose como resultado, de igual forma, una lista ordenada. Debido a este problema es común utilizar un umbral para dotar al sistema de robustez cuando se trabaja con conjuntos abiertos. Por el contrario y como ventaja, este tipo de sistemas no necesitan de un identificador adicional al rasgo biométrico para permitir el acceso de un usuario al sistema.

3.2 ERRORES EN LA VERIFICACIÓN Y MEDIDA DEL RENDIMIENTO

Dado que el modo de funcionamiento para el que se desarrolla este proyecto es el de verificación, conviene estudiar los errores derivados de éste para poder establecer una magnitud que defina el rendimiento del sistema de forma representativa. Esta magnitud es la EER (*Equal Error Rate*), que se define como el punto en el que la probabilidad de falso rechazo (PFR) y la probabilidad de falsa aceptación (PFA) son iguales, entendiendo ambas como la probabilidad de que el sistema rechace a un usuario genuino o acepte a un impostor de forma respectiva.

La **Figura** 3.4 muestra las funciones densidad de probabilidad (*fdp*) dada una puntuación O obtenida tras la comparación. Estas distribuciones corresponden a la hipótesis de que el usuario y el modelo almacenado con el que se compara son la misma persona (distribución *target*) y a la hipótesis en el que ambos no coinciden (distribución *non target*). Por lo tanto, si se compara el modelo del supuesto usuario frente a las características de test a identificar y la comparación da como resultado un valor mayor que un cierto umbral el usuario será aceptado. En el caso contrario éste será rechazado. En la misma figura también se representan la PFA y la PFR como el área azul bajo la curva *non target* a partir del umbral y el área de color rojo, bajo la curva *target*, anterior a este mismo umbral, respectivamente. La misma situación puede definirse mediante la curva representada en la **Figura** 3.5, donde se muestra la evolución de estas probabilidades en función del umbral establecido.

Respecto al umbral establecido, puede interesar definirlo con un valor mayor o menor en función de la aplicación:

- **Interesará un umbral alto**, que implica una PFA baja, en operaciones de máxima seguridad. Un ejemplo típico podría ser la autenticación para realizar una transferencia bancaria, escenario en el que es preferible que el usuario tenga que ingresar sus datos más veces contar de dotar al sistema de una cierta robustez.

- **Interesará un umbral bajo**, que implica una PFR baja, en aplicaciones donde la privacidad no sea un requisito indispensable. Un ejemplo de este caso podría ser el control de acceso mediante reconocimiento biométrico a un parque de atracciones, en el que interesa más que haya poca cola aun equivocándose más veces el sistema.

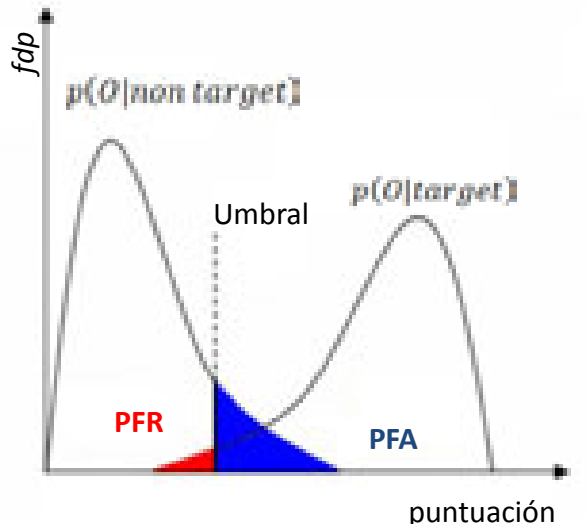


Figura 3.4. Densidad de probabilidad de usuarios e impostores.

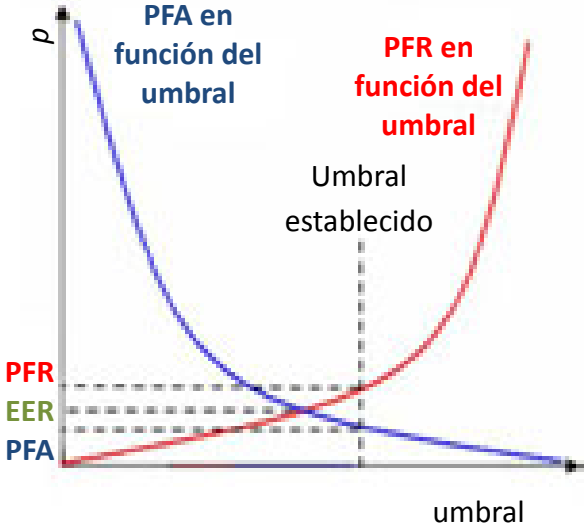


Figura 3.5. Probabilidad de falsa aceptación y falso rechazo en función del umbral.

La relación entre la PFA y la PFR se representa típicamente mediante las curvas DET (*Detection Error Trade-off*) (**Figura** 3.6), donde interesará que la curva representada esté más cerca del origen ya que significará que la discriminación entre las distribuciones *target* y *non target* es mayor, siendo así el sistema más robusto frente a los errores de clasificación.

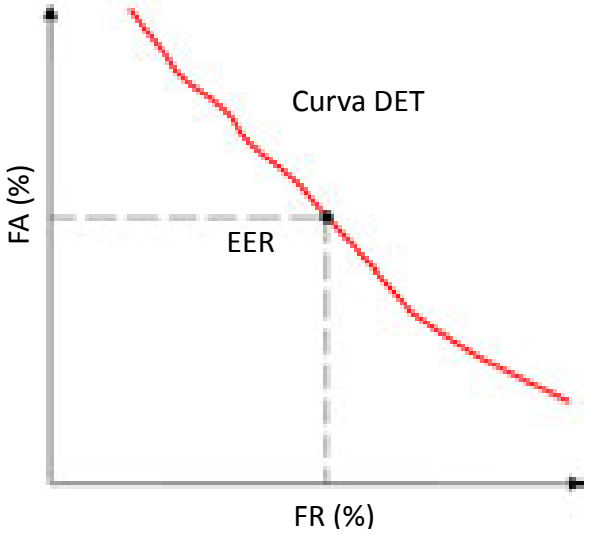


Figura 3.6. Curva DET.

Aunque el uso del EER como medida de rendimiento está ampliamente extendido (26), si el umbral establecido es distinto de este punto la medida puede no ser muy representativa, por lo que existen otras medidas alternativas como el $MinC_{llr}$, definido entre 0 y ,1 que tienen en cuenta lo bueno que es un sistema discriminando para cualquier umbral (representan de forma cuantitativa la curva DET completa y no sólo un punto como el EER). Esta medida de rendimiento se define cualitativamente de la siguiente forma: cuanto mayor $MinC_{llr}$ peor discriminación (correspondiendo el valor 1 a las distribuciones *target* y *non target* totalmente solapadas) y cuanto menor valor mejor discriminación (correspondiéndose el valor 0 a una discriminación total).

3.3 FUSIÓN DE SISTEMAS

Las personas se reconocen entre sí a través de distintas características biométricas ya sean físicas o conductuales. Por lo tanto es apropiado pensar que combinando un conjunto de sistemas biométricos complementarios el rendimiento del sistema será mayor (27). Esto se deberá a que se obtiene más información característica del sujeto o con menor distorsión que, combinándola de manera apropiada, mejorará el rendimiento del sistema global. A este concepto se le denomina *fusión de sistemas* y existen varias formas de implementarlo (28) (ver **Figura** 3.7):

- **a nivel de sensores**: utilizando diferentes instrumentos de medida que se distorsionan de forma diferente. En voz, pueden emplearse técnicas para fusionar datos de procedencia microfónica con telefónica.

- **a nivel de muestras**: utilizando diferentes muestras adquiridas bajo el mismo sensor en iguales o diferentes condiciones de entorno.

- **a nivel de instancias**: utilizando el mismo rasgo biométrico pero diferente instancia. En voz es similar a realizar reconocimiento de voz dependiente e independiente de texto (ver apartado 4.3), por ejemplo.

- **a nivel de sistemas para un mismo rasgo**: utilizando fusión multinivel o fusión de algoritmos. Es una técnica muy empleada ya que el rendimiento mejora de forma significativa si los sistemas son complementarios. En voz, una técnica típica es fusionar un sistema GMM (*Gaussian Mixture Model*) con un sistema SVM (*Support Vector Machine*) (ver capítulo 4.4).

- **a nivel de sistemas de reconocimiento rasgos biométricos distintos**: combinando, por ejemplo, el reconocimiento de voz con el de rostro para diseñar un sistema global más robusto. Estos sistemas solucionan en parte el problema de la no universalidad y el de la escasa disponibilidad de muestras biométricas de una determinada calidad. En cuanto al fraude, cabe destacar que hacen más complicado romper un sistema ya que requiere falsificar diferentes rasgos.

La dificultad de estos métodos radica en la elección de la estrategia de fusión a seguir y la definición de la adecuada combinación de los sistemas para obtener así un rendimiento superior. Además, habrá que tener en cuenta el coste de cada sistema y llegar a un compromiso entre el rendimiento del sistema global y su complejidad.

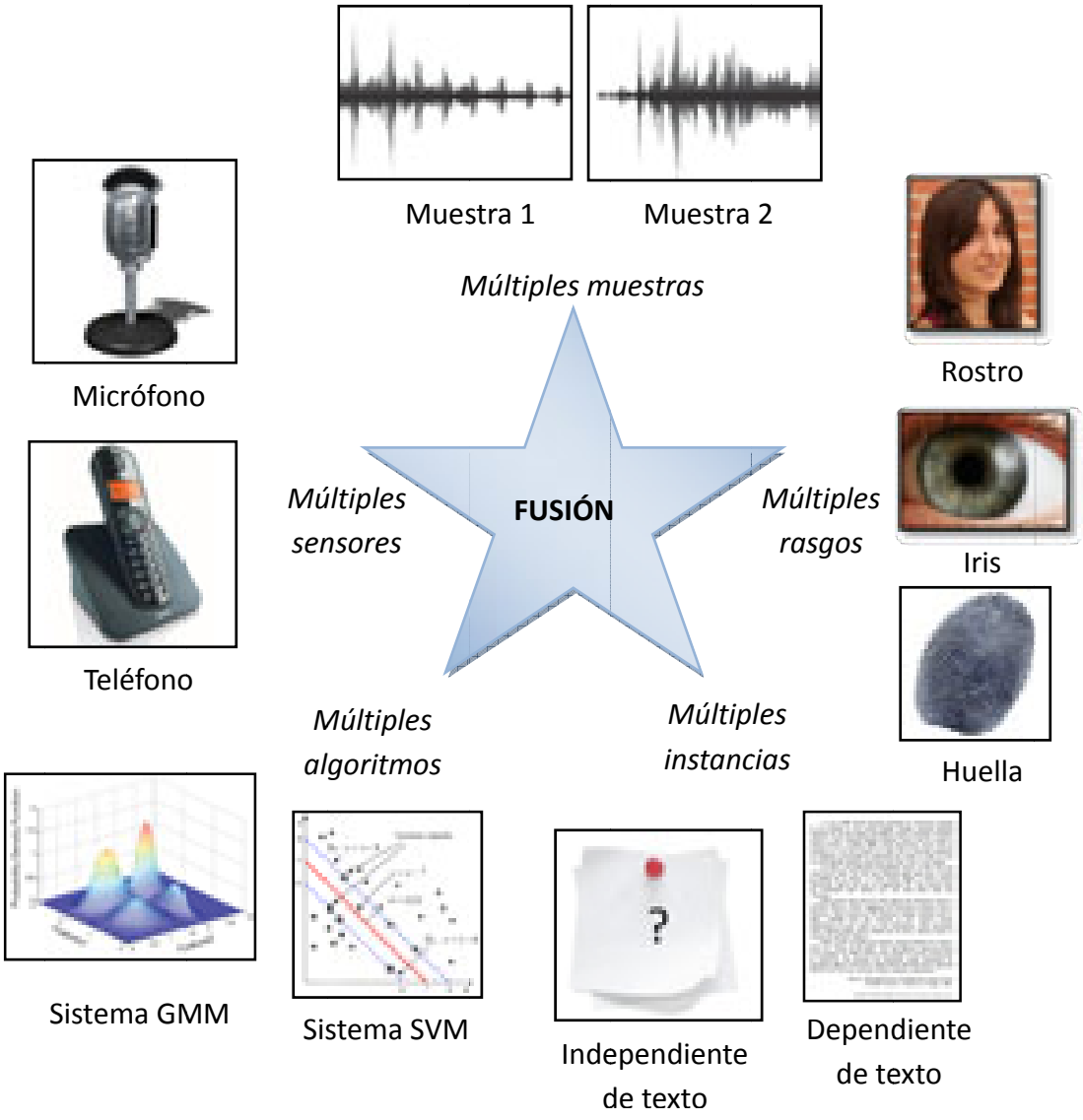


Figura 3.7. Escenarios de información para realizar fusión de sistemas. Figura adaptada de (28).

4 SISTEMAS DE RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR

4.1 DESCRIPCIÓN GENERAL

Un sistema de reconocimiento automático de locutor es un conjunto de algoritmos que tienen por objeto obtener información sobre la identidad de un usuario a partir de muestras de su voz.

En función del tipo de información que usan se puede distinguir entre dos grandes tipos de reconocedores (29):

- **Reconocedores de alto nivel,** que se centran en la información hablada como el tipo de lenguaje empleado, las pausas, la melodía o el número de veces que se repiten ciertas palabras o ciertos fonemas. Este tipo de sistemas se presuponen más robustos frente a ruido y distorsión, perturbaciones que sí afectan a los de bajo nivel. Las líneas de investigación actuales apuntan a algoritmos complejos que ofrecen un peor rendimiento que los de bajo nivel cuando la duración de la muestra es inferior a 10 minutos (30). Aún así, estas investigaciones son de suma importancia pues estos sistemas ofrecen gran complementariedad al reconocimiento de bajo nivel. Este grupo, a su vez, puede dividirse en varios niveles diferentes entre los que destacan el prosódico, el lingüístico y el fonético.

- **Reconocedores de bajo nivel**, que caracterizan al hablante desde la fase de producción de voz, etapa en la que intervienen los órganos ya estudiados en el apartado 2.3.1. Caracterizan la información asociada con el espectro de la señal, como el tono, las frecuencias de resonancia o formantes, la coarticulación y la concatenación de los sonidos. Éste nivel también se denomina acústico-espectral.

Para construir un buen sistema de reconocimiento automático de locutor, en adelante SRAL, es importante saber combinar de la manera adecuada los diferentes reconocedores. Por lo tanto, una división más precisa de los niveles citados se define de la siguiente manera:

- **Nivel lingüístico**: hace referencia a las características relacionadas con el uso del lenguaje y por lo tanto depende de aspectos como la educación, el origen y las condiciones sociológicas del hablante (30).

- **Nivel prosódico**: es el principal responsable de dotar a la voz de naturalidad y sentido a través de la combinación de la energía, la duración y el tono de los

fonemas.

- **Nivel fonético**: formado por la coarticulación de los fonemas y sus características.
- **Nivel acústico-espectral**: representa las características espectrales a corto plazo de la señal de voz, es decir, define las características de los órganos que intervienen en la generación del habla.

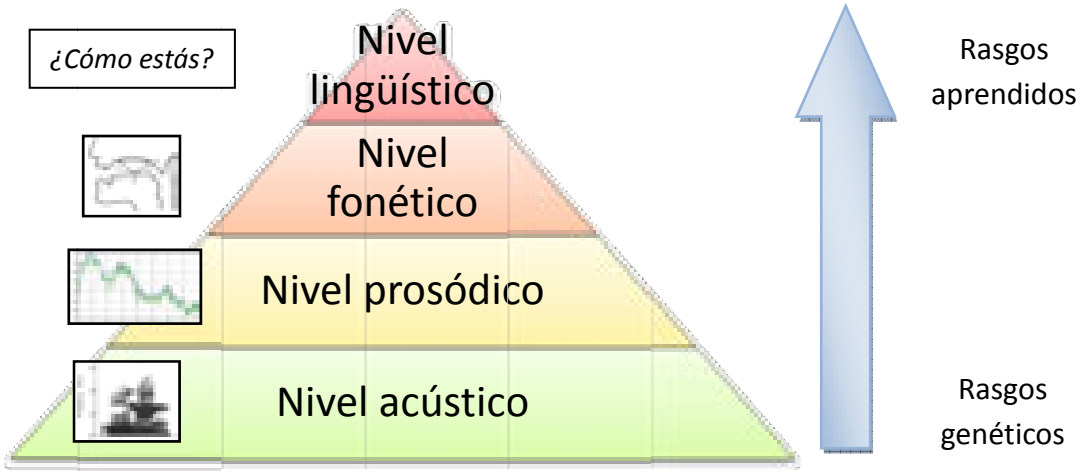


Figura 4.1. Niveles de identidad. Figura adaptada de (31).

4.2 CARACTERÍSTICAS IDENTIFICATIVAS EN LA SEÑAL DE VOZ

Diversas investigaciones llevadas a cabo hasta la fecha, han demostrado que los coeficientes en el dominio espectral representan mejor las características del nivel acústico de la voz, ya que se asemejan más a la percepción natural del oído humano (25). Por lo tanto, si las personas son capaces de reconocerse por medio de estas características posiblemente los sistemas automáticos presenten un buen rendimiento en este sentido.

Los sistemas de reconocimiento de locutores a nivel espectral, basan su eficiencia en la capacidad de parametrizar de forma adecuada la envolvente espectral de la señal de voz. A continuación se presentan dos técnicas ampliamente extendidas para la caracterización del locutor: MFCC (*Mel-Frecuency Cepstral Coefficients*) y LPC (*Linear Predictive Coefficients*).

COEFICIENTES MFCC (*MEL-FRECUENCY CEPSTRAL COEFFICIENTS*)

Las características representativas de la voz más comunes en el dominio espectral son los coeficientes MFCC (*Mel-Frecuency Cepstral*[2] *Coefficients*) (25) y derivados de ellos los coeficientes delta (Δ) y los coeficientes delta-delta (Δ-Δ), que dependen de la velocidad y aceleración con la que varían los mismos.

La extracción de estos coeficientes se realiza de la siguiente manera:

1. Se digitaliza la señal de voz (se muestrea y se cuantifica).

2. Se realiza un análisis localizado mediante la aplicación de ventanas, típicamente ventanas *Hamming* o rectangular de 20 ms de duración, que presentan en frecuencia un compromiso entre el tamaño del lóbulo principal, que debe ser estrecho, y unos lóbulos secundarios bajos. A continuación se presente la señal enventanada definida como:

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

donde $x(m)$ es el resultado de multiplicar la señal original de voz $s(n)$ por la ventana temporal $w(n)$ siendo *N* la duración de la ventana aplicada.

Con este tipo de ventanas se pretende reducir la distorsión en la mayor medida posible en señal final, ya que la multiplicación temporal se convierte en una convolución en frecuencia. La ventana *Hamming*, utilizada para la extracción de características del sistema utilizado en este proyecto tiene una estructura temporal en forma de coseno alzado (ver **Figura** 4.2):

$$w(n) = \begin{cases} 0.54 - 0.46 \cdot \cos\left({}^{2\pi n}/_{N-1}\right), & 0 \le n \le N-1 \\ 0, & en\ el\ resto \end{cases}$$

Dado que utilizar este tipo de ventanas supone minimizar las muestras de los extremos se aplican ventanas de forma solapada, siendo el área de solapamiento típica del 50% del total (10 ms), como se muestra en la **Figura** 4.3.

Una vez enventanada la señal, se suprimen los silencios mediante un detector de actividad de voz (VAD), desechando aquellas ventanas cuya energía media no supere un cierto umbral.

[2] El dominio cepstral se define como la transformada inversa de Fourier del logaritmo del módulo espectral, entendiéndose dicho módulo como la multiplicación de la respuesta del trato vocal (envolvente espectral) por la excitación glotal (estructura fina) (60).

3. Una vez obtenida la señal enventanada se realiza el módulo de su transformada de Fourier.

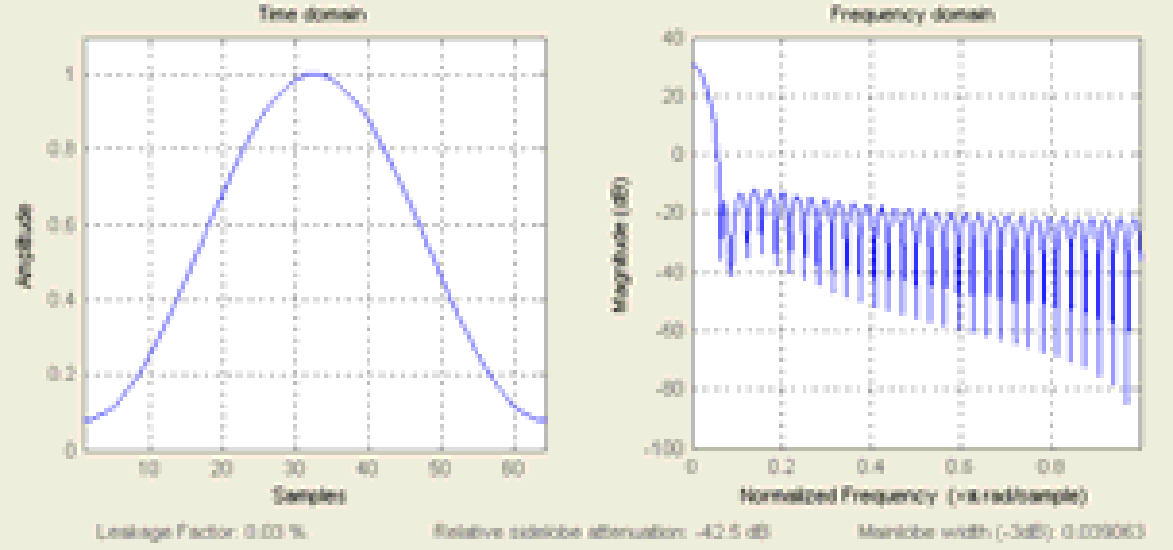


Figura 4.2. Ejemplo de ventana *Hamming* en el dominio del tiempo y la frecuencia.

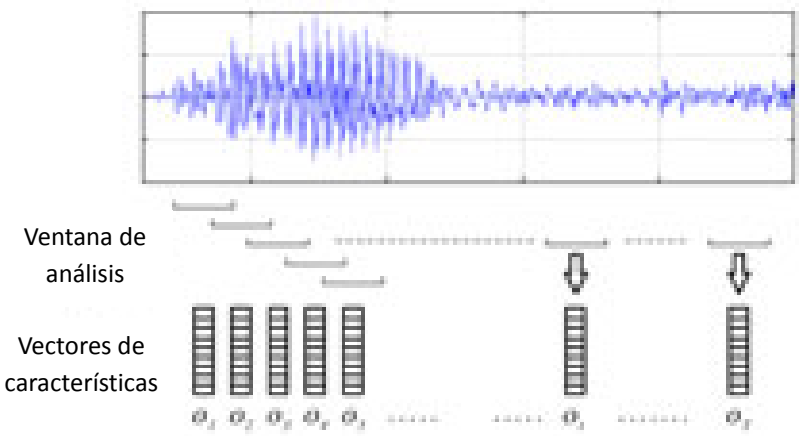


Figura 4.3. Enventanado de la señal de voz para la posterior extracción de características. Figura adaptada de (26).

4. El siguiente paso es filtrar la señal mediante un banco de filtros *Mel*, de tal forma que se otorgue una mayor resolución a las frecuencias más bajas como ocurre en el oído humano. El número de filtros utilizados es variable, y se considera un parámetro de diseño.

5. Se aplica el logaritmo a la señal para obtener la energía promedio de cada filtro.

6. Se realiza el suavizado de la envolvente mediante la DCT (*Discrete Cosin Transform)* de tal manera que se obtienen los coeficientes MFCC ortogonales entre sí.

Típicamente el número de coeficientes o parámetros característicos que se extraen son entre 13 y 19. Para el sistema utilizado para el desarrollo de este proyecto se eligieron 19 (apartado 5.2) ya que este número ha garantizado un buen rendimiento del sistema en evaluaciones anteriores.

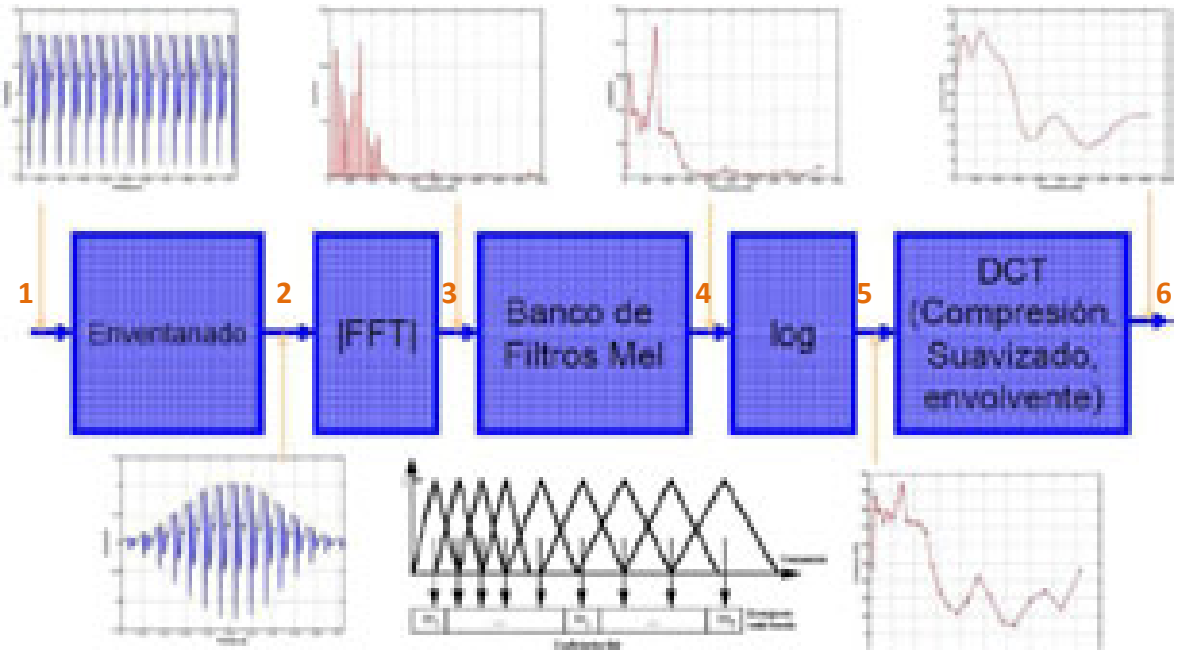


Figura 4.4. Extracción de los coeficientes MFCC.

COEFICIENTES LPC (*LINEAR PREDICTION COEFFICIENTS*)

Aunque los coeficientes más representativos son los MFCC existen otros como los LPC (*Linear Prediction Coefficients*), cuyo uso también se encuentra muy extendido en el ámbito del reconocimiento de voz. Estos coeficientes se basan en la fuerte correlación de muestras consecutiva en el habla, por lo que una muestra se puede definir como combinación lineal de las anteriores. De esta manera, y asemejándolo al modelo de producción de voz, formado por diferentes modelos matemáticos que definen de forma determinista el tracto vocal, se puede establecer una relación para definir un sistema lineal que responda a la envolvente espectral de la voz del usuario. Por lo tanto, y a través del análisis de predicción lineal (que queda fuera de este proyecto) se pueden definir los coeficientes LPC como los polos de dicho filtro, por lo que serán representativos de la identidad del usuario y de gran utilidad en el campo del reconocimiento de voz.

4.3 TIPOS DE RECONOCEDORES DE LOCUTOR

Las tecnologías de reconocimiento de voz pueden ser usadas en multitud de aplicaciones como el reconocimiento de habla espontánea, que puede ser usado como un sistema de control, un editor de texto o para realizar búsquedas de contenidos en grabaciones. Otra aplicación muy común es el reconocimiento de idioma, que por ejemplo puede ser incorporado a un número de emergencia para transferir al usuario que llama directamente a una persona que hable su lengua. Otras aplicaciones derivadas del reconocimiento de voz son la identificación de individuos, ya sea en ámbitos forenses o para garantizar la seguridad en un control de acceso. Todas estas tecnologías pueden llevarse a la práctica mediante dos tipos de reconocedores que a continuación se detallan, los dependientes y los independientes de texto.

RECONOCEDORES DEPENDIENTES DE TEXTO

La característica principal de estos sistemas es que el texto de la locución a reconocer es conocido. Al reducirse de manera drástica las posibilidades de habla, el sistema es menos complejo, pero, por el contrario, requieren de la colaboración del usuario. En la actualidad estos sistemas hacen uso de una topología de HMM de izquierda a derechas, la cual será descrito en el apartado 4.4.

Principalmente este tipo de sistemas se utilizan para el control de acceso, combinando autenticación biométrica con autenticación por conocimiento, ya que el texto a repetir es típicamente la contraseña de usuario. No obstante también se utilizan en aplicaciones de domótica o control de aparatos electrónicos.

RECONOCEDORES INDEPENDIENTES DE TEXTO

Este tipo de sistemas ha experimentado un gran desarrollo durante las dos últimas décadas ya que presentan un desafío en cuanto a conocimiento científico se refiere al ser desconocido el contenido lingüístico de la locución a reconocer. Son más comunes en reconocimiento automático de locutor ya que cualquier indicio de voz sirve para identificar a un usuario al no estar sujeto a un texto específico. Están basados, principalmente, en el nivel acústico (características espectrales de la voz). Los sistemas típicos que se utilizan en la actualidad incluyen los modelos de mezclas gaussianas (GMM), las máquinas de vectores soporte (SVM) y sistemas híbridos GMM-SVM (*SuperVectors*).

En la actualidad, la aplicación de este tipo de sistemas está muy extendida. Se utilizan para la identificación de personas en ámbitos forenses y al igual que los dependientes de texto para el control de acceso o el control remoto de aparatos electrónicos.

4.4 ESTADO DEL ARTE DE LOS SISTEMAS ACTUALES

El objetivo de un SRAL es autenticar a un usuario mediante una muestra de su voz. Para ello será necesario crear un modelo del locutor a través de las características de su locución, que será el que se compare o enfrente con la información extraída de los archivos de test cuya identidad se desea conocer. Como resultado se obtendrá una puntuación (*score*) que se comparará a un umbral establecido y en función de su valor se realizarán las acciones apropiadas dependiendo del modo de funcionamiento.

Algunos de los esquemas de modelado típicos en un SRAL que se utilizan en la actualidad se presentan a continuación, siendo posible la fusión de estos para obtener un rendimiento global mayor.

HMM (*HIDDEN MARKOV MODELS*)

Un HMM es un modelo estadístico en el que se asume que el sistema a modelar es una cadena de eventos (modelo de Markov de parámetros desconocidos) donde cada estado depende de las salidas probabilísticas de los eventos anteriores. Por lo tanto, un HMM queda definido por el número de estados (de carácter invariante) que lo componen, por las probabilidades de transición de un estado a otro, por sus probabilidades de estado inicial y por la probabilidad de observación en cada estado (36). La **Figura** 4.5 representa un modelo oculto de Markov de izquierda a derecha, en el que los términos a_{ij} simbolizan las probabilidades de transición entre estados y los $b_j(O_m)$ la probabilidad de observar la secuencia O_m en ese estado. En este caso, las secuencias a modelar son la sucesión de vectores de características (MFCC, LPC) extraídos de la señal de voz.

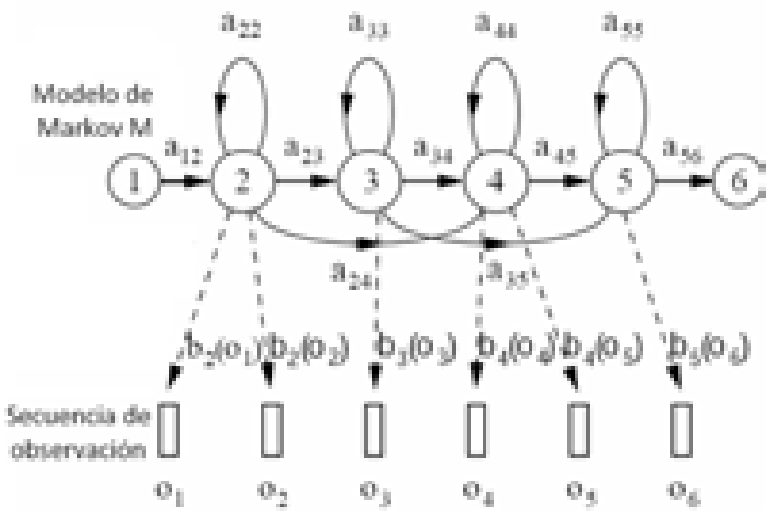


Figura 4.5. Modelos Ocultos de Markov.

No obstante, habiendo definido un HMM se presentan 3 cuestiones básicas a resolver a la hora de implementar un sistema de reconocimiento de locutor (36):

1. **Problema de puntuación**: ¿Cómo se calculan las probabilidades de observar una secuencia que defina las características de un locutor dado un modelo oculto de Markov?

   La solución directa a este problema sería calcular todos los caminos probabilísticos posibles, que se convierte en una solución impracticable cuanto mayor es el número de estados. Por lo tanto, se utilizan otros algoritmos como los *Forward* o *Backward* que consiguen llegar a una solución computacionalmente aceptable.

2. **Problema de reconocimiento de estados**: ¿Cómo encontrar la secuencia de estados que mejor se ajusta a las observaciones?

   Puede ocurrir que la secuencia de estados no tenga sentido (si por ejemplo la coarticulación de esos fonemas no exista en ese idioma). Por lo tanto habrá que elegir el camino de estados con mayor probabilidad total. Comúnmente este problema se resuelve mediante el algoritmo de Viterbi, que obtiene el resultado de mayor probabilidad de la forma menos costosa.

3. **Problema de entrenamiento**: ¿cómo encontrar las matrices de probabilidad para maximizar la probabilidad de observar el conjunto de entrenamiento?

   La solución a este problema pasa por aplicar el algoritmo *Baum-Welch* o criterio de máxima verosimilitud, en el que se maximiza la probabilidad de la observación dado el modelo. Para ello se inicializa el modelo de forma aleatoria y se itera sobre los datos disponibles hasta cumplir con el criterio de máxima verosimilitud fijado.

Un sistema basado en HMM modela de forma estadística la acústica de la voz teniendo presente la direccionalidad de los fonemas, por lo que la definición de un modelo de izquierda a derechas los hace ideales para el reconocimiento de voz dependiente de texto. No obstante, definiendo de forma apropiada la topología de los HMM se pueden utilizar también para reconocimiento de locutor independiente de texto.

GMM (*GAUSSIAN MIXTURE MODELS*)

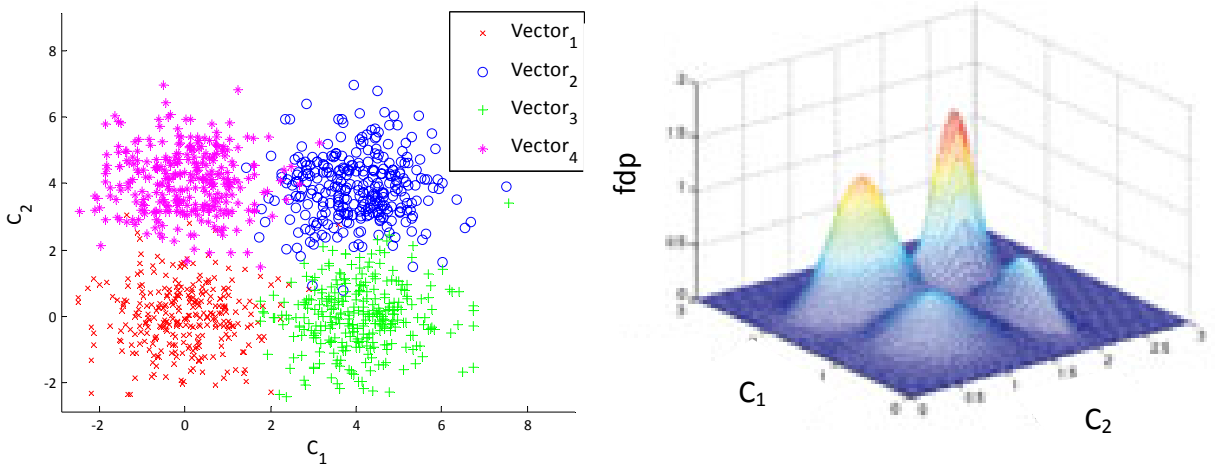
Cada estado del HMM puede considerarse, a su vez, como un modelo de mezclas gaussianas o GMM. Un GMM es un modelo estadístico que define las distribuciones de probabilidad de las características espectrales de la voz para discriminar a los posibles locutores. Está formado por G gaussianas multivariadas definidas por su peso ω_i, su vector de medias μ_i y su matriz de covarianzas Σ_i. El modelo se representa a continuación:

$$\theta = \{\omega_i, \mu_i, \Sigma_i\}$$

Ante una observación x desconocida, el modelo GMM asigna una puntuación relacionada con la verosimilitud entre el modelo y la muestra observada. La siguiente fórmula define la probabilidad de observar el parámetro x de dimensión *D,* dada la hipótesis *target,* como una suma ponderada de las funciones densidad de probabilidad de dichas gaussianas características del usuario (**Figura** 4.6):

$$p(x|target) = \sum_{i=1}^{G} w_i \cdot p_i(x)$$

$$p_i(x) = \frac{1}{(2\cdot\pi)^{D/2}\cdot|\Sigma|^{1/2}} \cdot e^{\left(-\frac{1}{2}\cdot(x-\mu_i)^T\cdot(\Sigma)^{-1}\cdot(x-\mu_i)\right)}$$



a) Distribución espacial de los coeficientes espectrales 1 y 2 para distintas muestras de voz del mismo locutor.

b) GMM entrenado a partir de la distribución espacial de los coeficientes.

Figura 4.6. Representación espacial de las características espectrales del locutor mediante GMMs.

Mediante el teorema de Bayes se puede demostrar que la decisión óptima viene dada por el cociente de probabilidad que sigue (denominado *relación de verosimilitud* o LR, en inglés *Likelihood Ratio*), siendo *u* el umbral establecido.

$$\frac{p(x|target)}{p(x|non\ target)} \begin{cases} \leq u & \rightarrow se\ apuesta\ a\ que\ x\ corresponde\ al\ locutor. \\ > u & \rightarrow se\ apuesta\ a\ que\ x\ no\ corresponde\ al\ locutor. \end{cases}$$

Para estimar $p(x|non\ target)$ se hace uso del modelo universal o *UBM* (*Universal Background Model*). El UBM es un modelo de 1024 gaussianas entrenado con toda la información de voz disponible, de tal manera que las gaussianas obtenidas representan a

todos los usuarios.

Como típicamente los archivos para generar el modelo del locutor son de una duración corta y por lo tanto contienen poca información para su identificación (si se generase un GMM éste sería poco general), lo que se hace es calcular las mezclas más representativas del GMM del usuario y adaptar las mezclas más pesadas del UBM, del tal forma que se obtenga un modelo semejante al UBM, pero que resalte las diferencias del locutor.

Aunque existen diversas técnicas para logar esta adaptación, para este proyecto se utiliza la **adaptación MAP** (*Maximum a Posteriori*) al disponerse de una gran cantidad de datos, que hace uso del algoritmo EM (*Expectation Maximization*) (35) para transformar las diferentes gaussianas de forma independiente variando su media y covarianza (**Figura** 4.7). Esta adaptación opera parámetro a parámetro de los GMMs y considera que el coeficiente del modelo independiente del locutor es la información a priori sobre dicho parámetro para después, con la nueva información observada del locutor, estimarlo. La nueva media para la gaussiana *m* en el estado *j* se calculará mediante la siguiente fórmula, donde τ es el peso de la información a priori, *N* es la probabilidad de ocupación de los datos de adaptación, μ_{jm} se corresponde con la media del modelo independiente de locutor y $\bar{\mu}_{jm}$ es la media de los datos de adaptación:

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\bar{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm}$$

De la fórmula se puede extraer que si la probabilidad de ocupación de una componente gaussiana N_{jm} es pequeña, entonces la estimación de la media será parecida a la media de la componente independiente de locutor.

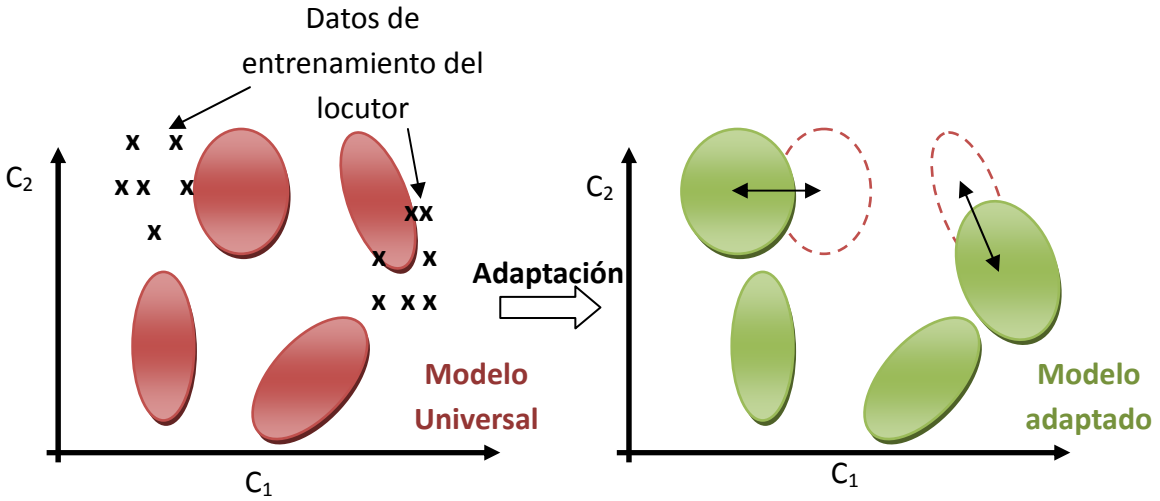


Figura 4.7. Representación gráfica de la adaptación del UBM al modelo del locutor.

Dado que los UBM estiman la densidad de probabilidad de las observaciones de todos los usuarios, la probabilidad de que el modelo universal (λ_{UBM}) observe la secuencia de

parámetros *O* de test, relacionada con la probabilidad de que el modelo de entrenamiento (λ_t) observe esos mismos parámetros se define como puntuación (*score*), y será una magnitud fiable de pertenencia de estos coeficientes característicos siempre que se disponga de las probabilidades condicionadas (el resultado de enfrentar ese fichero de test frente a modelos conocida su pertenencia o no pertenencia).

$$S(O,\lambda_t) = \log\big(p(O,\lambda_t)\big) - \log\big(p(O,\lambda_{UBM})\big)$$

La principal limitación de estos sistemas se encuentra en que sólo modela la característica de la voz, por lo que en ocasiones se combinan con reconocedores de alto nivel basados en HMMs para aportar fiabilidad al reconocimiento.

SISTEMAS BASADOS EN SVM (*SUPPORT VECTOR MACHINES*)

Las Máquinas de Vectores Soporte conforman una etapa de aprendizaje discriminativo. Su principal objetivo es establecer una frontera de separación entre clases en el dominio transformado (36). De esta manera, al evaluar una muestra de voz dado su modelo de entrenamiento, si sus características caen del lado *target* se puede decir que muy probablemente el usuario de la muestra de voz y el modelo enfrentado sean la misma persona. De igual modo se podrá decir de la clase *non target* si las características del fichero de test se encuentran al otro lado de la frontera. En este tipo de sistemas la puntuación se expresa como la distancia de los vectores al hiperplano de separación. Dada su gran flexibilidad y buen comportamiento es una de las técnicas más populares en la actualidad en el ámbito del reconocimiento de voz.

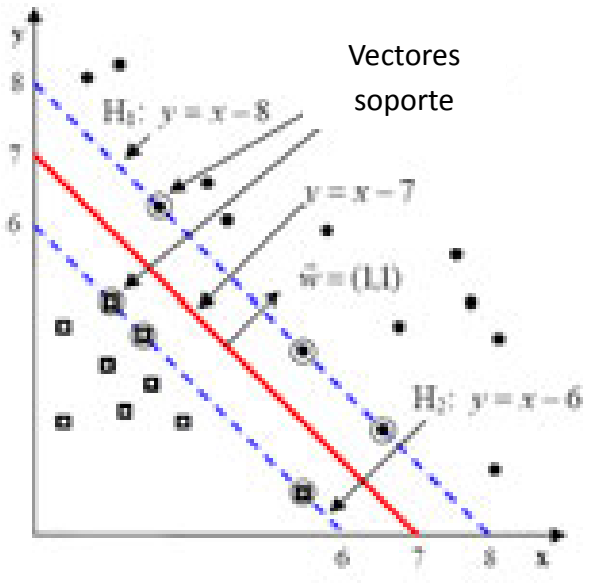
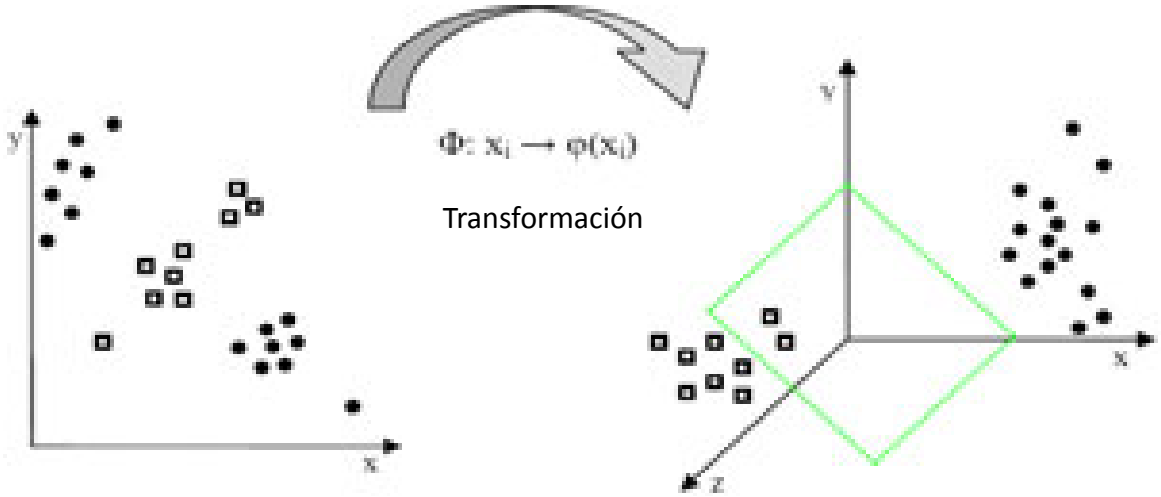


Figura 4.8. Ejemplo de vectores soporte.

En la mayoría de ocasiones los vectores MFCC o LPCC (LPC en el dominio *cepstral*) de dos locutores no son separables en el espacio de características por un hiperplano entrenado con un SVM, por lo que estos vectores se transforman usando técnicas como GLDS (*Generalized Linear Sequence Kernel*), aumentando su dimensión hasta encontrar un hiperplano que separe dichos conjuntos de características. La **Figura 4.8** muestra dos vectores de características pertenecientes a dos locutores diferentes, uno representado mediante círculos negros y el otro mediante cuadrados blancos. Como se puede observar, estos conjuntos no son separables mediante un hiperplano. Sin embargo, llevando dichos

vectores a una dimensión mayor mediante la transformación adecuada se consigue generar un hiperplano que diferencie a ambos locutores, es decir, si otro vector de características puntúa por debajo del hiperplano en el espacio transformado muy probablemente corresponderá al locutor cuyo modelo se rija por los parámetros representados mediante cuadrados (enfrentamiento *target*), y en caso contrario tendrá muchas posibilidades de pertenecer al otro usuario (enfrentamiento *non target*).



a) Espacio de características de 2 dimensiones. **b) Espacio de características de 3 dimensiones.**

Figura 4.9. Representación de la transformación de espacio de características de 2 a 3 dimensiones.

HÍBRIDOS GMM-SVM (*SV O SUPERVECTORS*)

SuperVectors es una técnica híbrida que aprovecha las propiedades de modelo generativo de los sistemas GMM y del modelo discriminativo de los sistemas SVM (37) (38).

El concepto que define esta técnica es el siguiente: mediante un sistema SVM se modelan las desviaciones de los vectores de medias de los modelos, que se obtienen del modelado bajo GMM de los diferentes locutores.

De forma resumida, el procedimiento para implementar este sistema es el siguiente:

1. Se entrena un GMM para cada locución que intervenga en el enfrentamiento.
2. De cada GMM entrenado se extrae el vector de medias de cada gaussiana ponderado por su peso y covarianza para después agruparlos todos en un SV.
3. La discriminación entre SVs *target* y *non target* se realiza mediante un SVM.
4. La puntuación final se obtiene de enfrentar el modelo del locutor correspondiente con el *SV* de la locución de test.

4.5 TÉCNICAS DE COMPENSACIÓN DE VARIABILIDAD INTERSESIÓN

En este apartado se explican diferentes técnicas que se utilizan en la actualidad para minimizar el efecto del ruido y otras perturbaciones introducidas por el canal de transmisión. Estas perturbaciones se engloban en lo que se denomina *variabilidad inter-sesión*, ya explicada en el apartado 2.3.2.

Si bien la disminución de las perturbaciones es un parámetro importante para definir su rendimiento, éste también está caracterizado por el dominio en el que se aplican estas técnicas (a nivel de parámetros, modelos o *scores*), si requieren de entrenamiento o si necesitan trabajar mediante los datos etiquetados previamente del canal.

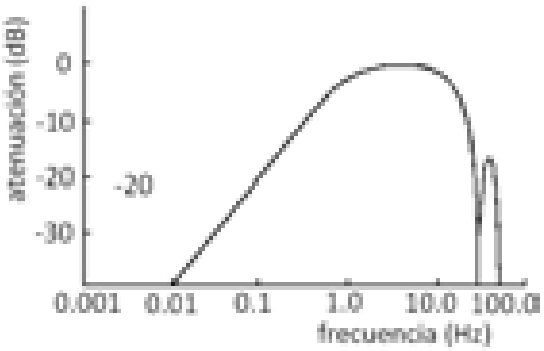
CMN (*CESPTRAL MEAN NORMALIZATION*)

La normalización por media cesptral es una técnica muy popular en el ámbito de normalización de canal debido a su sencillez. Trabaja en el dominio de los parámetros característicos del hablante y no requiere ni entrenamiento ni etiquetado previo del canal. Consiste en, supuesto el canal lineal invariante durante toda la locución, eliminar el efecto del canal en el dominio cesptral eliminando de cada coeficiente su media ponderada calculada a través de toda la locución (39).

FILTRADO RASTA

Explota las diferencias entre las propiedades temporales del habla y las propiedades del canal. Consiste en filtrar cada componente *cepstral* mediante un filtro paso banda bajo la hipótesis de que cualquier constante o componente de variación demasiado lenta o rápida no es habla. El filtro a aplicar se define a continuación (40):

$$H(z) = 0.1 \cdot z^{4} \cdot \frac{2 + z^{-1} - z^{-3} - 2 \cdot z^{-4}}{1 - 98 \cdot z^{-1}}$$



Aunque es también un método que trabaja en el dominio de los parámetros y no requiere ni de entrenamiento ni de etiquetado del canal como CMN, su complejidad es mayor.

Figura 4.10. Respuesta en frecuencia del filtro RASTA.

FILTRADO WIENER

El filtrado *Wiener* es una técnica conocida desde los años 40 pero que no se ha utilizado en reconocimiento de voz hasta la última década. Es un filtro ecualizador diseñado para minimizar el error cuadrático medio entre la señal estimada sin ruido y la señal con ruido. La experiencia ha demostrado que es una técnica que ofrece buenos resultados cuando se procesan datos microfónicos.

FW (*FEATURE WARPING*)

Este método tiene por objeto ajustar la distribución de los parámetros a una distribución gaussiana de media nula y varianza unidad. Se basa en que la distorsión de canal modifica la distribución real de los coeficientes cepstrales en cortos periodos de tiempo (41).

Trabaja en el dominio de los parámetros y no requiere ni entrenamiento ni etiquetado de canal.

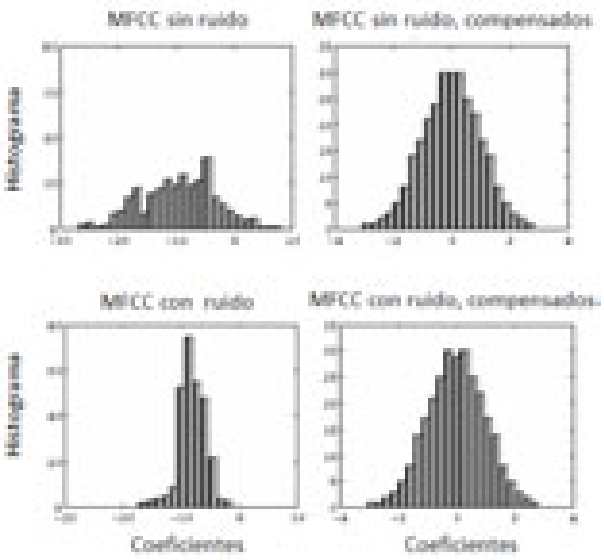


Figura 4.11. Histograma de los coeficientes cepstrales con y sin ruido, antes y después de la compensación.

FM (*FEATURE MAPPING*)

Esta técnica de normalización presenta unos resultados mejores que FW (42). Es una técnica que entrena un conjunto de transformaciones no lineales para asignar un espacio de características neutral a uno dependiente del canal. Esto se logra entrenando un GMM con información de todos los posibles canales convirtiéndolo en un modelo libre de contexto, para posteriormente adaptar ese modelo universal a los parámetros del modelo original. Teniendo un conjunto de GMMs dependientes del contexto, las locuciones de voz pueden enfrentarse con cada uno de ellos y determinar el más parecido para después mapear estas locuciones a un espacio de características sin información de canal. Originalmente la generación de los GMMs dependientes de contexto se hace etiquetando los diferentes canales a modelar.

Se puede concluir que esta técnica trabaja en el dominio de los parámetros como las anteriores pero requiere entrenamiento y etiquetado de canal, lo que la convierte en una técnica más robusta pero a la vez más compleja y costosa.

JFA (*JOINT FACTOR ANALISYS*)

Es una técnica de compensación desarrollada en los últimos años que ha demostrado reducir de forma significativa la influencia del canal en las locuciones (9). Consiste en modelar las direcciones de máxima variabilidad interlocutor e intra-locutor de las características extraídas del habla. A partir de esta información se trata de compensar aquellas variaciones relacionadas con la variabilidad intra-locutor y mantener las variaciones interlocutor. La gran limitación de esta técnica es que su rendimiento depende en gran medida de la disponibilidad de un corpus apropiado deseablemente con las mismas condiciones de la voz a reconocer, lo cual no es frecuente en aplicaciones reales por lo que motiva la definición de nuevos algoritmos de compensación como se presentan en este trabajo (ver capítulo 8).

NAP (*NUISANCE ATTRIBUTE PROJECTION*)

La proyección de atributos indeseados es una técnica de compensación de la variabilidad intersesión (canal, entorno, etc.) (43), y es propiamente una variante de JFA usada con sistemas SVM. Para implementar esta técnica se pueden distinguir dos pasos importantes: por un lado, la creación de una matriz de referencia que se usa para compensar los vectores y por otro lado la compensación en sí de estos. Está demostrado que la variabilidad intersesión sólo afecta, principalmente, a algunas dimensiones de los datos de entrada, por lo que realizando una proyección de los coeficientes en el dominio *cepstral* se eliminará en gran parte el efecto nocivo del canal, pero eso sí, introduciendo una pequeña distorsión en los parámetros de entrada. La matriz de referencia deberá contener la mayor cantidad de datos posible, cuantos más usuarios y más locuciones por usuario mejor se recogerá la variabilidad intersesión.

RESUMEN DE CARACTERÍSTICAS DE LOS MÉTODOS DE COMPENSACIÓN

Los algoritmos más eficientes, en cuanto a coste computacional se refiere, serán los que no requieran ni entrenamiento ni etiquetado del canal, hecho que conlleva a la obtención de peores resultados, por lo que habrá que llegar a un compromiso entre coste computacional y rendimiento obtenido según la disponibilidad de una base de datos apropiada (variable en cuanto a calidad y con diferentes canales de transmisión). Este es el caso de *factor analysis*, que ofrece un rendimiento notable y un coste computacional no muy elevado en la realización de las comparaciones mediante las técnicas de implementación actuales (44). Otro aspecto a considerar de gran importancia es el dominio de trabajo: el rendimiento de los algoritmos que trabajan en el dominio del modelado estará siempre sujeto al sistema usado (GMM, SVM, etc.), mientras que aquellos que trabajen en el dominio de los parámetros serán independientes del sistema.

Es evidente que la mejor solución para lograr eliminar el efecto del canal y poder comparar el modelo del locutor con los parámetros de test en un espacio de características común, es utilizar varios de los métodos brevemente descritos de forma complementaria. De hecho, en el estado del arte es frecuente encontrar sistemas que combinan las técnicas de *Feature Warping* y *Joint Factor Analysis.*

Técnica	Dominio	Requiere entrenamiento	Requiere etiquetado del canal
CMN	Parámetros	NO	NO
RASTA	Parámetros	NO	NO
Wiener	Parámetros	NO	NO
FW	Parámetros	NO	NO
FM	Parámetros	SÍ	SÍ
JFA	Modelado	SÍ	NO
NAP	Parámetros	SÍ	NO

Tabla 4.1. Resumen de las características importantes de los métodos de compensación.

4.6 TÉCNICAS DE NORMALIZACIÓN DE PUNTUACIONES (*SCORES*)

La normalización se define como una transformación de los *scores* de salida, de un sistema de reconocimiento de locutor, con el objetivo de reducir el desalineamiento de las distribuciones *target* y *non target* debido a variaciones en las condiciones de los enfrentamientos (45). Principalmente se emplea para eliminar la dependencia del enfrentamiento con el canal de adquisición pero también sirve para homogeneizar los rangos de los *scores* de cara a la fusión de distintos sistemas (uno puede medir similitudes como GMM y otro distancias como hacen los sistemas SVM). Adicionalmente, como es el caso que en este proyecto se estudia, las normalizaciones ayudan a situar las distribuciones en el mismo rango de cara a situar un umbral común para el sistema y no para cada locutor. A continuación se presentan diferentes técnicas de normalización utilizadas por la comunidad científica.

T-NORM

La técnica de Test-Normalización es ampliamente utilizada en el campo del reconocimiento de locutor (4). Se centra en el fichero de test y su comportamiento frente a otros modelos evitando el desajuste de las distribuciones *non target* (ver **Figura** 4.12). El procedimiento a seguir se define a continuación: a la vez que se enfrenta el fichero de test al modelo bajo estudio, se enfrenta también a una cohorte de modelos de impostores (usuarios distintos al fichero de test), de cuya distribución se obtienen la media y varianza que se aplicará a los enfrentamientos a analizar obteniéndose así un alineamiento de la distribución de probabilidad *non target* dependiente del fichero de test a identificar. Por lo tanto, dicha normalización queda definida mediante la siguiente fórmula:

$$s_{Tnorm} = \frac{s_{raw} - \mu_{Tnorm}}{\sigma_{Tnorm}}$$

donde s_{raw} hace referencia al *score* sin normalización, μ_{Tnorm} y σ_{Tnorm} equivalen a los parámetros de la distribución de los *scores non target* supuesta gaussiana y obtenida a través del enfrentamiento del modelo de test y la cohorte de impostores y s_{Tnorm} se define como el *score* normalizado. A la hora de aplicar T-Norm la selección de la cohorte de modelos es un elemento importante y sujeto a investigación. Debe utilizarse un número elevado de modelos de características similares a los modelos de usuario. Para este trabajo se ha realizado T-Norm con una cohorte de modelos de impostores de la base de datos de NIST SRE 2005, con aproximadamente 100 modelos para cada sexo (la normalización se hace para hombres y mujeres por separado al ser sus características espectrales distintas).

Z-NORM

La técnica de Zero-Normalización explota la misma idea que T-Norm (46): en este caso es el modelo de entrenamiento el que se compara a una cohorte de ficheros de test de usuarios impostores para obtener la distribución *non target* de la cual se extraen la media y varianza para realizar el escalado deseado. Esta normalización viene definida por la siguiente fórmula:

$$s_{Znorm} = \frac{s_{raw} - \mu_{Znorm}}{\sigma_{Znorm}}$$

donde s_{raw} hace referencia al score sin normalización, μ_{Znorm} y σ_{Znorm} equivalen a los parámetros de la distribución de los scores *non target* extraídos del enfrentamiento entre el modelo bajo estudio y la cohorte de ficheros de test y s_{Znorm} se define como el score normalizado. El resultado de aplicar esta técnica será el alineamiento de los scores *non target* dependientes del modelo de entrenamiento para cualquier enfrentamiento del

sistema. Al igual que sucedía con la cohorte de T-Norm la selección de la cohorte de ficheros necesaria para Z-Norm es también muy importante. De nuevo, estos ficheros de voz deben presentar las mismas características que el fichero de test de enfrentamiento para que el escalado resulte en el alineamiento de distribuciones buscado.

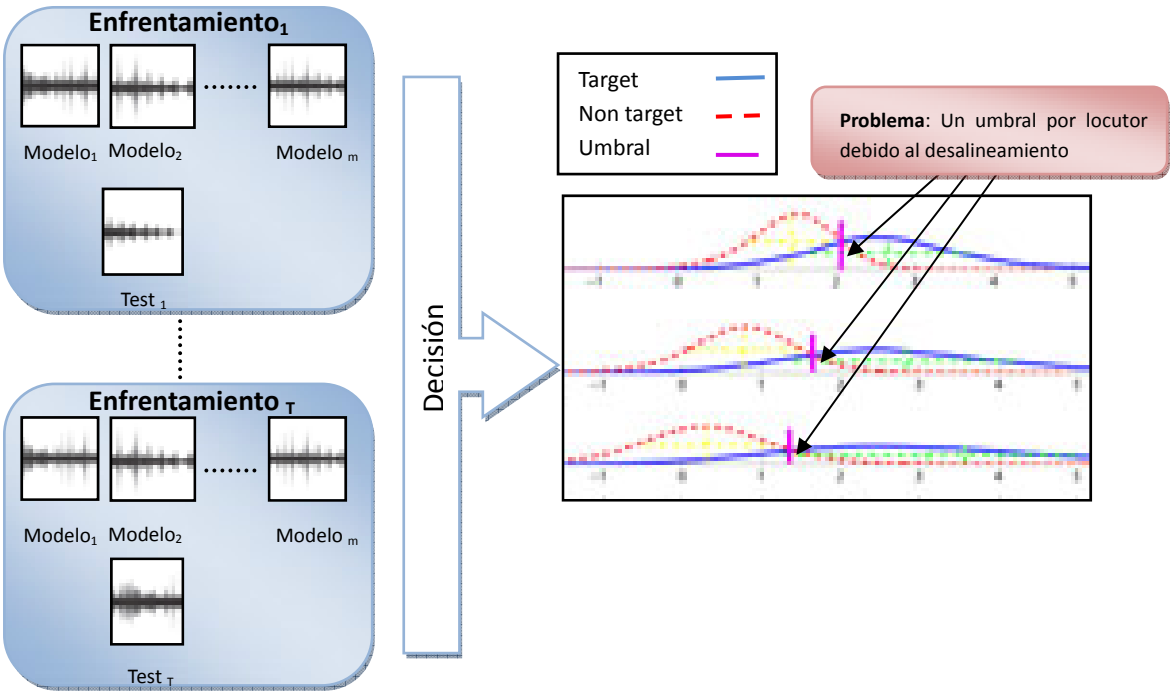


Figura 4.12. Efecto del desalineamiento y motivación de la de T-Normalización.

OTRAS NORMALIZACIONES

En la actualidad, existen numerosas técnicas basadas en T-Norm y en Z-Norm para compensar la variabilidad inter-sesión a nivel de puntuación. Algunas de ellas normalizan por dispositivo (*Handset-Norm*) o por canal (*Channel-Norm*) (49), donde la media y desviación se obtiene de una cohorte de segmentos de test con variabilidad correspondiente a cada caso (dispositivo o canal). Sin embargo, existen otras técnicas que consideran a la vez la información del modelo de entrenamiento y del fichero de test para la normalización de las puntuaciones y en consecuencia, la eliminación de la variabilidad de las muestras, como son las normalizaciones adaptativas KL-T-Norm (*Kullback-Leibler-Test-Normalization*) (33) o AT-Norm (*Adaptative Test-Normalization*) (48). Por otra parte, es acertado pensar que las técnicas anteriores se pueden combinar de diferentes formas para obtener buenos resultados. No obstante, en ocasiones, el coste computacional que introducen es demasiado elevado en comparación al beneficio que se obtiene, lo que lleva a utilizar una sola normalización (como en el sistema inicial de este proyecto) que compense la variabilidad existente. Algunas de estas normalizaciones son ZT-Norm, HT-Norm o CT-Norm.

5 MARCO EXPERIMENTAL

En esta sección se definen los protocolos experimentales seguidos para analizar y compensar el impacto de la variabilidad en las condiciones del habla (duración y otras medidas de calidad), se detallan las bases de datos utilizadas y se describe el sistema mediante el cual se han obtenido los resultados de análisis (capítulos 1 y 7) y compensación (capítulo 9).

5.1 PROTOCOLOS: EVALUACIONES NIST SRE

El organismo norteamericano NIST (*National Institute of Standards and Technology*) (8) organiza evaluaciones bianuales abiertas de carácter competitivo en las que se elaboran bases de datos y se definen una serie de tareas o protocolos para medir de manera objetiva el rendimiento, bajo las mismas condiciones, de los sistemas presentados. Las bases de datos utilizadas para el desarrollo de este trabajo, derivadas de las de NIST *Speaker Recognition Evaluation* (SRE) y diferentes para la compensación de variabilidad en duración y calidad, se presentan a continuación.

BASES DE DATOS PARA EL ANÁLISIS Y COMPENSACIÓN DE LA VARIABILIDAD EN DURACIÓN

Para medir el análisis del rendimiento de los sistemas actuales en función de la duración de las muestra de voz, es necesario disponer de una base de datos específica que contenga suficientes archivos de cualquier longitud. Para ello, se han creado dos bases de datos, denominadas *DurTelSRE06* y *DurTelSRE08* a partir de las originales de las evaluaciones NIST, las cuales que se describen a continuación.

Las bases de datos originales utilizadas son NIST SRE 2006 protocolo *1con4w-1conv4w* (9), utilizada para entrenar los modelos de compensación, y NIST SRE 2008 protocolo short2-*short3* parte telefónica (3) para testearlos. En ambos casos los archivos provienen de hablantes masculinos y femeninos. Son de carácter conversacional y de aproximadamente 5 minutos de duración una vez suprimidos lo silencios, de los cuales típicamente 2,5 minutos corresponden a cada locutor. La composición de estas bases puede verse en la **Tabla** 5.1 y **Tabla** 5.2.

NIST SRE 2006						NIST SRE 2008					
Hombres		Mujeres		Ambos		Hombres		Mujeres		Ambos	
Modelos	Tests	Modelos	Tests	Modelos	Tests	Modelos	Tests	Modelos	Tests	Modelos	Tests
352	1.604	462	2.127	814	3.731	648	895	1.140	1.678	1.788	2.573

Tabla 5.1. Número de archivos de las bases de datos de NIST SRE 2006 y 2008.

NIST SRE 06		NIST SRE 08	
Target	Non target	Target	Non target
3.613	47.376	3.832	33.218

Tabla 5.2. Enfrentamientos para las bases de datos de NIST SRE 2006 y 2008 (parte telefónica).

Para generar las bases de datos *DurTelSRE06* y *DurTelSRE08* se han dividido los ficheros de audio originales para obtener ficheros de habla con una duración menor tras eliminar los silencios. En concreto, se han generado archivos de 3, 10, 15, 20, 30, 40, 50, 60, 100 y 150 segundos de habla neta (después de eliminar los silencios). El número de archivos y enfrentamientos resultantes se presenta en la **Tabla** 5.3 y la **Tabla** 5.6 de forma respectiva, donde la columna *Modelos* hace referencia al número de ficheros de voz creados para entrenar el modelo del locutor y *Tests* al número de ficheros de habla a identificar. Observando los cuadros, se puede apreciar el esfuerzo y tiempo computacional necesario para la realización de estos experimentos (**Tabla** 5.3, **Tabla** 5.4, **Tabla** 5.6 y **Tabla** 5.5).

	DurTelSRE06						DurTelSRE08					
	Hombres		Mujeres		Ambos		Hombres		Mujeres		Ambos	
Duración	Modelos	Tests	Modelos	Tests	Modelos	Tests	Modelos	Tests	Modelos	Tests	Modelos	Tests
3	352	1604	461	2125	813	3729	648	895	1140	1678	1788	2573
10	351	1604	461	2124	812	3728	648	895	1140	1678	1788	2573
15	350	1604	461	2123	811	3727	648	895	1140	1678	1788	2573
20	349	1604	461	2123	810	3727	648	895	1140	1677	1788	2573
30	349	1599	459	2120	808	3719	646	894	1135	1672	1781	2566
40	347	1586	456	2107	803	3693	641	887	1131	1664	1772	2551
50	345	1568	451	2089	796	3657	634	866	1123	1638	1757	2504
60	341	1535	445	2048	786	3583	612	844	1097	1599	1709	2443
100	243	1166	324	1572	567	2738	435	593	842	1224	1277	1818
150	80	380	92	510	172	890	142	205	304	463	446	668

Tabla 5.3. Número de archivos de las bases de datos de *DurTelSRE06* y *DurTelSRE08*.

A continuación, se representa de forma cuantitativa (**Tabla** 5.4) y visual (**Figura** 5.1) la cantidad total de ficheros creados para cada base de datos, siendo interesante observar el escaso número de archivos de 150 segundos reales resultante.

	DurTelSRE06		DurTelSRE08	
Duración	Modelos	Tests	Modelos	Tests
3	813	3729	1788	2573
10	812	3728	1788	2573
15	811	3727	1788	2573
20	810	3727	1788	2573
30	808	3719	1781	2566
40	803	3693	1772	2551
50	796	3657	1757	2504
60	786	3583	1709	2443
100	567	2738	1277	1818
150	172	890	446	668

Tabla 5.4. Número de archivos generados para las bases de datos DurTel06 y DurTel08.

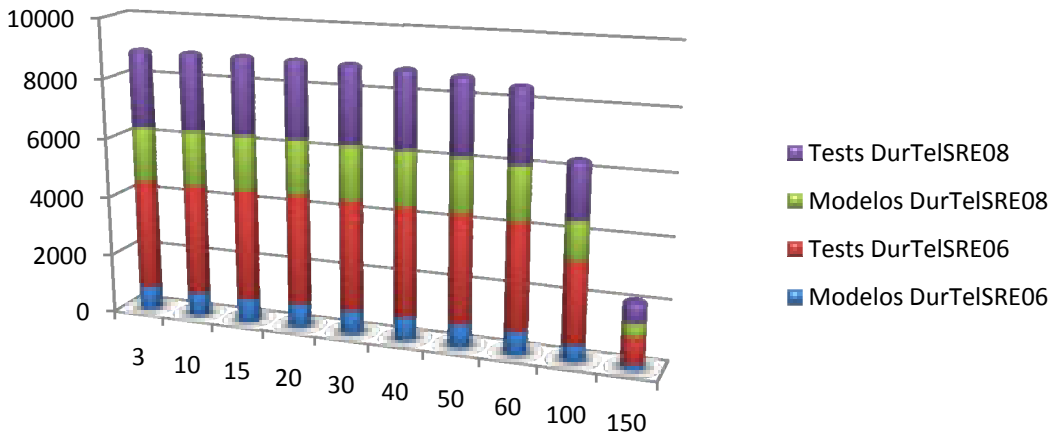


Figura 5.1. Número de modelos de entrenamiento y ficheros de test de las bases de datos *DurTelSRE06* y *DurTelSRE08*.

Por último y para entender la magnitud de estos experimentos se presentan la **Tabla** 5.6 y la **Tabla** 5.5, donde se indica el número de enfrentamientos totales y el número de ficheros por conjunto de enfrentamiento dada la longitud o duración del modelo de entrenamiento y del fichero de test a identificar. Los experimentos globales incluyen la caracterización del rendimiento de todos los modelos frente a todos los ficheros tal y como se indica en las evaluaciones de NIST SRE 2006 y 2008.

DurTelSRE06		DurTelSRE08	
Target	**Non target**	**Target**	**Non target**
286.556	3.726.243	306.853	2.633.685

Tabla 5.5. Número total de enfrentamientos para las bases de datos DurTelSRE06 y DurTelSRE08.

	DurTelSRE06		DurTelSRE08	
Duración del enfrentamiento	Target	Non target	Target	Non target
3,3	3613	47367	3832	33218
10,10	3610	47314	3832	33218
15,15	3610	47222	3832	33218
20,20	3605	47177	3832	33218
30,30	3593	46951	3807	32989
40,40	3550	46404	3769	32684
50,50	3496	45567	3697	31885
60,60	3406	44095	3523	30330
100,100	1970	24281	2297	17233
150,150	312	2314	449	2289

Tabla 5.6. Enfrentamientos para las bases de datos DurTelSRE06 y DurTelSRE08.

BASES DE DATOS PARA EL ANÁLISIS Y COMPENSACIÓN DE LA VARIABILIDAD EN CALIDAD

Para realizar el análisis del impacto de la calidad se han introducido muestras microfónicas en los experimentos con el objetivo de aumentar la variabilidad de los indicadores de degradación, dando lugar a dos tipos de experimentos que utilizan bases de datos distintas:

- **Compensación de ficheros telefónicos:** entrenando los algoritmos con la base de datos de NIST SRE 2006 protocolo *1con4w 1conv4w* y evaluando los resultados sobre la base de datos de NIST SRE 2008 protocolo *short2-short3 (parte telefónica).*

- **Compensación de ficheros telefónicos y microfónicos**: en este caso se ha utilizado únicamente la base de datos de NIST SRE 2008 protocolo *short2-short3* en sus cuatro variantes debido a la no existencia de muestras microfónicas de gran variabilidad en 2006. Las cuatro variantes o condiciones de 2008, que se enuncia n a continuación, se diferencian en función de la procedencia de la voz a utilizar para entrenar los modelos de locutor y los ficheros de test a reconocer:

  - ***tel-tel***: en el que el modelo de entrenamiento y el fichero de test han sido adquiridos de un canal telefónico.

  - ***tel-mic***: en el que el fichero de audio con el que se entrena el modelo del usuario a identificar ha sido adquirido a través de un canal telefónico y el fichero de test mediante un micrófono.

  - ***mic-tel***: en el que el modelo se ha extraído de una grabación con

micrófonos y el archivo de enfrentamiento se ha capturado a través de la red telefónica.

- ***mic-mic***: en la que el modelo y fichero de test han sido capturados con un dispositivo microfónico.

Cabe destacar que en esta base de datos el habla microfónica ha sido capturada con distintos micrófonos y a distintas distancias para introducir variabilidad en las muestras. El formato de estas muestras puede ser similar al anterior (conversacional de 5 minutos para los dos locutores) o tipo entrevista, en el que el fichero de voz es de aproximadamente 3 minutos de duración donde principalmente habla el entrevistado. Es importante remarcar que tanto los ficheros de entrenamiento como los de enfrentamiento de tipo microfónico, presentan un filtrado *Wiener* ya que se ha demostrado en el grupo ATVS que este tipo de procesado ayuda a mejorar el rendimiento de sistemas que trabajan con estas muestras.

La elección de estas bases de datos se ha llevado a cabo debido a que los protocolos NIST se encuentran ampliamente extendidos en la actualidad por lo que han sido fuertemente estudiados por los distintos grupos de reconocimiento biométrico existentes en el panorama internacional como es el ATVS.

El análisis detallado de las bases de datos se abordará en el apartado 7.3, una vez explicadas las medidas de calidad de las cuales dependen.

5.2 SISTEMA UTILIZADO

El sistema de desarrollo que ha permitido analizar el impacto de la variabilidad en duración y calidad y experimentar con los métodos propuestos, es el presentado por el grupo ATVS en la evaluación NIST SRE 2008 (ver **Figura 5.2**). Este sistema está basado en un modelo GMM de 1024 mezclas de las siguientes características.

EXTRACCIÓN DE CARACTERÍSTICAS

- Uso de ventanas *Hamming* de 20 ms solapadas al 50% (10 ms).
- Extracción de MFCC: 20 filtros espaciados según la escala MEL de 300 a 3300 Hz.
- 38 parámetros en total: 19 MFCC y 19 Δ.

COMPENSACIÓN

- CMN (*Cesptral Mean Normalization).*
- Filtrado RASTA.
- Filtrado *Wiener* para todas las condiciones salvo para *tel-tel*.
- *Feature Warping.*
- *NAP (Nuisance Attribute Projection).*
- *Joint Factor Analysis*
- T-Norm (dependiente de género) a través de una cohorte de modelos de NIST SRE 2005 compuesta por 100 locutores para las condiciones *tel-tel* y *tel-mic,* y 100 locutores telefónicos y 240 microfónicos para las condiciones *mic-tel* y *mic-mic*.

MODELADO GMM

- Entrenado con dos UBMs distintos, uno para la condición *tel-tel* y otro para las condiciones *tel-mic, mic-tel* y *mic-mic*. Estos UBMs se han generado mediante modelos de 1024 mezclas gaussianas y estimación ML a través del algoritmo EM.
- Adaptación MAP.

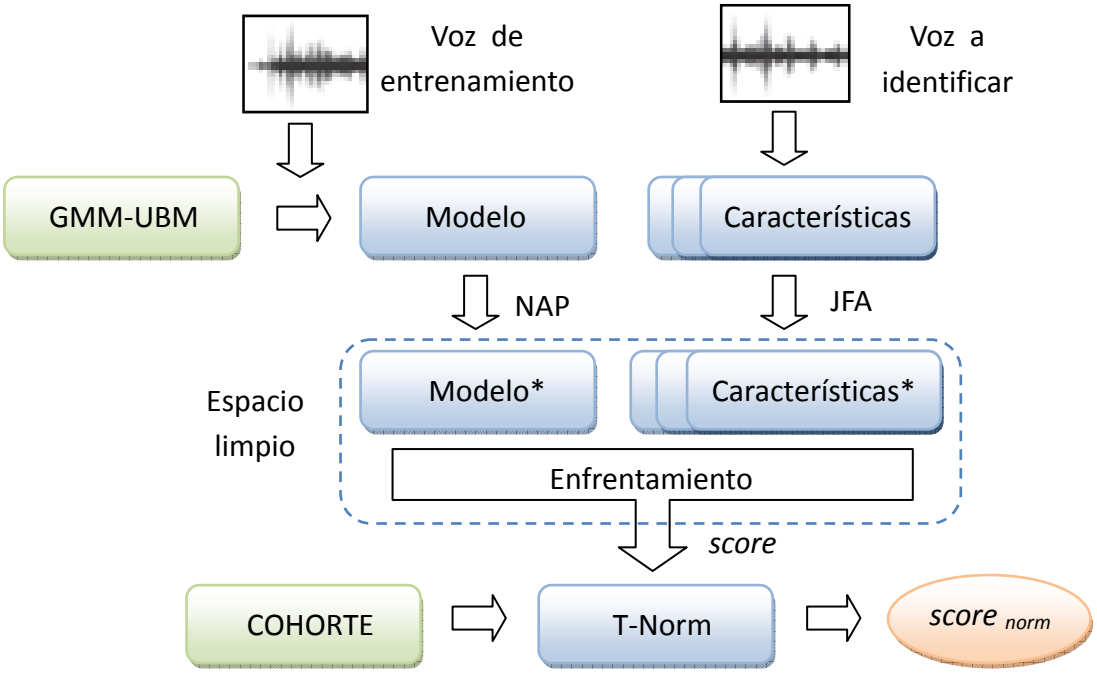


Figura 5.2. Diagrama de bloques del sistema utilizado.

6 ANÁLISIS DEL IMPACTO DE LA DURACIÓN

6.1 DESCRIPCIÓN GENERAL DEL PROBLEMA

En este capítulo se analiza el impacto de la variabilidad en la duración de las muestras de entrenamiento y los ficheros de test a reconocer en el rendimiento de los sistemas de verificación de locutor.

A la fecha de hoy, no existen muchos estudios en esta línea de investigación que incluyan algoritmos específicos de compensación o incluso analicen el problema de forma precisa, lo cual constituye una contribución misma de este proyecto. Este hecho se debe principalmente a la propia configuración de las tareas en las evaluaciones NIST, en las que existe mucha información de variabilidad en cuanto calidad pero apenas varía la duración de los archivos de voz a utilizar, oscilando estos alrededor de los 150 segundos. No obstante, existe un gran conjunto de escenarios de aplicación donde la longitud de las muestras de voz a reconocer puede variar, como es el caso de las aplicaciones forenses (50), donde la variación de la locución a identificar implicaría crear un escenario similar a las condiciones de test para poder reconocer de forma robusta al sujeto, lo que resulta virtualmente impracticable debido a la muy posible no cooperación de éste.

6.2 IMPACTO DE LA VARIABILIDAD EN DURACIÓN

El impacto de la variación en la longitud del habla, entendiendo longitud del habla como el número de muestras de su fichero de audio, se traduce en un desalineamiento de las distribuciones *target* y *non target* de *scores* en función de la duración de los archivos involucrados en el enfrentamiento, lo cual supone una degradación del rendimiento global del sistema cuando se evalúa de forma conjunta esas puntuaciones dependientes de la longitud de los archivos (6). Con el objetivo de estudiar la influencia de la configuración de las muestras en relación al sistema, el estudio llevado a cabo presenta dos objetivos principales:

- El principal objetivo de este análisis es determinar el comportamiento de la variabilidad de la duración en test habiendo usado T-Norm definido mediante una cohorte de modelos de impostores de duración fija para la normalización de *scores*. Como se verá a lo largo de este apartado, debido a que el alineamiento de las distribuciones *non target* no es total, causa que el EER, aunque alcance valores

relativamente aceptables para cada subconjunto de puntuaciones dependientes de la duración del fichero de test, sea peor para el conjunto de *scores* global. Pero este desalineamiento no sólo afecta a la discriminación del sistema en conjunto, sino que también se presenta como un problema a la hora de establecer un único umbral, o de homogeneizar las distribuciones en un mismo rango para fusionar este sistema con otros (50). Este hecho motiva el uso de algoritmos que se estudiarán en la sección 8 para compensar a nivel de *score* esta variación.

- Como objetivo secundario y ya que la base de datos se ha diseñado para que tenga variabilidad en cuanto a duración del fichero de test y del modelo de entrenamiento, se ha procedido estudiar el impacto de la duración de éste último sin tener en cuenta la duración del test. En este caso el EER global se verá a lo largo del capítulo que es aún peor (33.47%) debido a la no implementación de una normalización que alinee las distribuciones dependiente de la duración del modelo como Z-Norm. Este hecho permitirá observar el funcionamiento de los algoritmos de compensación propuestos dependiendo de si se utiliza normalización de puntuaciones dependiente de modelo o no en trabajos futuros.

Con el objetivo de realizar un estudio minucioso del impacto de la variabilidad en duración se han analizado varios tipos de figuras:

- **EER en 3D**: en función de la duración del modelo y del fichero de test.
- **MinCllr en 3D**: en función de la duración del modelo y del fichero de test.
- **Curvas de nivel para la gráfica EER en 3D**: proyección de la gráfica EER en 3D.
- **Curvas de nivel para la gráfica MinCllr en 3D**: proyección de la gráfica MinCllr en 3D.
- **Curvas DET D_{test}:** sobre los subconjuntos de *scores* dependientes de duración agrupados por longitud del fichero de test ignorando la del modelo (es decir, todas las duraciones de los diferentes ficheros que entrenaron los modelos se consideran de forma simultánea).
- **Curvas DET D_{modelo}**: sobre los subconjuntos de *scores* dependientes de duración agrupados por longitud del modelo de entrenamiento ignorando la del test (es decir, todas las duraciones de los diferentes ficheros de test se consideran de forma simultánea).
- **Distribuciones *Kernel* D_{test}:** distribuciones de probabilidad sobre los subconjuntos de *scores* dependientes de duración agrupados por longitud del fichero de test ignorando la del modelo.
- **Distribuciones *Kernel* D_{modelo}:** distribuciones de probabilidad sobre los subconjuntos de *scores* dependientes de duración agrupados por longitud del modelo de entrenamiento ignorando la del test.

- **Curvas de media (μ) y desviación típica (σ) D_{test}:** para las distribuciones *target* y *non target*.
- **Curvas de media (μ) y desviación típica (σ) D_{modelo}:** para las distribuciones *target* y *non target*.
- **Curvas en 3D de media (μ) y desviación típica (σ) D_{test}:** para las distribuciones *target* y *non target*.
- **Curvas en 3D de resta de medias:** entre las medias *target* y *non target*.
- **Curvas en 3D de resta de desviaciones:** entre las desviaciones típicas *target* y *non target*.

Para entender en profundidad el problema a tratar se representa de la **Figura 6.1** a la **Figura 6.4**, donde se ilustra el efecto en cuanto a rendimiento en condiciones de duración extrema para las bases de datos *DurTelSRE06* y *DurTelSRE08*:

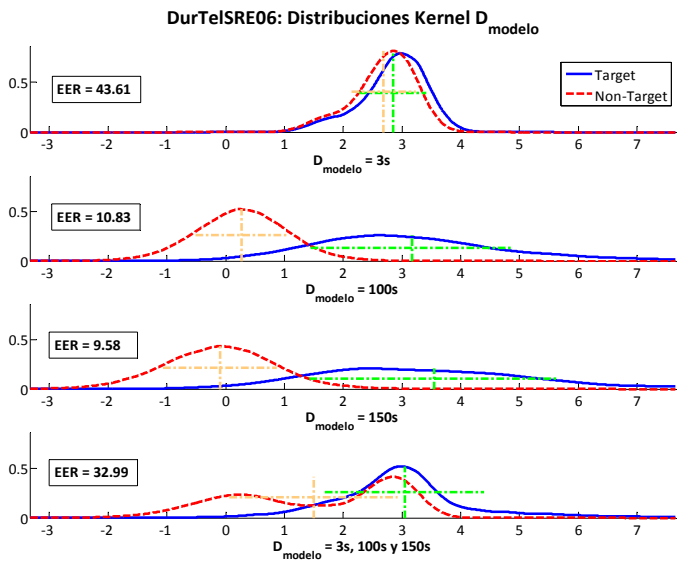
- La figura a), en cada caso, presenta las distribuciones *Kernel*[3] de *scores target* y *non target* generadas mediante el sistema GMM-UBM del ATVS implementado para las evaluaciones de NIST SRE 2008 (apartado 5.2). Cada par de distribuciones ha sido generado por el mismo conjunto de modelos enfrentados a los mismos archivos de test de duraciones extremas de 3, 100 y 150 segundos, dando lugar a un desalineamiento en éstas como se indica por la representación de la media de las distribuciones *target* y *non target* en color naranja y verde respectivamente.

- En la figura b), en cada caso, se encuentran sus curvas DET equivalentes.

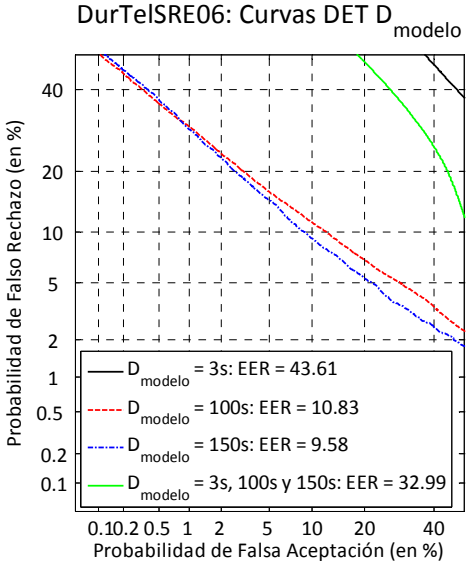
Interpretando estas representaciones, se puede ver cómo a medida que aumenta la duración del archivo de test o del modelo la EER disminuye aumentando la discriminación del sistema (ver curvas DET) ya que se dispone de más información característica perteneciente al usuario. Por otra parte, observando las distribuciones se diferencia claramente como cuando la duración del archivo de audio es pequeña, ya sea para entrenar el modelo (**Figura 6.1** y **Figura 6.3**) o el fichero de test a reconocer (**Figura 6.2** y **Figura 6.4**), sus distribuciones se encuentran casi totalmente solapadas derivando en cerca del peor EER posible (50%). Sin embargo, cuando la duración aumenta éstas se separan contribuyendo a la robustez del sistema. Por lo tanto, viendo el fuerte desalineamiento en cualquiera de los cuatro casos (*DurTelSRE06 y 08* para D_{modelo} y D_{test}) se puede entender la magnitud del problema. Por citar un ejemplo, sería muy subóptimo establecer un umbral de decisión para las distribuciones de la **Figura 6.1** ya que existe una diferencia en EER de 33 puntos entre las locuciones de 3 y 150 segundos, por lo que situar el umbral en

[3] Distribución de probabilidad. Para más información ver (77).

el punto adecuado para la distribución de 3 segundos implicaría considerar todos los usuarios como impostores si su modelo se entrenara con más información (100 y 150 segundos en el ejemplo). Este efecto puede verse con claridad en la última fila de las figuras ya que en ella se representa el desajuste considerando un sistema que trabaje con cualquier duración de entrada.

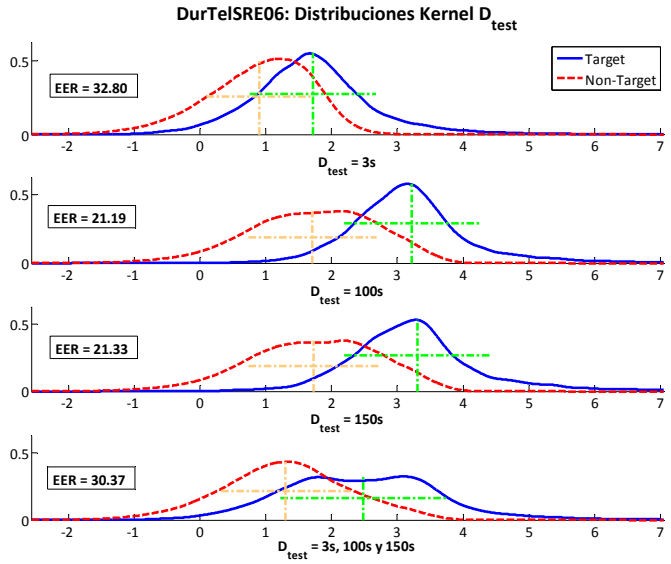


a) **Distribuciones *Kernel* para *DurTelSRE06* y duración variable en el modelo.**

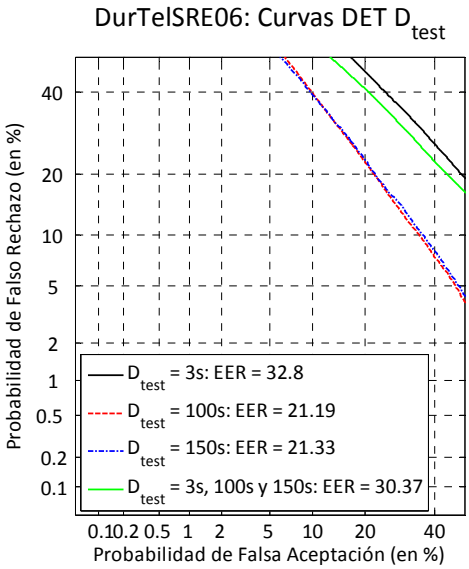


b) **Curvas DET para *DurTelSRE06* y duración variable en el modelo.**

Figura 6.1. Efecto del desalineamiento debido a la variabilidad en la duración del modelo (*DurTelSRE06*).

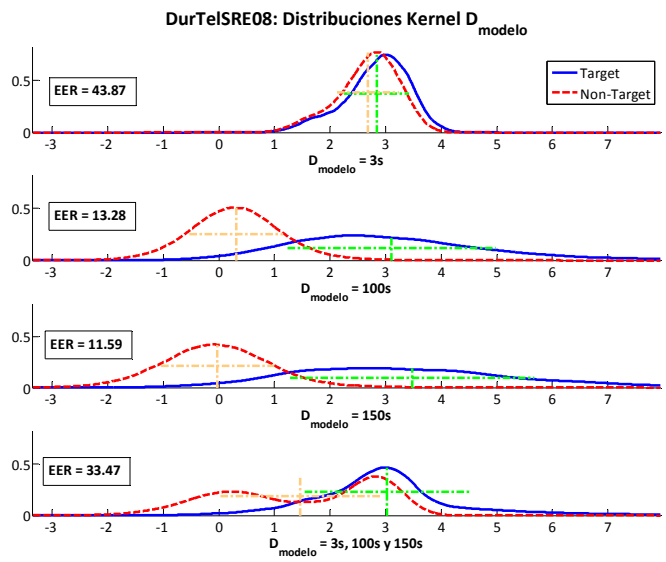


a) **Distribuciones *Kernel* para *DurTelSRE06* y duración variable en el test.**

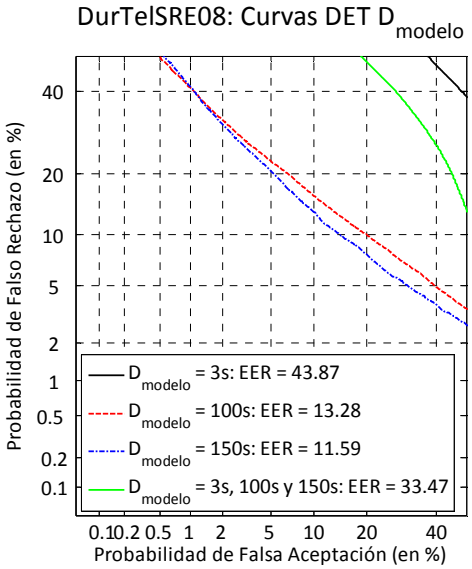


b) **Curvas DET para *DurTelSRE06* y duración variable en el test.**

Figura 6.2. Efecto del desalineamiento debido a la variabilidad en la duración del test (*DurTelSRE06*).

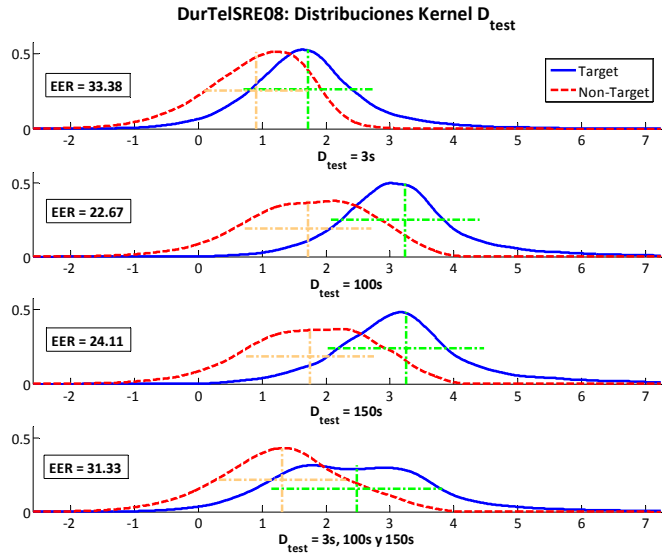


a) **Distribuciones *Kernel* para *DurTelSRE08* y duración variable en el modelo.**

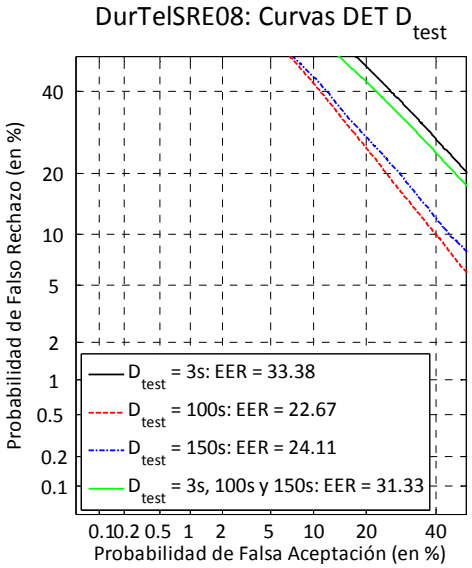


b) **Curvas DET para *DurTelSRE08* y duración variable en el modelo.**

Figura 6.3. Efecto del desalineamiento debido a la variabilidad en la duración del modelo (*DurTelSRE08*).



a) **Distribuciones *Kernel* para *DurTelSRE08* y duración variable en el test.**



b) **Curvas DET para *DurTelSRE08* y duración variable en el test.**

Figura 6.4. Efecto del desalineamiento debido a la variabilidad en la duración del test (*DurTelSRE08*).

Por lo tanto, y a la vista de los resultados obtenidos, cuando se disponga de un sistema que puede trabajar con archivos de voz de múltiples duraciones, es recomendable disgregar el problema en tantas partes como longitudes se dispongan para aplicar la normalización pertinente a cada locución y después evaluar el sistema de forma conjunta.

Otra forma de ver la influencia de la duración en el rendimiento del sistema es evaluar la tendencia de la media y la desviación típica de las distribuciones en función de la duración. De esta manera, se puede apreciar cómo según aumenta la duración las medias de las distribuciones se alejan reduciendo la PFR y la PFA (área de solape dado un umbral) aumentando la discriminación del sistema. Este aumento en la discriminación se ve afectado de forma negativa por un aumento en la desviación típica de las curvas, el cual es menos significativo que el aumento de la diferencia entre las medias si se observa la mejora en el rendimiento global del sistema.

La **Figura** 6.5 y la **Figura** 6.6 representan esta tendencia. En cada caso, la gráfica superior hace referencia a la media de las distribuciones representas en color negro y rojo (*target* y *non target* respectivamente) y la gráfica inferior a su desviación típica. Cabe destacar que, para evaluar la tentendicia de estas variables, se ha representado una curva continua entre 3 y 150 segundos realizada a partir de la interpolación cúbica de los estadísticos de los subconjuntos de *scores* dependientes de duración, siendo en el primer caso (**Figura** 6.5) la longitud del modelo ignorada (se evalúan todas las disponibles de forma simultánea) y en el segundo la del test (**Figura** 6.6).

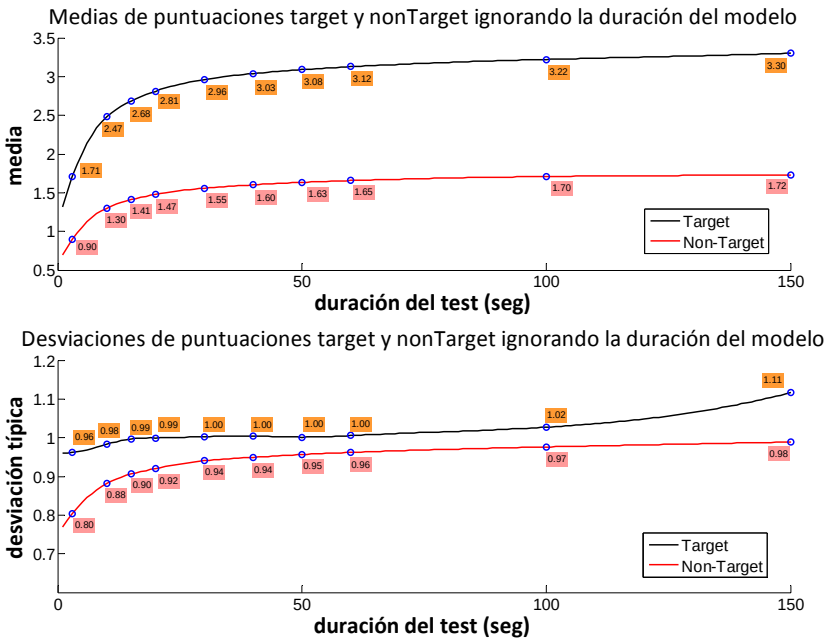


Figura 6.5. Media y desviación de las distribuciones en función de la duración del test (DurTelSRE06).

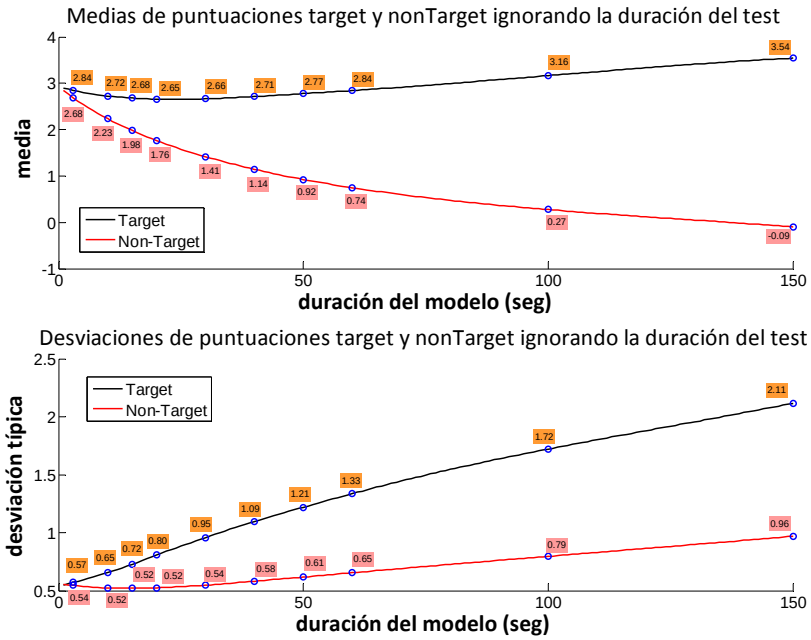
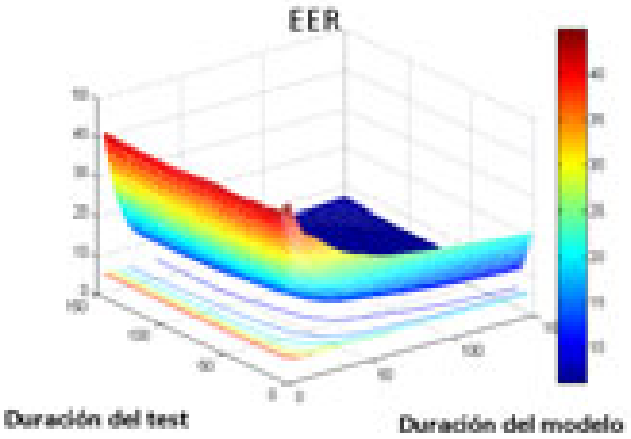
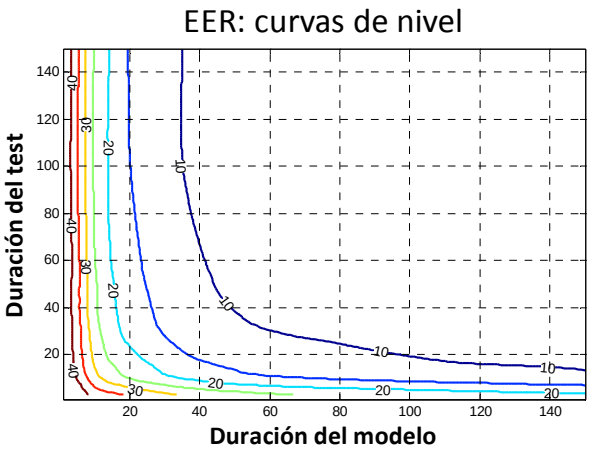


Figura 6.6. Media y desviación de las distribuciones en función de la duración del modelo (DurTelSRE06).

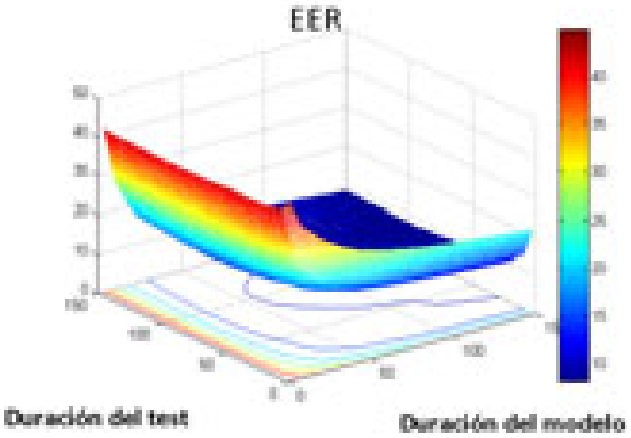
Por otra parte, se sabe que la variación de la duración del modelo y del test influye de manera negativa cuando de reconocer a un usuario se trata, pero *¿cuál de las dos variaciones es más influyente? ¿Cuál es más crítica? ¿Es preferible disponer de una muestra de voz del locutor de una duración alta para modelar de forma correctamente sus características espectrales sin importar la duración del archivo a autenticar?* El siguiente conjunto de figuras responde a esta pregunta. En ellas se puede ver la evolución del EER y del $MinC_{llr}$ en función de la duración del modelo de entrenamiento y del fichero de test, observándose un peor rendimiento cuando la duración del modelo disminuye de forma drástica. Sin embargo, que la duración del test sea baja sólo influye de forma muy negativa cuando la del modelo también lo es. Por lo tanto, se puede concluir que la autenticación será más fiable cuanta más información se tenga del modelo y que la falta de longitud de la muestra de test es menos crítica debido a la T-Normalización llevada a cabo.



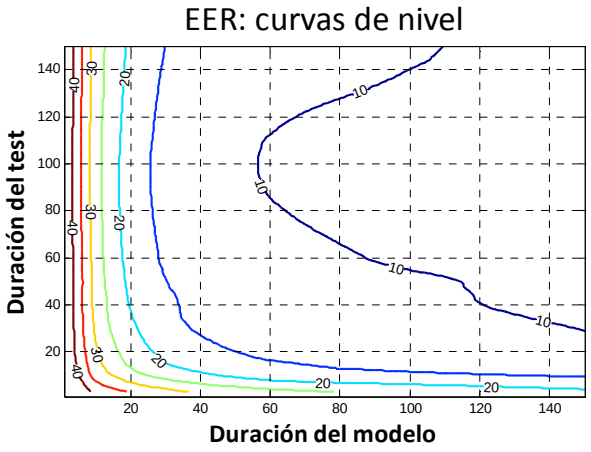
a) **EER en función de la duración del modelo y del test para *DurTelSRE06*.**



b) **EER en función de la duración del modelo y del test para *DurTelSRE06*.**

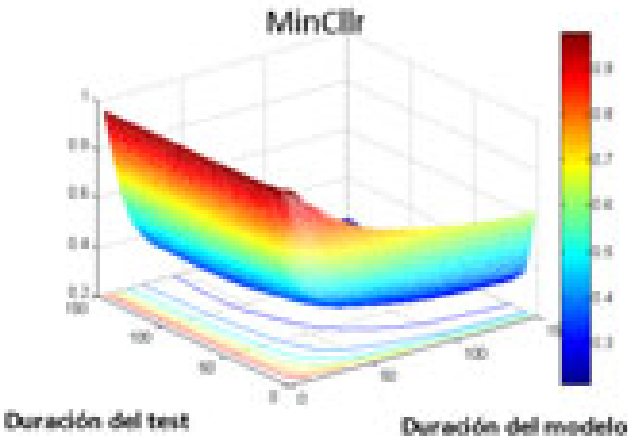


c) **EER en función de la duración del modelo y del test para *DurTelSRE08*.**

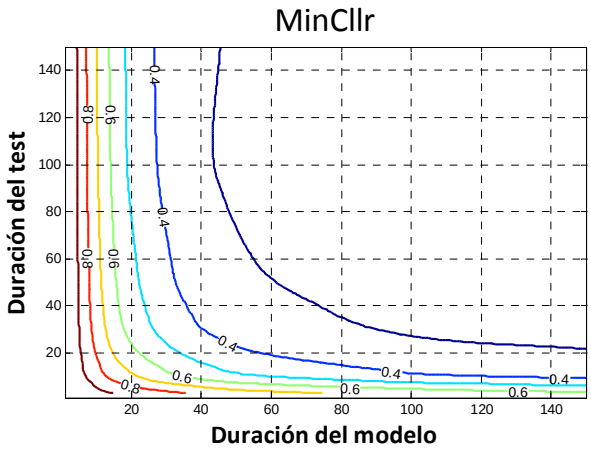


d) **EER en función de la duración del modelo y del test para *DurTelSRE08*.**

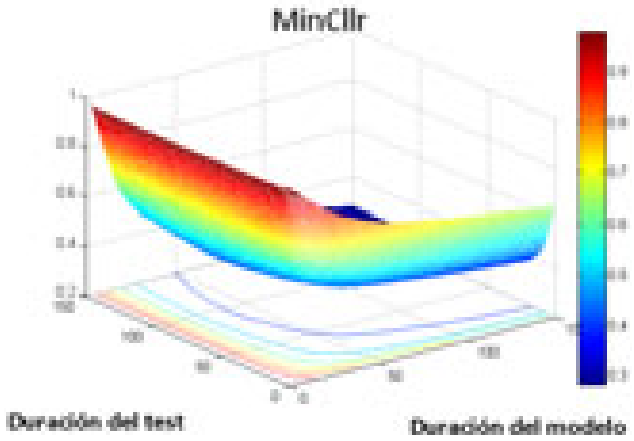
Figura 6.7. Evolución del EER en función de la duración del modelo y del test (DurTelSRE06 y DurTelSRE08).



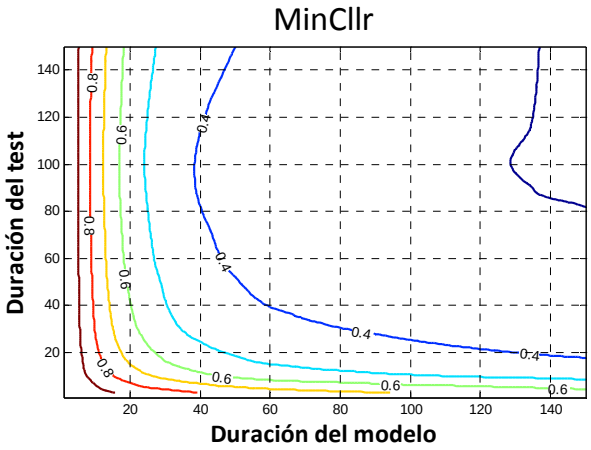
a) **MinCllr en función de la duración del modelo y del test para *DurTelSRE06*.**



b) **MinCllr en función de la duración del modelo y del test para *DurTelSRE06*.**



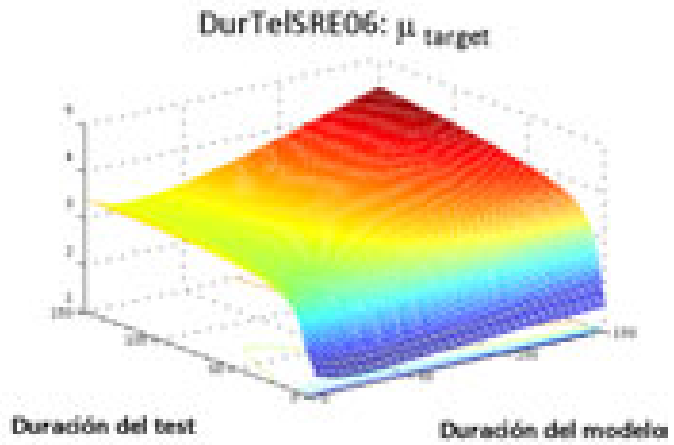
c) **MinCllr en función de la duración del modelo y del test para *DurTelSRE08*.**

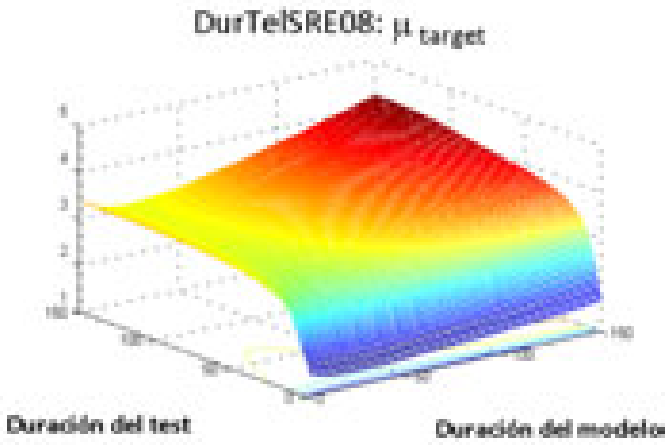


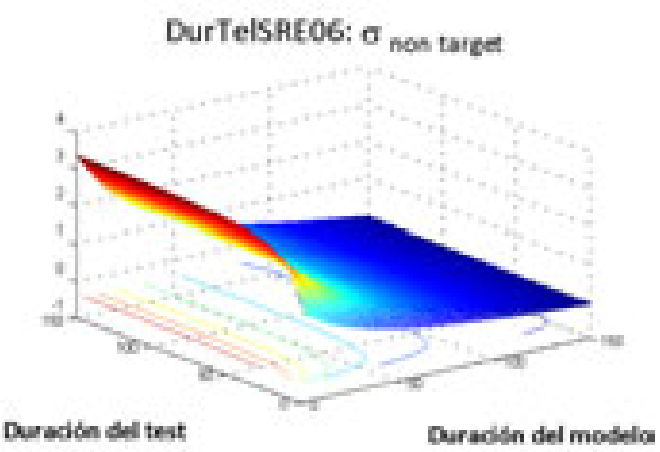
d) **MinCllr en función de la duración del modelo y del test para *DurTelSRE08*.**

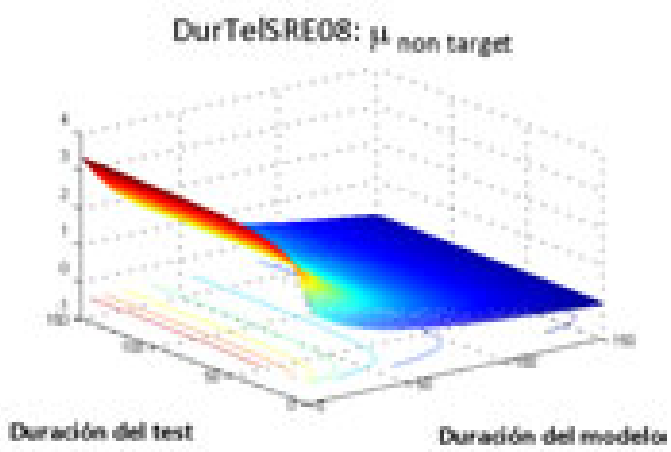
Figura 6.8. Evolución del $MinC_{llr}$ en función de la duración del modelo y del test (DurTelSRE06 y DurTelSRE08).

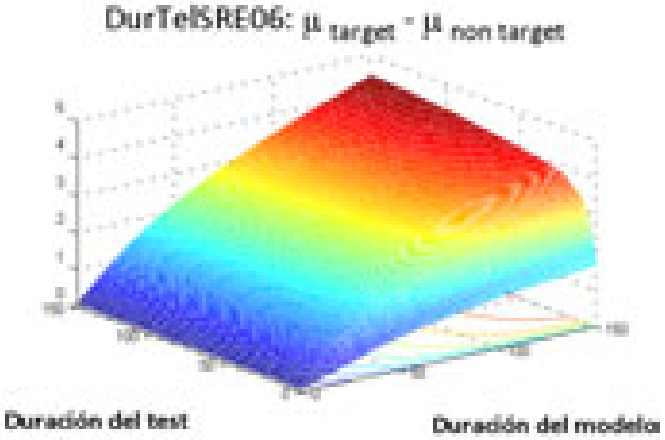
Otra forma de visualizar el buen comportamiento del sistema cuando la duración de las muestras de voz se enfrentan es elevada, es mediante la representación en 3D de las medias y desviaciones típicas de las distribuciones *target* y *non target* en función de la duración del modelo de entrenamiento y el fichero de test (**Figura** 6.9). Para observar la evolución de éstas, también se representa la resta de medias y desviaciones para las dos distribuciones. De nuevo, y para ambas bases de datos, se puede observar cómo la tendencia de las medias es la de separarse, contribuyendo en mayor medida que el aumento de las desviaciones, a que exista una menor área de solape entre las distribuciones respectivas siendo la discriminación del sistema mayor. Hay que destacar que al igual que para el caso en 2D las curvas han sido obtenidas mediante la interpolación de los estadísticos derivados de la evaluación de las duraciones creadas para la base de datos *DurTelSRE06*.

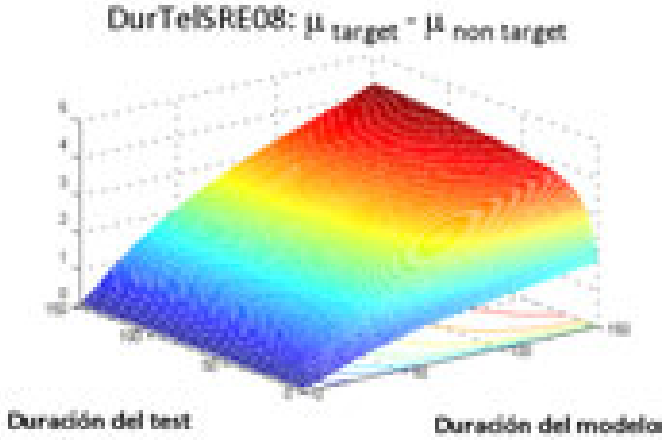


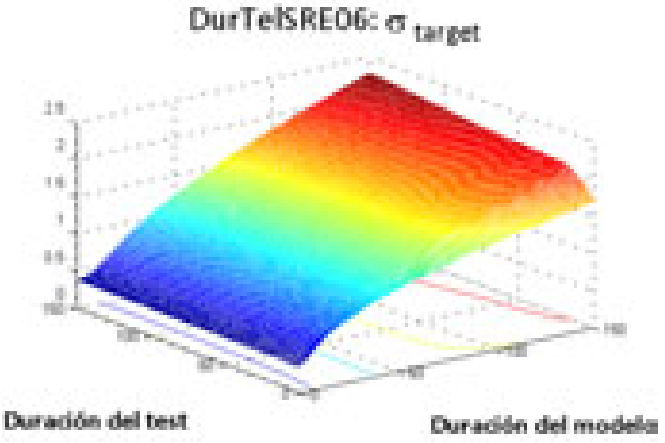


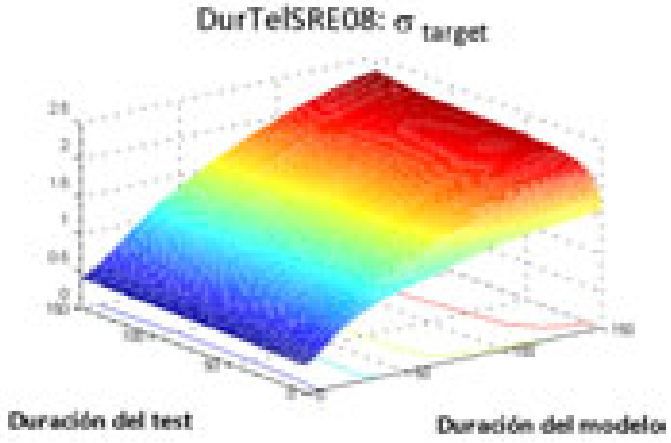

DurTelSRE06: σ non target
Duración del test
Duración del modelo


DurTelSRE08: μ non target
Duración del test
Duración del modelo


DurTelSRE06: μ target * μ non target
Duración del test
Duración del modelo


DurTelSRE08: μ target * μ non target
Duración del test
Duración del modelo


DurTelSRE06: σ target
Duración del test
Duración del modelo


DurTelSRE08: σ target
Duración del test
Duración del modelo

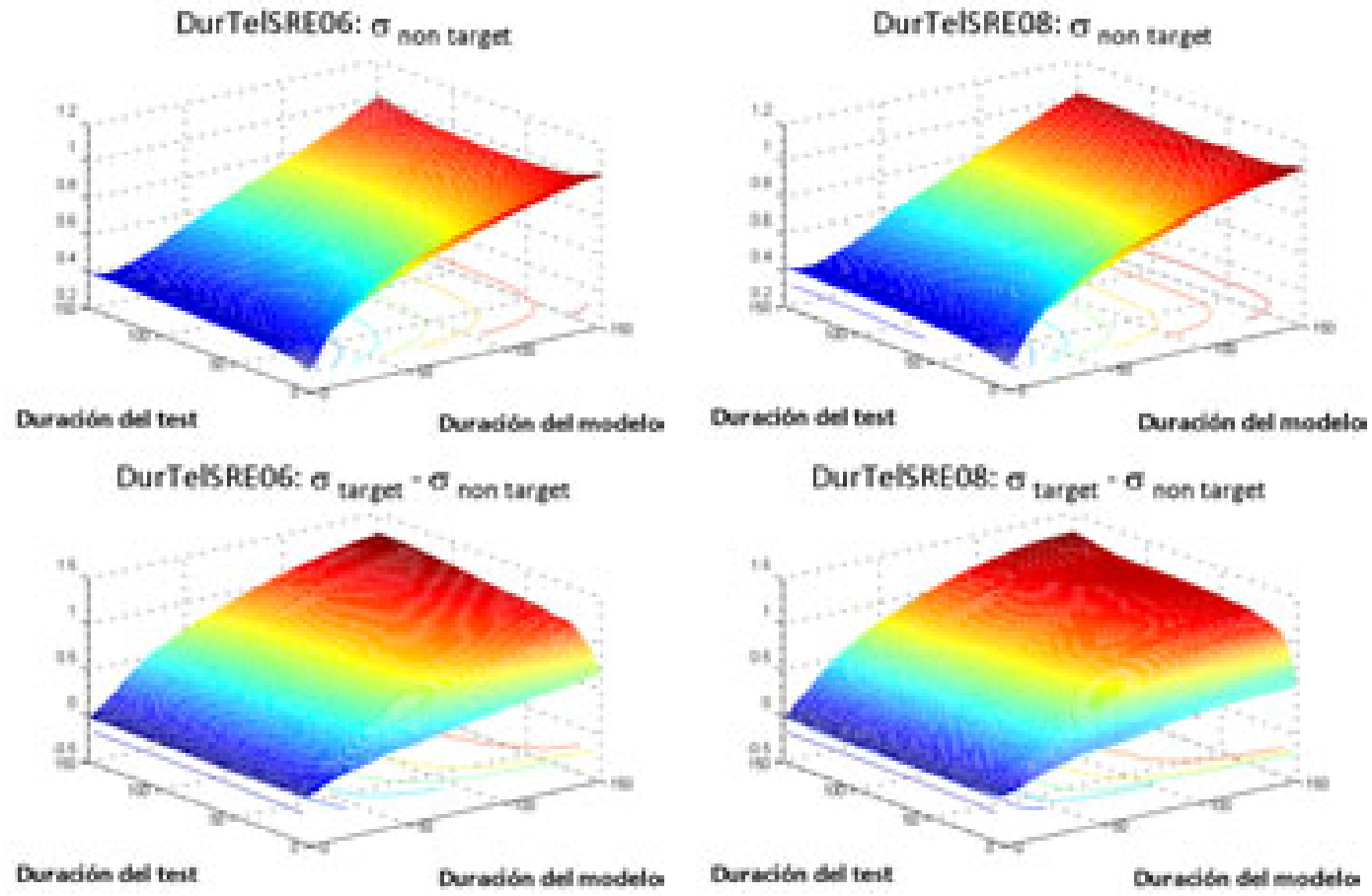


Figura 6.9. Para las bases de datos *DurTelSRE06* (columna izquierda) y *DurTelSRE08* (en la derecha): Media *target* y *non target* y su diferencia, desviación típica target y non target y su diferencia.

Para concluir este capítulo, se muestra efecto final del desalineamiento en forma de curva DET y EER en la **Figura** 6.10, donde se ha evaluado el rendimiento del sistema ATVS utilizado mediante las bases de datos originales de NIST SRE de 2006 y 2008 y las conformadas con variabilidad en duración para este estudio, *DurTelSRE06* y *DurTelSRE08*. La consecuencia de este desalineamiento se traduce en la obtención de un rendimiento entre tres y cuatro veces inferior al original, hecho que de nuevo motiva la implementación de algoritmos para reducir este efecto (apartado 8).

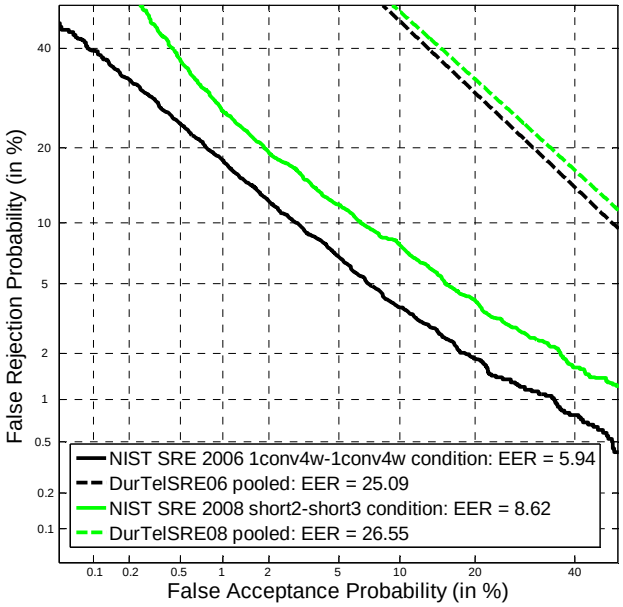


Figura 6.10. Rendimiento del sistema utilizado mediante las bases de datos (parte telefónica) de NIST SRE 2006 y 2008, DurTelSRE06 y DurTelSRE08.

7 ANÁLISIS DEL IMPACTO DE LA CALIDAD

7.1 DESCRIPCIÓN GENERAL DEL PROBLEMA

La idea de que la calidad de una muestra de voz puede afectar al rendimiento de un sistema de verificación automática de locutor es bastante intuitiva (52). De hecho la medida y compensación de la calidad de una señal de audio ha sido una tarea en el que se ha invertido un gran esfuerzo en el ámbito científico biométrico en los últimos años (52). Inicialmente este esfuerzo viene por la necesidad de controlar la calidad de la voz en las redes telefónicas pero en la actualidad se ha transformado en la definición de medidas de calidad y algoritmos de calibración que permitan predecir el rendimiento de un sistema de reconocimiento biométrico.

La calidad de voz está basada en el conocimiento de la señal mediante la cual se puede predecir tanto el rendimiento de un sistema de verificación de locutor como un posible desalineamiento de las puntuaciones del mismo debido a cambios en dicho indicador de degradación. Es por ello que, al igual que con la variabilidad en duración, forma parte de este proyecto el estudio de distintos tipos de calidad. Las medidas de calidad estudiadas son: la KLPC (*Kurtosis of Linear Prediction Coefficients*), la KCEP (*Kurtosis Cepstral*), la UBML (*Universal Background Model Likelihood*), la SNR (*Signal to Noise Ratio*) y la recomendación P.563 de la ITU (*International Telecommunication Union*)[4].

7.2 MEDIDAS DE CALIDAD EMPLEADAS

7.2.1 DEFINICIÓN DE LAS MEDIDAS DE CALIDAD

La calidad de una muestra biométrica viene definida por tres criterios básicos según (53), un borrador estándar de calidad propuesto por NIST sobre dicho tipo de muestras:

- La **fidelidad**, que se refiere a la exactitud y precisión con la que una muestra biométrica es capturada, procesada y almacenada en el sistema.
- El **carácter**, entendido como la actitud o predisposición del usuario a que se

[4] www.itu.int

capture su muestra biométrica (factores conductuales).

- La **utilidad**, definida como la característica para evaluar y predecir el rendimiento de un sistema de reconocimiento biométrico.

Si bien es interesante estudiar estos tres criterios, que definen de forma concisa la calidad relativa de una muestra como predictor del rendimiento de un sistema, este trabajo pretende evaluar mediante la *utilidad* (54) el rendimiento global del sistema bajo diferentes condiciones de calidad. Posteriormente se utilizará esta información de calidad para mejorar la discriminación del sistema mediante diferentes algoritmos detallados en el capítulo 8.

Es importante destacar que las medidas de calidad bajo estudio han sido seleccionadas por su fuerte relación con el rendimiento del sistema como se ha demostrado en estudios anteriores (7) y por comportarse de manera consistente en cuanto a bases de datos y sistemas utilizados, es decir, el rendimiento en función de la calidad elegida presenta la misma tendencia para diferentes bases de datos y sistemas. Otro criterio de elección es el nivel de correlación entre los indicadores de degradación propuestos (7), que indican cuanto de complementaria es la información que aporta cada medida de calidad para después combinarlas.

RELACIÓN SEÑAL A RUIDO (SNR)

Como su nombre indica, la SNR expresa la relación entre la potencia de la señal de voz y la potencia de ruido que la corrompe. Por lo tanto, queda definida mediante la siguiente fórmula:

$$SNR = 10 \cdot \log\left(\frac{E_{voz}}{E_{silencio}}\right)$$

siendo E_{voz} y $E_{silencio}$ la energía media de las zonas de voz y silencio del fragmento de audio respectivamente.

Aunque existen varias maneras de estimar la SNR, en este proyecto se ha calculado mediante un detector de actividad de voz (VAD), cuyo principal problema es que la fiabilidad de la calidad dependerá de la precisión de este sistema siendo una no muy buena referencia como indicador de degradación si el diseño de éste no es el apropiado. No obstante es un indicador de degradación ampliamente extendido y utilizado.

Siguiendo las recomendaciones de (54) y (55), para trabajar de forma homogénea con cualquier tipo de calidad, es necesario expresar todo indicador de degradación entre 0 y 1 mediante una función de mapeo $Q(x)$, siendo 0 el valor mínimo de calidad y 1 el máximo.

La función de mapeo usada en este caso es la siguiente:

$$Q_{SNR}(x) = \frac{x}{60}$$

donde x corresponde al valor de SNR obtenido en un rango de $(0 - 60)dB$ (56).

SIMILITUD A UN MODELO DE HABLA UNIVERSAL (UBML)

La UBML es una medida de calidad que trata de aproximar la similitud de una locución al modelo de habla universal utilizado para la generación del modelo estadístico de un locutor. Es una medida, considerada de forma reciente en (56), se extrae de forma obligatoria, sin suponer un coste computacional adicional, en los sistemas GMM para calcular la puntuación de similitud:

$$S(O, \lambda_t) = \log\big(p(O, \lambda_t)\big) - \log\big(p(O, \lambda_{UBM})\big)$$

donde $S(O, \lambda_t)$ es el *score* de similitud entre el modelo del locutor y el universal y $p(O, \lambda_t)$ y $p(O, \lambda_{UBM})$ son las funciones densidad de probabilidad del modelo de usuario y universal respectivamente. Por lo tanto, la UBML queda definida de la siguiente forma:

$$UBML = \log\big(p(O, \lambda_{UBM})\big)$$

Su función de mapeo puede definirse como:

$$Q_{UBML}(x) = \frac{x + 13}{8}$$

donde x corresponde al valor de la UBML obtenido perteneciente al rango de $(-13, -5)$ estimado de forma experimental (56).

RECOMENDACIÓN ITU P.563 (P.563)

Es un método de evaluación objetivo de calidad subjetiva definido por la ITU (52). Esta recomendación surgió por la necesidad de monitorizar la calidad de la señal de voz en las redes telefónicas y en la actualidad se utiliza como indicador de calidad en algunos sistemas biométricos. Esta medida tiene en cuenta la mayoría de factores degradantes de la señal de voz en las redes de comunicaciones actuales tales como el ruido, los ecos, la diafonía y la interferencia. Este algoritmo, de gran complejidad, calcula 51 parámetros de la señal de voz a partir de los cuales estima el factor de degradación dominante (hace uso de indicadores como la KLPC, la KCEP o la SNR), para después calcular una puntuación sobre la escala MOS (*Mean Opinion Score*) que será representativa de la calidad subjetiva que un oyente daría en media a la señal. Esta escala está definida entre 1 y 5, donde 1

corresponde a la peor calidad y 5 a la mejor. Su función de mapeo $Q(x)$ se define a continuación (56):

$$Q_{P.563}(x) = \frac{x-1}{4}$$

KURTOSIS DE LOS COEFICIENTES DE PREDICCIÓN LINEAL (KLPC)

La *kurtosis* es una medida estadística que mide el parecido de una distribución dada a una distribución gaussiana a través de la energía de sus colas y su simetría (57). Esta medida, aplicada en este caso a los P coeficientes LPC, se calcula como el momento estadístico de cuarto orden de la distribución a estudiar:

$$k = \frac{1}{p} \cdot \sum_{p=1}^{P} \left(\frac{a_p - \frac{1}{p} \cdot \sum_{p=1}^{P} a_p}{\sigma} \right)^4 - 3$$

siendo a_p los valores de la distribución de los P coeficientes y σ su desviación típica. Por lo tanto y dado que la función está normalizada, una distribución gaussiana tendrá *kurtosis* 0. En el caso de la *kurtosis* LPC se evalúa el parecido de la distribución de estos coeficientes a una gaussiana. Para este estudio el número de coeficientes utilizados ha sido de 21, obtenidos de cada trama de voz de 20 ms una vez suprimidos los silencios.

De nuevo, es pertinente definir una función de mapeo que represente de manera general la degradación de la muestra de voz siendo 0 su valor de calidad peor y 1 el mejor:

$$Q_{KLPC}(x) = 1 - \frac{x-3}{8}$$

donde x corresponde con el valor de *kurtsosis* de los coeficientes LPC de la muestra de voz (56).

KURTOSIS CEPSTRAL (KCEP)

La *kurtosis cepstral* mide el parecido entre la distribución de los coeficientes MFCC (definidos en el dominio *cepstral*) y una distribución gaussiana de igual forma que se hacía para los coeficientes LPC. Su función de mapeo $Q(x)$ se define a continuación:

$$Q_{KCEP}(x) = \frac{x-10}{6}$$

De nuevo, el factor x hace referencia el indicador de degradación antes de ser transformado (56).

7.2.2 CONSIDERACIONES A TENER EN CUENTA

En este apartado se detallan algunas consideraciones a tener en cuenta extraídas de (7) que serán de gran utilidad a la hora de compensar la variabilidad de las locuciones de voz mediante las técnicas descritas en el capítulo 8 y extraer conclusiones:

- La correlación entre los identificadores de degradación P.563 y SNR es baja. Este hecho se debe a que las bases de datos de experimentación (NIST SRE 2006 y 2008) presentan una SNR alta que hace que el factor dominante en la P.563 sea distinto de este factor de calidad. Por lo tanto, combinar la P.563 con la SNR puede aportar mejoras en el sistema.

- La P.563 mantiene una correlación moderada con la KLPC y la KCEP, ya que se basa en coeficientes del mismo espacio de características para indicar la calidad de una muestra.

- La UBML mantiene una correlación alta con la SNR, por lo que descarta a la SNR para combinarla a la hora de aplicar métodos de compensación. Esto es debido a que la UBML presenta una alta sensibilidad frente al nivel de ruido ya que el UBM frente al cual fue calculada se entrenó con un nivel poco representativo de éste.

- Entre las calidades estudiadas la UBML presenta una mayor utilidad en cuanto a predicción del rendimiento del sistema.

- En condiciones en las cuales el modelo de entrenamiento es de procedencia telefónica y el fichero de test de procedencia microfónica la KCEP presenta utilidad como predictor del rendimiento.

- La P.563 está diseñada para medir la calidad de la voz transmitida por la red telefónica, por lo que considerarla para medir la degradación de señales de voz adquiridas de otro canal puede no ser útil (datos microfónicos).

- La SNR, aun siendo un buen indicador de degradación para datos telefónicos, se comporta mejor con datos de procedencia microfónica, ya que estos sí que tienen una gran variabilidad de esta medida de calidad en la base de datos utilizada, no siendo así para los datos telefónicos.

- La degradación de las señales se traduce en una pérdida de información biométrica (58). Es más probable que dicha degradación haga que dos locuciones de un mismo individuo parezcan diferentes a que se asemejen contribuyendo a un empeoramiento del rendimiento del sistema.

7.3 BASES DE DATOS CON VARIABILIDAD EN CALIDAD

Para evaluar el impacto de las medidas de calidad sobre el rendimiento de los sistemas se ha utilizado la base de datos de NIST SRE 2008. En este caso se ha utilizado tanto para el entrenamiento de los modelos como para los ficheros de test muestras de voz de procedencia telefónica y microfónica (ver apartado 5.1).

A continuación se representa mediante tablas y diagramas de barras el número de archivos (modelos y tests) y enfrentamientos (*target* y *non target*) de las bases de datos utilizadas dependientes de las medidas de calidad presentadas en la sección 7.2. Destacar que para la representación, se han agrupado los subconjuntos de *scores* en función del valor de su indicador de degradación (que toma valores de 0 a 1) definiéndolo en cuatro posibles intervalos o *bins*: de 0 a 0.25 para el primero, de 0.25 a 0.50 para el segundo, de 0.50 a 0.75 para el tercero y de 0.75 a 1 para el cuarto y último.

TT	KLPC		KCEP		UBML		SNR		P.563	
Bins	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**
1º	11	148	232	3117	22	345	15	254	570	7981
2º	635	8449	875	12060	2103	26799	658	8748	2765	35776
3º	2709	35505	1679	22218	1488	20232	2538	32929	278	3619
4º	258	3274	827	9981	0	0	402	5445	0	0

Tabla 7.1. Número de enfrentamientos para la base de datos de NIST SRE 2006: condición *tel-tel*.

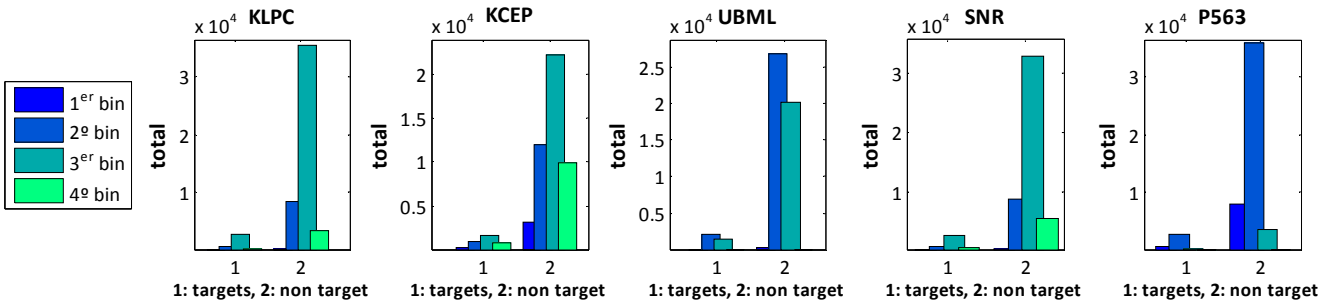


Figura 7.1. Diagrama de barras: enfrentamientos para la base de datos de NIST SRE 2006: condición *tel-tel*.

TT	KLPC		KCEP		UBML		SNR		P.563	
Bins	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**
1º	7	91	260	2130	60	582	70	729	841	7743
2º	664	5491	1037	8958	2445	21092	1500	13346	2760	23648
3º	2727	24074	1580	14207	1327	11544	1976	16449	230	1810
4º	434	3562	955	7923	0	0	286	2694	1	17

Tabla 7.2. Número de enfrentamientos para la base de datos de NIST SRE 2008: condición *tel-tel*.

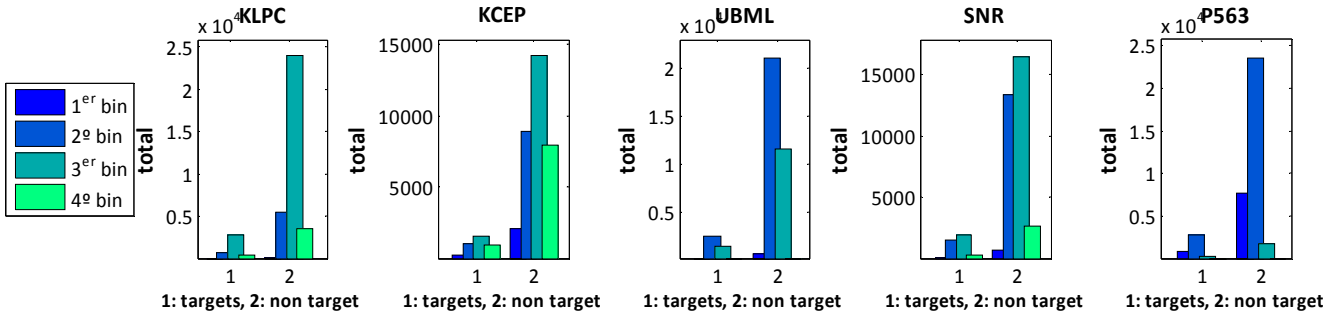


Figura 7.2. Diagrama de barras: enfrentamientos para la base de datos de NIST SRE 2008: condición *tel-tel*.

TM	KLPC		KCEP		UBML		SNR		P.563	
Bins	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**
1º	41	112	0	0	34	152	0	0	988	3043
2º	737	2182	16	32	2589	7303	541	1349	2914	8581
3º	2797	8385	2297	6728	1346	4347	2948	8706	67	178
4º	394	1123	1656	5042	0	0	480	1747	0	0

Tabla 7.3. Número de enfrentamientos para la base de datos de NIST SRE 2008: condición *tel-mic*.

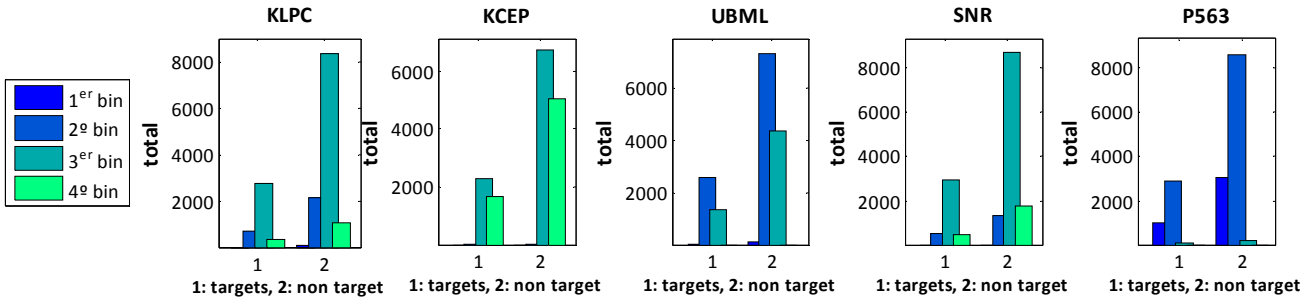


Figura 7.3. Diagrama de barras: enfrentamientos para la base de datos de NIST SRE 2008: condición *tel-mic*.

MT	KLPC		KCEP		UBML		SNR		P.563	
Bins	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**
1º	1	43	158	2003	278	2413	259	2438	578	5798
2º	359	3299	173	1477	815	8107	824	8138	485	4441
3º	600	6081	340	2991	12	116	2	32	40	382
4º	145	1213	434	4165	0	0	20	28	2	15

Tabla 7.4. Número de enfrentamientos para la base de datos de NIST SRE 2008: condición *mic-tel*.

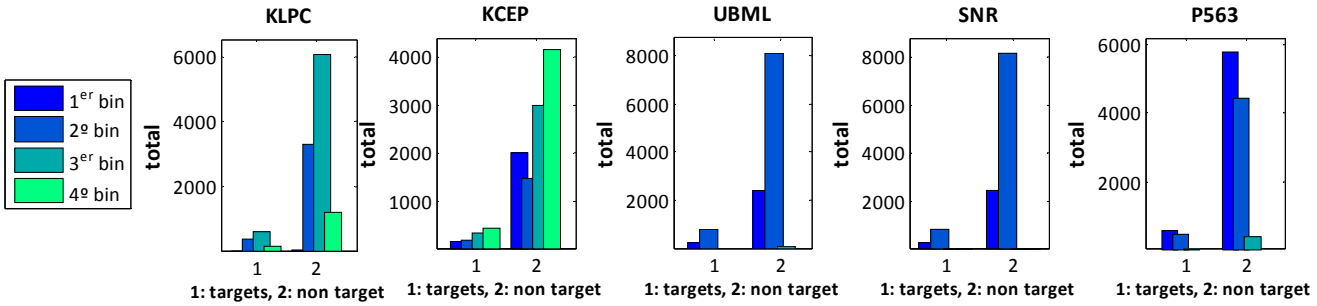


Figura 7.4 Diagrama de barras: enfrentamientos para la base de datos de NIST SRE 2008: condición *mic-tel*.

MM	KLPC		KCEP		UBML		SNR		P.563	
Bins	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**	**Target**	**Non target**
1º	55	104	2111	4129	2517	4821	2714	5321	6270	12323
2º	3646	6997	1643	3308	8872	17440	8747	17096	4803	9306
3º	6518	12831	3295	6324	136	260	32	104	430	864
4º	1306	2589	4476	8760	0	0	32	0	22	28

Tabla 7.5. Número de enfrentamientos para la base de datos de NIST SRE 2008: condición *mic-mic*.

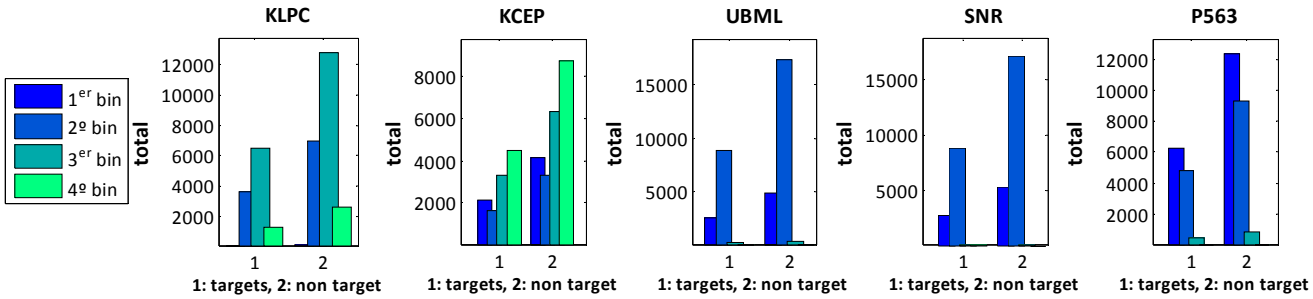


Figura 7.5. Diagrama de barras: enfrentamientos para la base de datos de NIST SRE 2008: condición *mic-mic*.

TT	KLPC		KCEP		UBML		SNR		P.563	
Bins	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**
1º	159	205	3349	3335	367	855	269	673	8551	11657
2º	9084	8255	12935	11607	28902	31081	9406	11485	38541	35948
3º	38214	35626	23897	22005	21720	19053	35467	33495	3897	3384
4º	3532	6903	10808	14042	0	0	5847	5336	0	0

Tabla 7.6. Número de archivos de la base de datos de NIST SRE 2006: condición *tel-tel*.

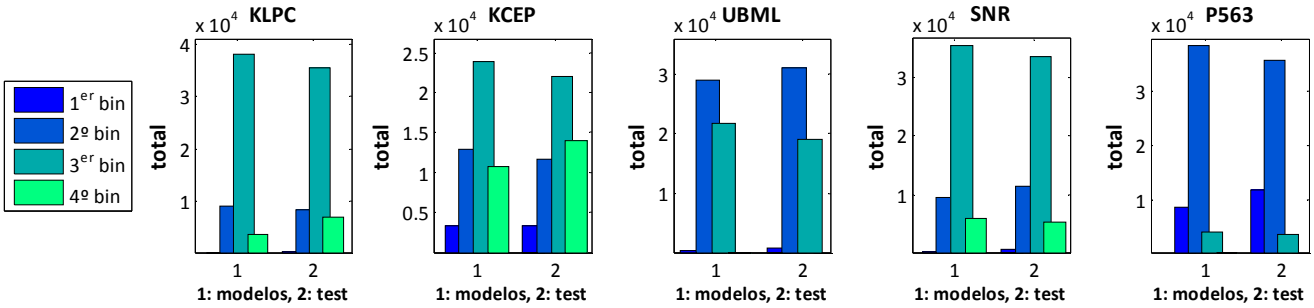


Figura 7.6. Diagrama de barras: archivos de la base de datos de NIST SRE 2006: condición *tel-tel*.

TT	KLPC		KCEP		UBML		SNR		P.563	
Bins	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**
1º	98	40	2390	2288	642	1000	799	10	8584	8885
2º	6155	5392	9995	8871	23537	24202	14846	7773	26408	26255
3º	26801	26591	15787	16350	12871	11848	18425	26322	2040	1910
4º	3996	5027	8878	9541	0	0	2980	2945	18	0

Tabla 7.7. Número de archivos de la base de datos de NIST SRE 2008: condición *tel-tel*.

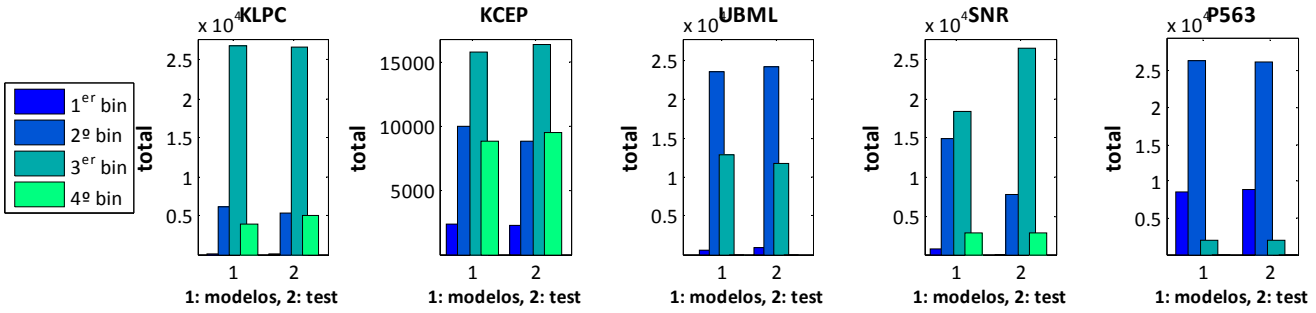


Figura 7.7. Diagrama de barras: archivos de la base de datos de NIST SRE 2008: condición *tel-tel*.

TM	KLPC		KCEP		UBML		SNR		P.563	
Bins	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**
1º	153	559	0	2048	186	3725	0	2252	4031	11194
2º	2919	7933	48	2766	9892	11831	1890	12829	11495	4225
3º	11182	6894	9025	8712	5693	215	11654	690	245	341
4º	1517	385	6698	2245	0	0	2227	0	0	11

Tabla 7.8. Número de archivos de la base de datos de NIST SRE 2008: condición *tel-mic*.

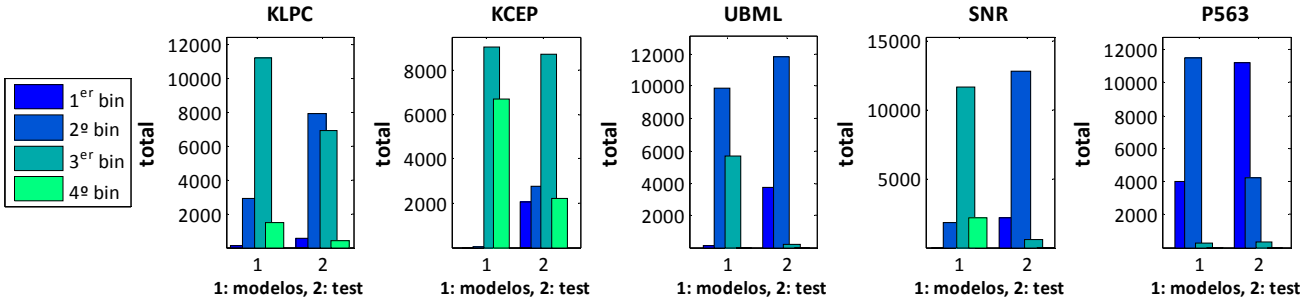


Figura 7.8. Diagrama de barras: archivos de la base de datos de NIST SRE 2008: condición *tel-mic*.

MT	KLPC		KCEP		UBML		SNR		P.563	
Bins	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**
1º	44	408	2161	0	2691	35	2697	0	6376	3183
2º	3658	1336	1650	0	8922	7432	8962	2866	4926	8050
3º	6681	8281	3331	6590	128	4274	34	7705	422	508
4º	1358	1716	4599	5151	0	0	48	1170	17	0

Tabla 7.9. Número de archivos de la base de datos de NIST SRE 2008: condición *mic-tel*.

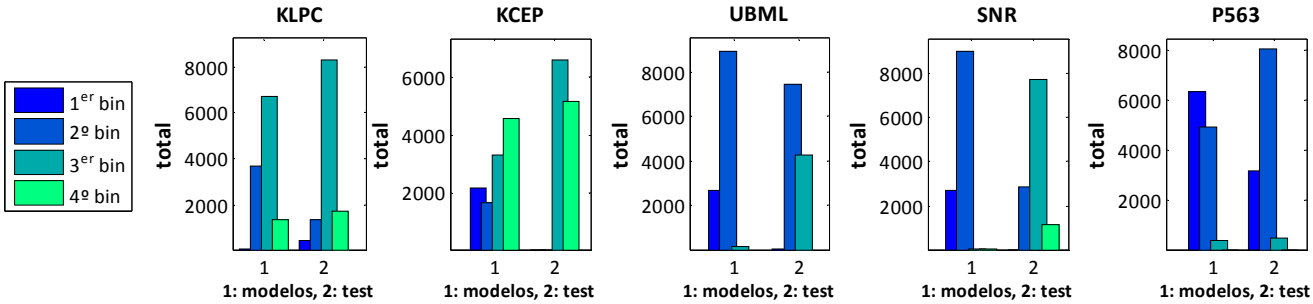


Figura 7.9. Diagrama de barras: archivos de la base de datos de NIST SRE 2008: condición *mic-tel*.

MM	KLPC		KCEP		UBML		SNR		P.563	
Bins	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**	**Modelos**	**Tests**
1º	159	1323	6240	199	7338	6383	8035	4248	18593	21923
2º	10643	15826	4951	4077	26312	27310	25843	29708	14109	10909
3º	19349	15294	9619	22849	396	353	136	90	1294	1209
4º	3895	1603	13236	6921	0	0	32	0	50	5

Tabla 7.10. Número de archivos de la base de datos de NIST SRE 2008: condición *mic-mic*.

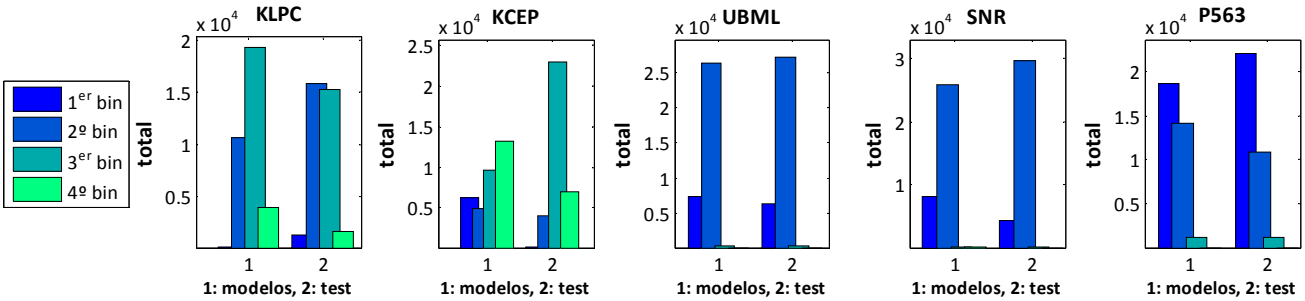


Figura 7.10. Diagrama de barras: archivos de la base de datos de NIST SRE 2008: condición *mic-mic*.

Observando las diferentes tablas y diagramas, y notando que en muchos casos la base de datos no presenta suficiente variabilidad en cuanto a valores de calidad extremos (las muestras, en general, presentan un valor entre 0.25 y 0.75 para todas las medidas de calidad no existiendo muestras para los intervalos de 0 a 0.25 y de 0.75 a 1), se decidió agrupar las puntuaciones por el valor de sus indicadores de degradación mediante el uso de cuartiles[5] (4 intervalos que contienen el 25% de los enfrentamientos cada uno ordenados por valor de calidad) para estudiar su relación con el rendimiento del sistema (ver apartado 7.4).

[5] Referido a los intervalos derivados de los valores de calidad equivalentes a los percentiles 25, 50 y 75.

7.4 IMPACTO DE LA VARIABILIDAD EN CALIDAD

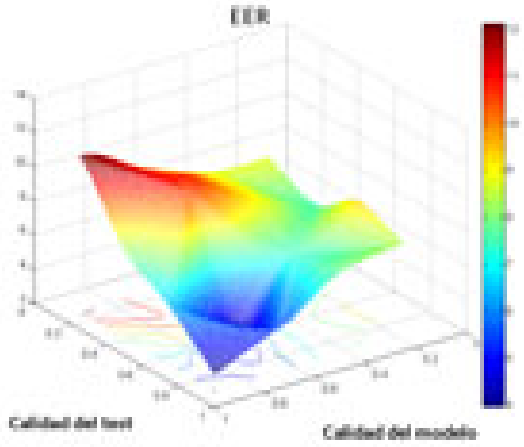
Con el objetivo de realizar un estudio minucioso del impacto de la variabilidad en calidad se han analizado varios tipos de figuras:

- **EER en 3D**: en función de la calidad del modelo de entrenamiento y del fichero de test.
- **Curvas DET Q_{test}:** sobre los subconjuntos de calidades agrupados por cuartil de calidad de test ignorando la calidad del modelo (considerando todas las muestras dependientes de la calidad del modelo de forma simultánea al igual que con variabilidad en duración, sección 6.2).
- **Curvas DET Q_{modelo}**: sobre los subconjuntos de calidades agrupados por cuartil de calidad de modelo ignorando la calidad del test (considerando todas las muestras dependientes de la calidad del test de forma simultánea al igual que con variabilidad en duración, sección 6.2).
- **Distribuciones Kernel Q_{test}**: sobre los subconjuntos de calidades agrupados por cuartil de calidad de test ignorando la calidad del modelo.
- **Distribuciones Kernel Q_{modelo}**: sobre los subconjuntos de calidades agrupados por cuartil de calidad de modelo ignorando la calidad del test.

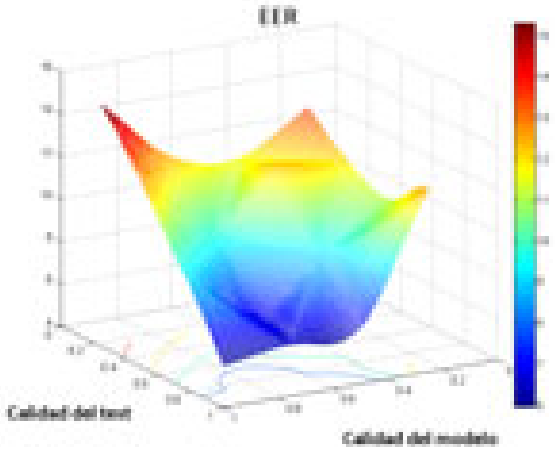
Antes de entrar en detalle en el análisis de cada calidad es importante remarcar que debido a la poca variabilidad existente en cuanto a calidad, las curvas interpoladas presentan un carácter ondulatorio, por lo que en más de una ocasión la tendencia de éstas no es del todo clara debido al efecto de la interpolación cúbica. De igual manera se puede comprobar que, al igual que con la variabilidad en duración, cuanto mayor es la calidad del modelo y del test mejor discriminación entre las distribuciones existe (EER más próximo a 0).

IMPACTO DE LA VARIABILIDAD EN UBML

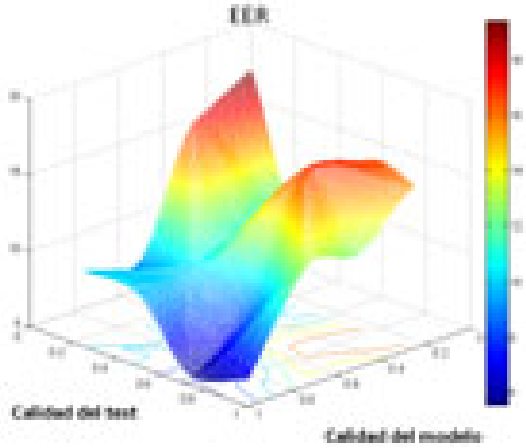
Para las distintas condiciones, al contrario que en duraciones, la calidad del test es más crítica que la del modelo. Posiblemente este hecho se explique observando los histogramas para la calidad UBML de la **Figura** 7.2 a la **Figura** 7.5, donde se indica que antes de realizar el mapeo a cuartiles la calidad está distribuida entre 0.25 y 0.75, no existiendo apenas muestras de calidad superior e inferior. Por lo tanto, al realizar subconjuntos de *scores* de 4 cuartiles estos se repartirán dicho rango de calidades siendo los grupos de indicadores similares, dando lugar a distribuciones no muy desalineadas. Aunque la tendencia presenta un cierto carácter ondulatorio se puede decir que es la medida de calidad estudiada junto con la SNR que mejor refleja el rendimiento del sistema (es decir, la más útil).



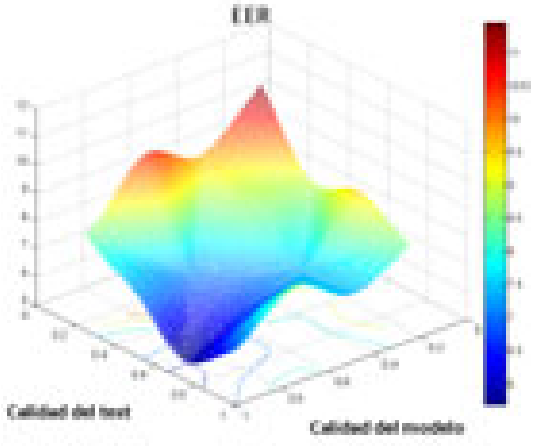
a) EER en función de la calidad del modelo y del test: condición *tel-tel*.



b) EER en función de la calidad del modelo y del test: condición *tel-mic.*



c) EER en función de la calidad del modelo y del test: condición *mic-tel.*



d) EER en función de la calidad del modelo y del test: condición *mic-mic.*

Figura 7.11. EER en función de la calidad UBML del modelo y del test para las 4 condiciones.

Para las curvas DET Q_{test} (ver **Figura** 7.12) existe una mayor diferencia entre el EER del primer y segundo cuartil (hasta 4.15 puntos de diferencia) que para las curvas DET Q_{modelo} (hasta 1.52 puntos de diferencia). Por lo tanto, este hecho apunta a que la calidad del test es más crítica que la del modelo. No obstante, observando la pronunciada separación de las curvas DET para todos los casos salvo para la condición *mic-mic* (con calidad de modelo y de test) se reitera el hecho de que esta medida de calidad es un buen indicador de degradación, por lo que su compensación debería ser mayor (capítulo 9.2).

Por otra parte y al igual que en duraciones, aunque esta vez de forma menos abrupta, se observa un desalineamiento siendo éste más acusado en las distribuciones *target* (ver **Figura 7.13** y la **Figura** 7.14), efecto que tratará de compensarse como los métodos descritos en el capítulo 8 para mejorar el rendimiento global del sistema.

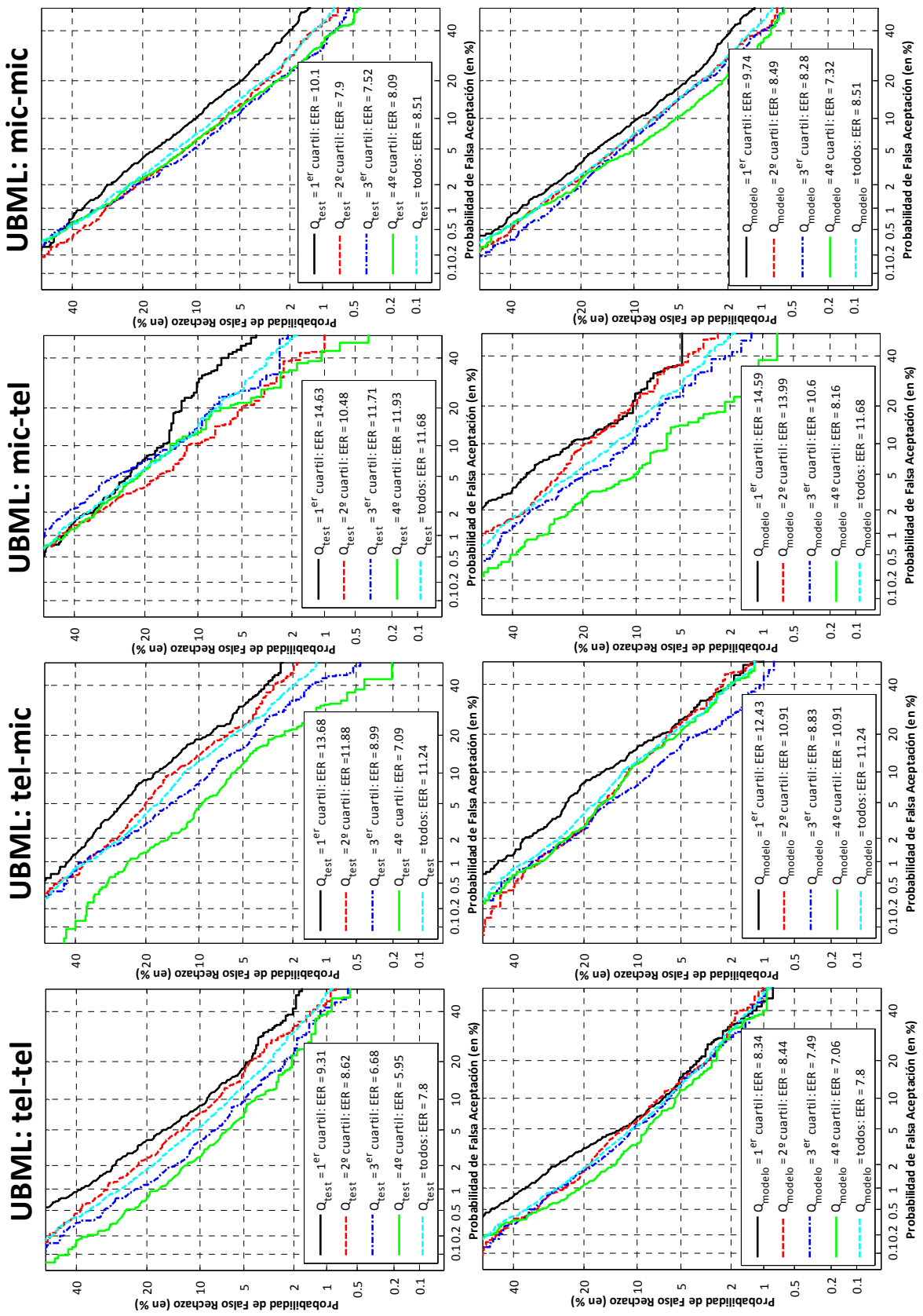


Figura 7.12. Curvas DETS para los subconjuntos dependientes de calidad UBML.

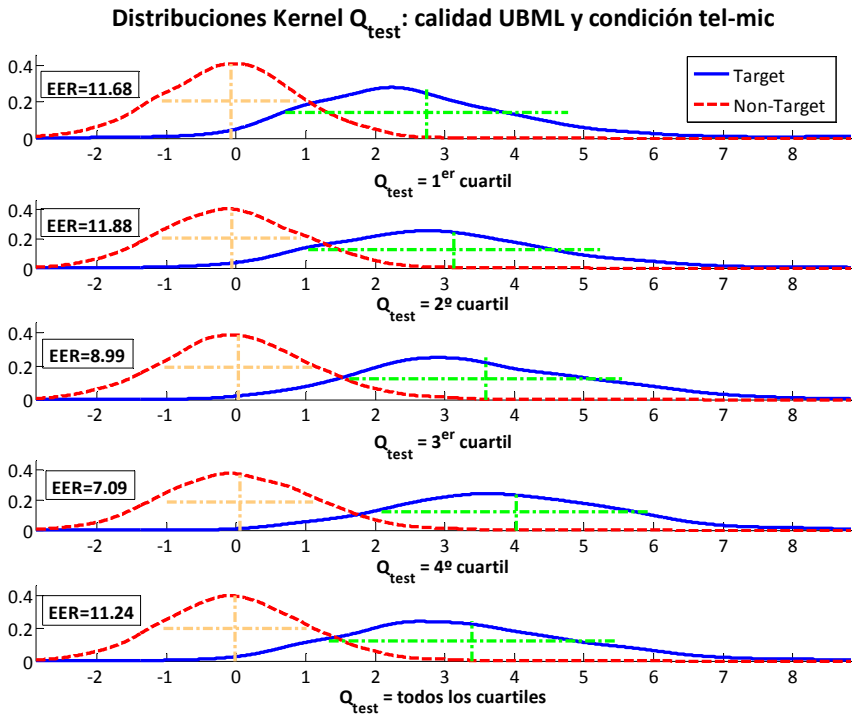


Figura 7.13. Desalineamiento para la Q_{test} UBML: condición *tel-mic*.

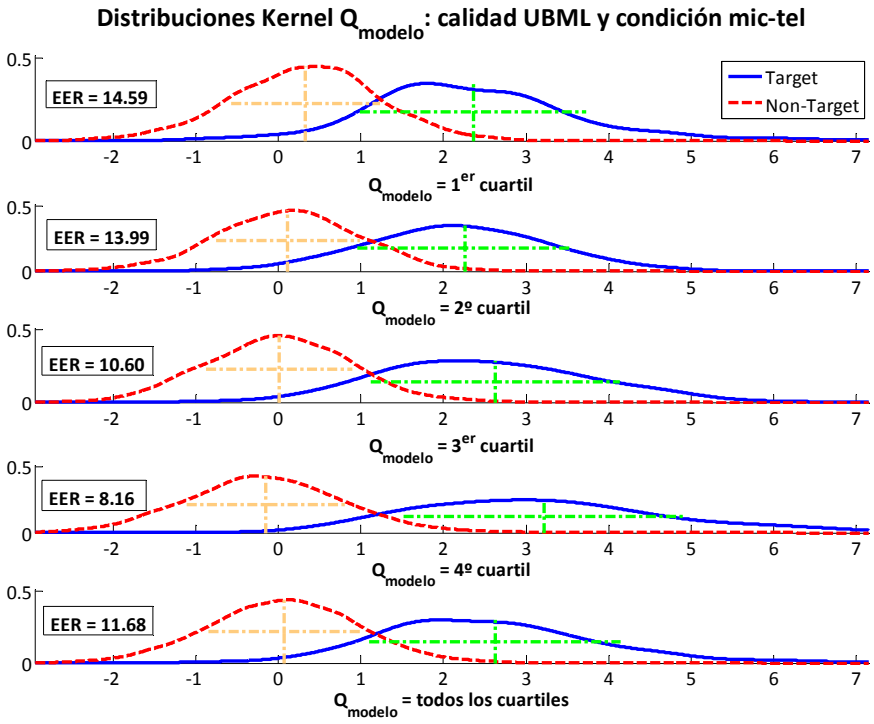
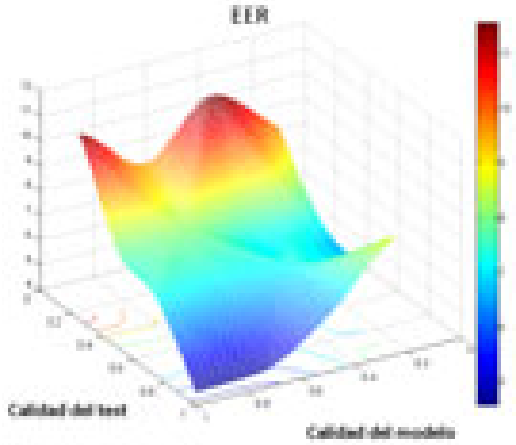


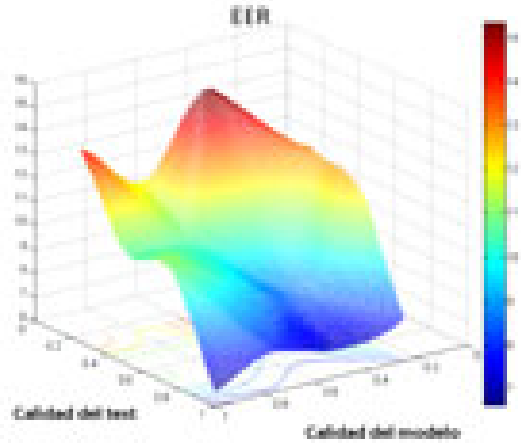
Figura 7.14. Desalineamiento para la Q_{modelo} UBML: condición *mic-tel*.

IMPACTO DE LA VARIABILIDAD EN SNR

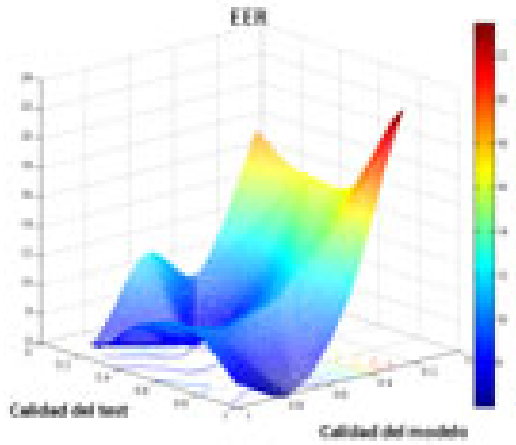
Al estar la SNR y la UBML correladas de forma notable, la tendencia del EER es similar siendo la SNR una buena medida de rendimiento del sistema. Como se comentó en el apartado 7.2.2, esta medida ofrece una buena estimación de la EER cuando se evalúa sobre muestras procedentes de canales microfónicos. Este efecto se puede observar de forma más pronunciada en la **Figura** 7.16, donde se aprecia cómo el cambio en el rendimiento es mayor cuando el modelo es microfónico (mayor separación entre las curvas DET), lo que viene a confirmar que la SNR es un buen predictor de rendimiento en este tipo de ensayos. Si se observa la **Figura 7.18** se notará de forma más intuitiva el efecto de la degradación del rendimiento en forma de alineamiento a compensar al igual que en el resto de indicadores o medidas (existen un cambio de18.95% al 11.78% al pasar del primer cuartil al segundo).



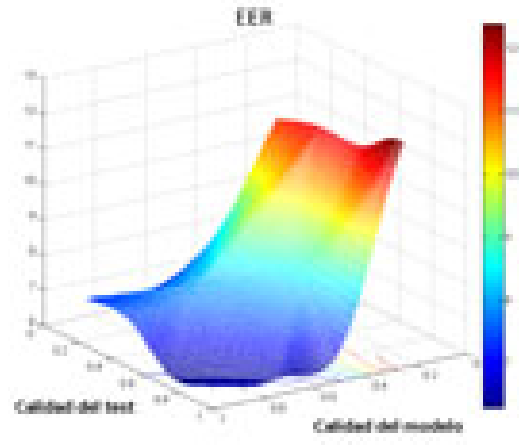
a) EER en función de la calidad del modelo y del test: condición *tel-tel.*



b) EER en función de la calidad del modelo y del test: condición *tel-mic.*



c) EER en función de la calidad del modelo y del test: condición *mic-tel.*



d) EER en función de la calidad del modelo y del test: condición *mic-mic.*

Figura 7.15. EER en función de la calidad SNR del modelo y del test para las 4 condiciones.

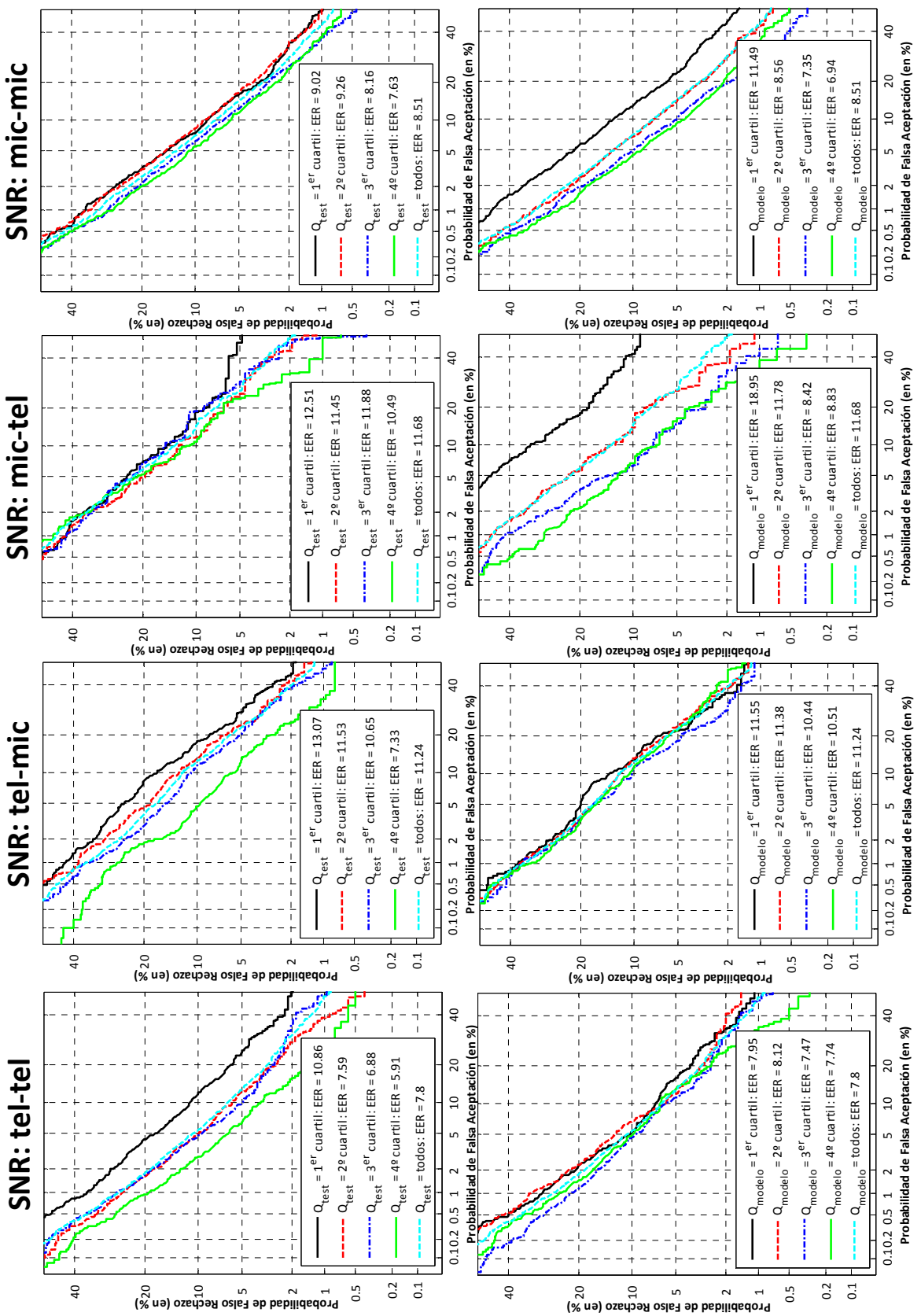


Figura 7.16. Curvas DETS para los subconjuntos dependientes de calidad SNR.

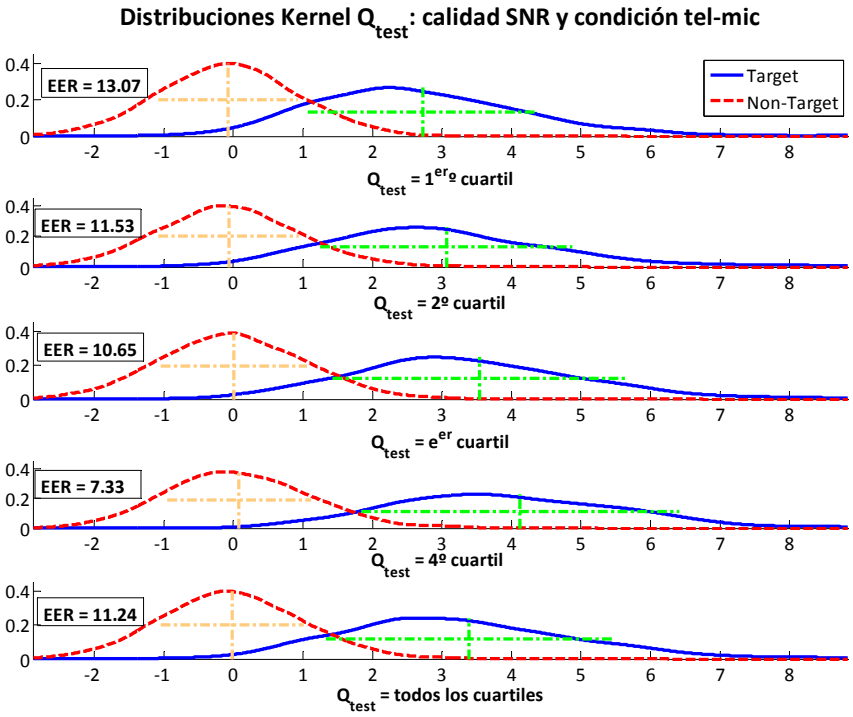


Figura 7.17. Desalineamiento para la Q_{test} SNR: condición *tel-mic*.

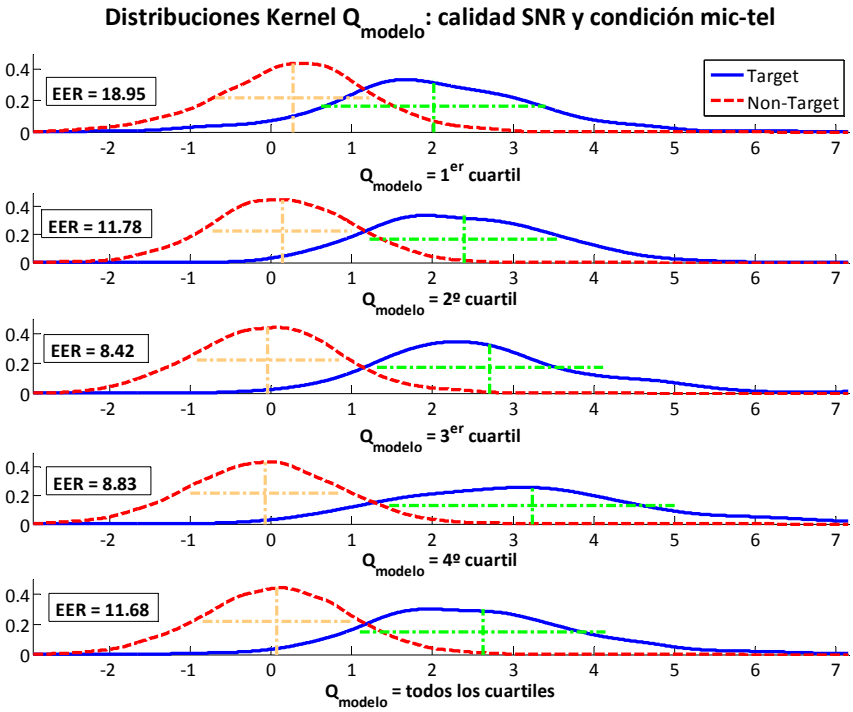
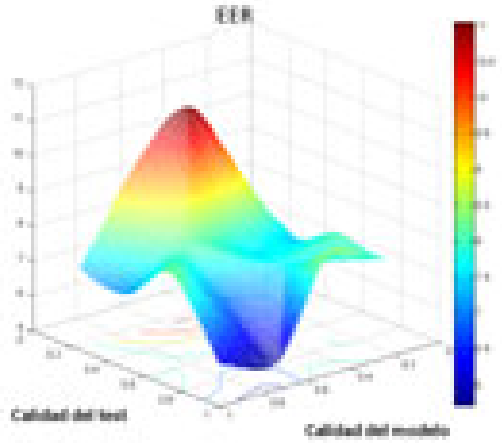


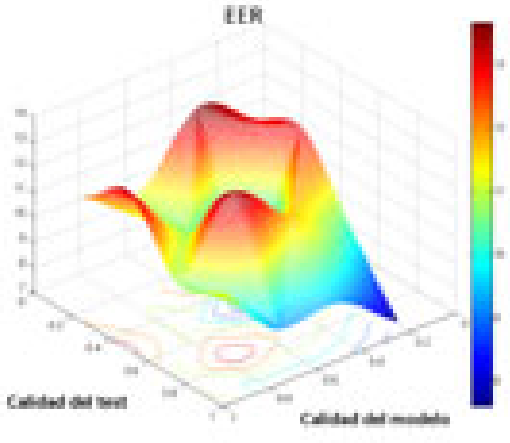
Figura 7.18. Desalineamiento para la Q_{modelo} SNR: condición *mic-tel*.

IMPACTO DE LA VARIABILIDAD EN P.563

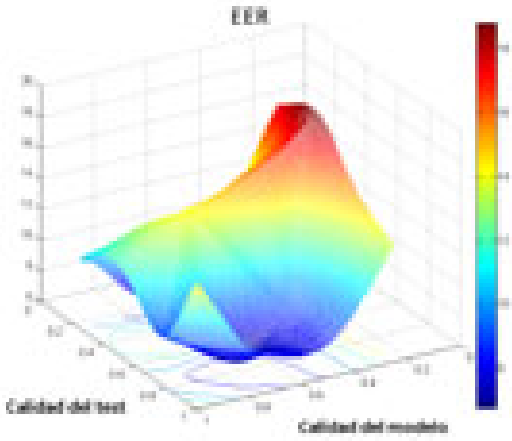
El caso de la P.563 es diferente: la tendencia clara de aumento de rendimiento con la calidad puede verse en la condición *tel-tel* (**Figura** 7.19, **Figura** 7.20 y **Figura** 7.21), donde de nuevo existe mayor diferencia entre las curvas que en los otros casos y cuyo rendimiento es mayor (EER del 6.29% en el mejor de los casos) que para las medidas de calidad ya analizadas (por encima del 7%). Sin embargo, el resto de condiciones presentan un carácter diferente al visto hasta ahora y es porque esta medida de calidad está diseñada por la ITU de forma específica para medir la calidad de las redes de comunicación definidas por unas perturbaciones diferentes a las que se pueden encontrar en un canal microfónico.



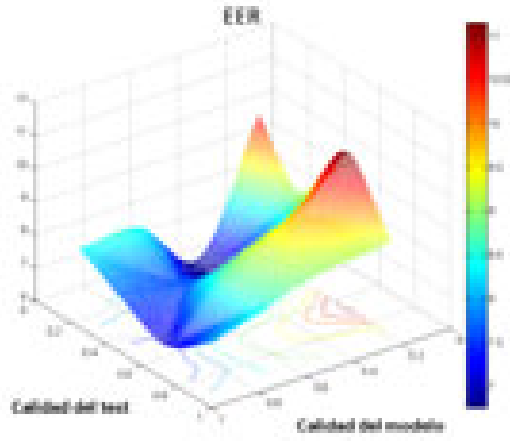
a) EER en función de la calidad del modelo y del test: condición *tel-tel.*



b) EER en función de la calidad del modelo y del test: condición *tel-mic.*



c) EER en función de la calidad del modelo y del test: condición *mic-tel.*



d) EER en función de la calidad del modelo y del test: condición *mic-mic.*

Figura 7.19. EER en función de la calidad P.563 del modelo y del test para las 4 condiciones.

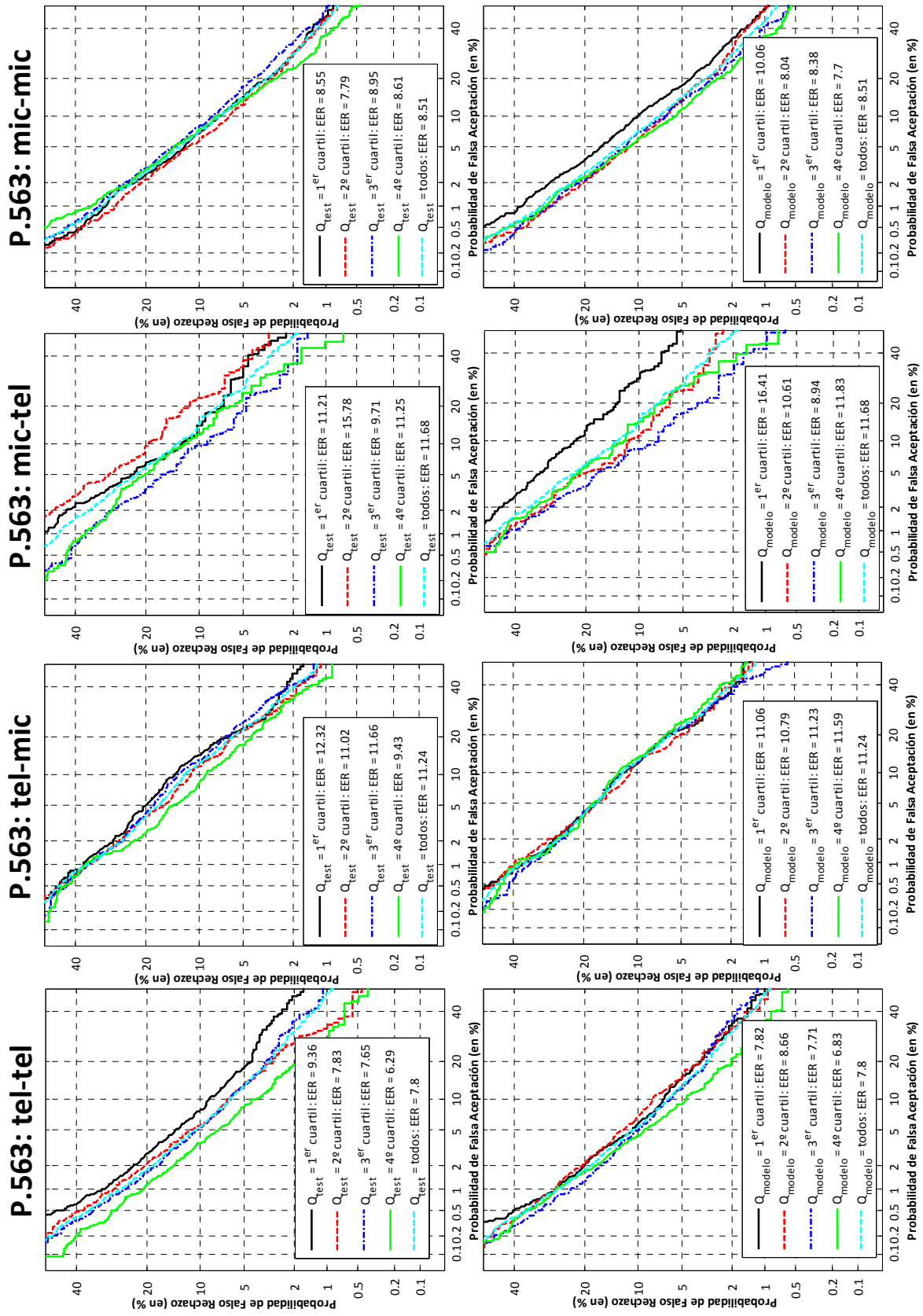


Figura 7.20. Curvas DETS para los subconjuntos dependientes de calidad P.563.

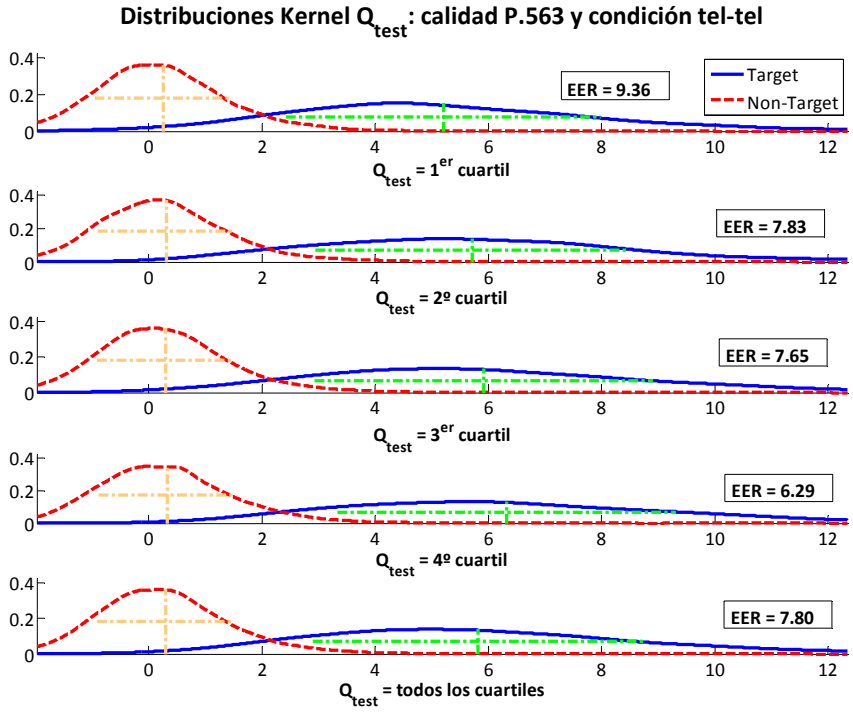


Figura 7.21. Desalineamiento para la Q_{test} P.563: condición *tel-tel.*

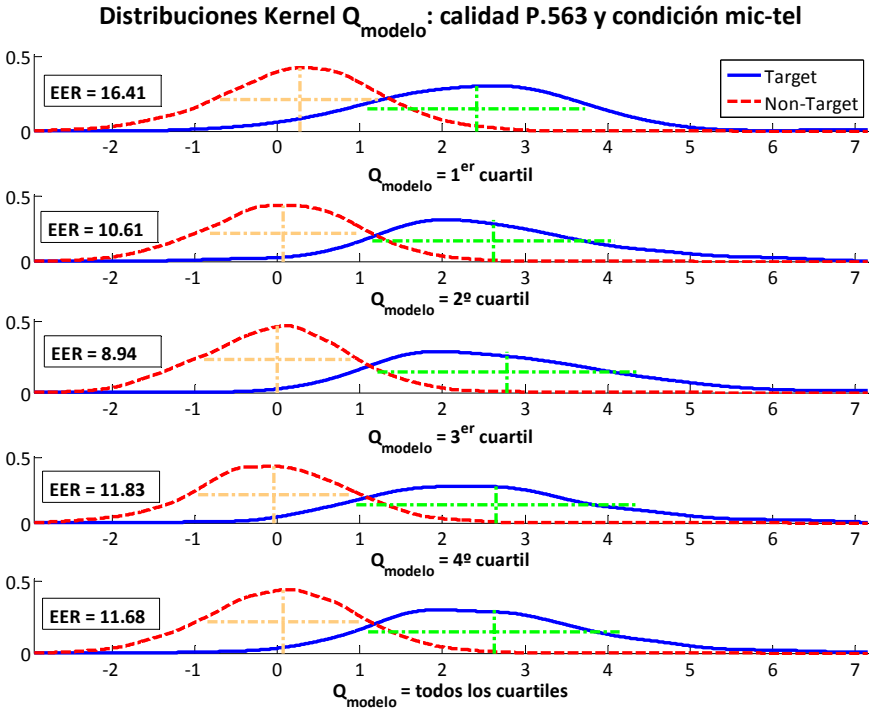
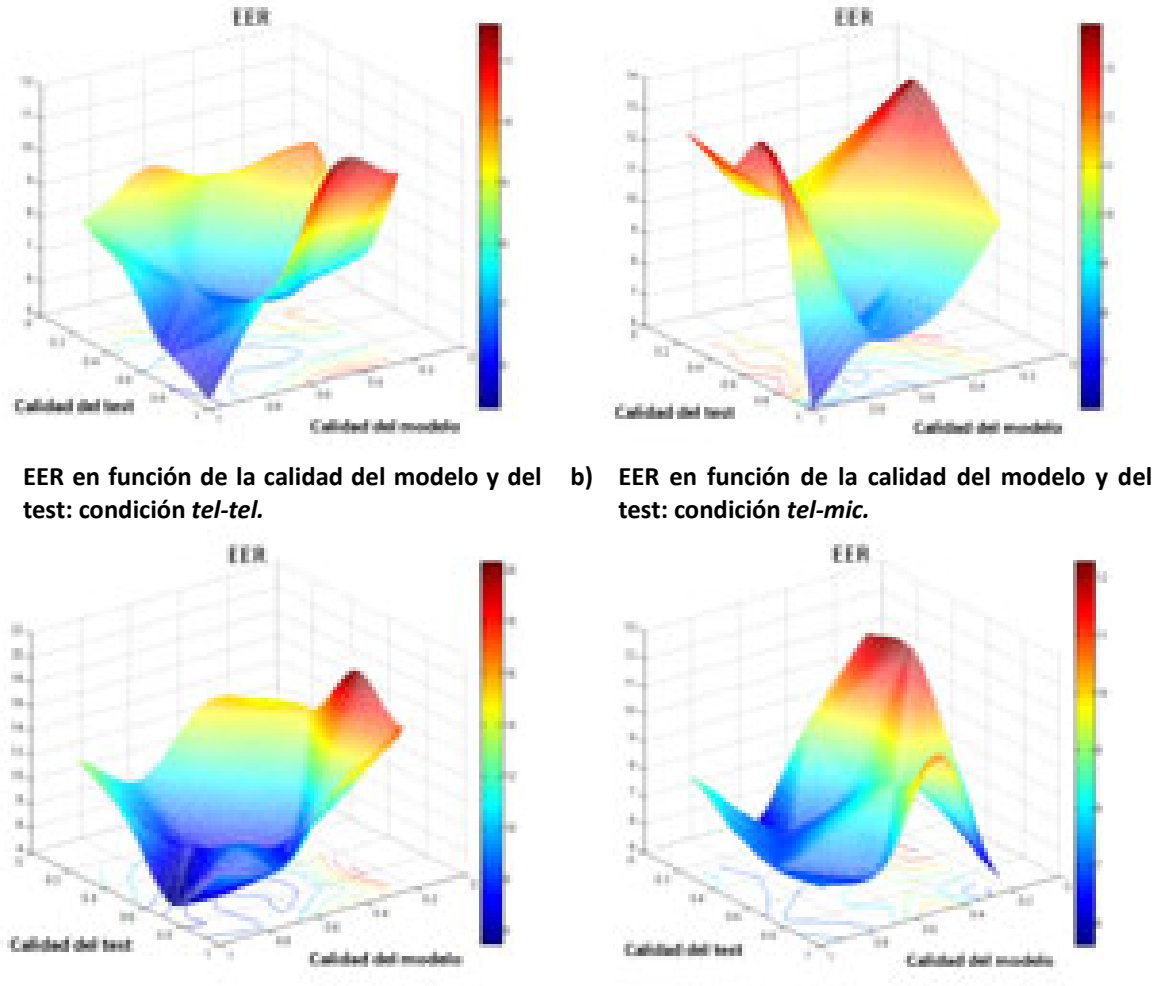


Figura 7.22. Desalineamiento para la Q_{modelo} P.563: condición *mic-tel*.

IMPACTO DE LA VARIABILIDAD EN KCEP

La *kurtosis cepstral* se presenta como un indicador de nivel de degradación moderado. Esta afirmación se fundamenta observando la no muy clara tendencia de las curvas en 3D de la figura **Figura** 7.23, debido en parte al efecto de la interpolación cúbica de los estadísticos usada para la representación pero no a la escasa variabilidad como ocurría con las medidas de calidad UBML y SNR (ver que existe una buena distribución en cuanto al número de archivos y enfrentamientos desde la **Figura** 7.1 a la **Figura** 7.10 y en sus respectivas tablas). Igualmente, al ser un indicador propuesto por (7) se estudiará la compensación de su rendimiento fruto del desalineamiento típico de la variabilidad (**Figura** 7.25 y **Figura 7.26**) y la dependencia con las bases de datos telefónicos y microfónicos.



a) EER en función de la calidad del modelo y del test: condición *tel-tel.*

b) EER en función de la calidad del modelo y del test: condición *tel-mic.*

c) EER en función de la calidad del modelo y del test: condición *mic-tel.*

d) EER en función de la calidad del modelo y del test: condición *mic-mic.*

Figura 7.23. EER en función de la calidad KCEP del modelo y del test para las 4 condiciones.

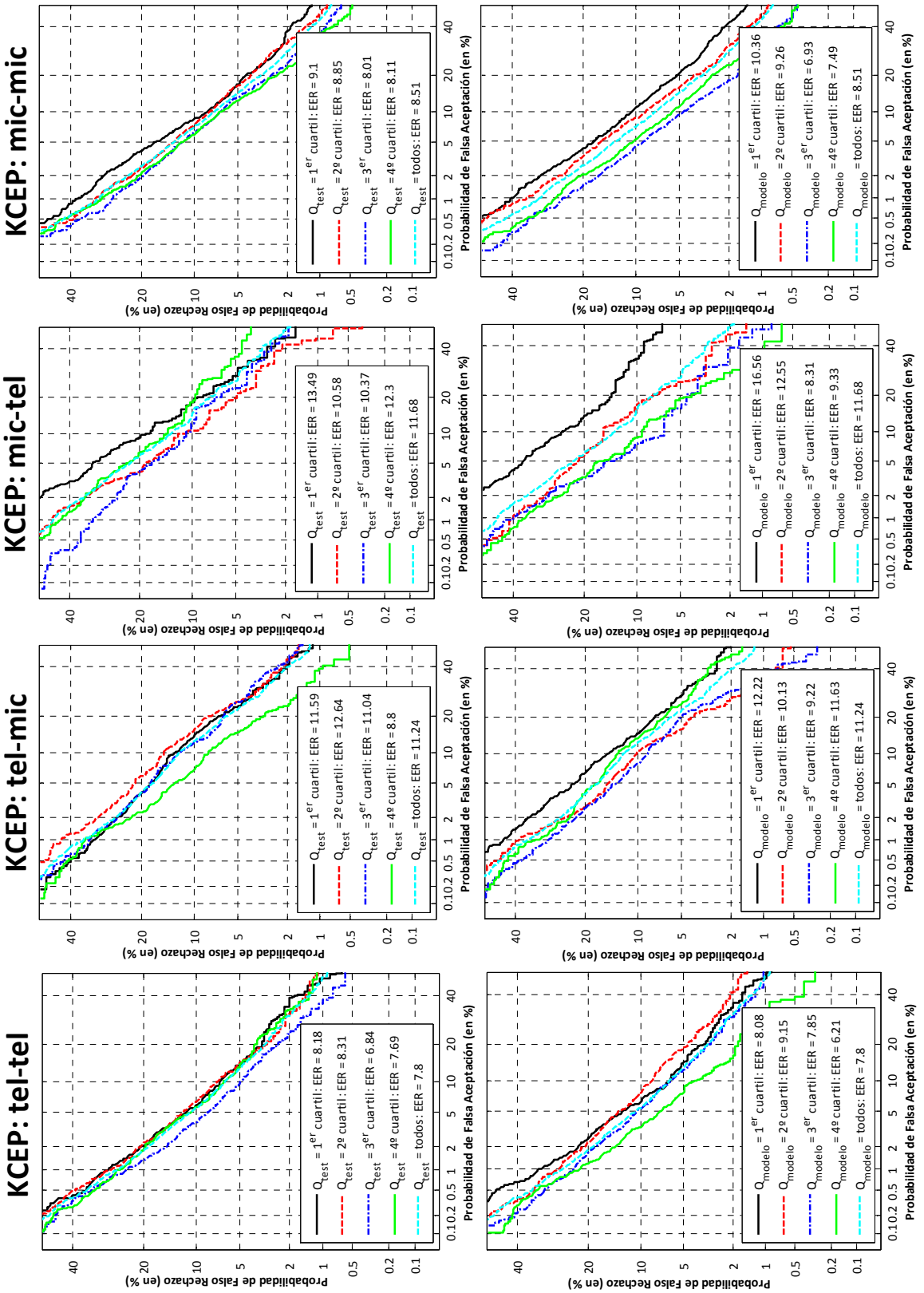


Figura 7.24. Curvas DETS para los subconjuntos dependientes de calidad KCEP.

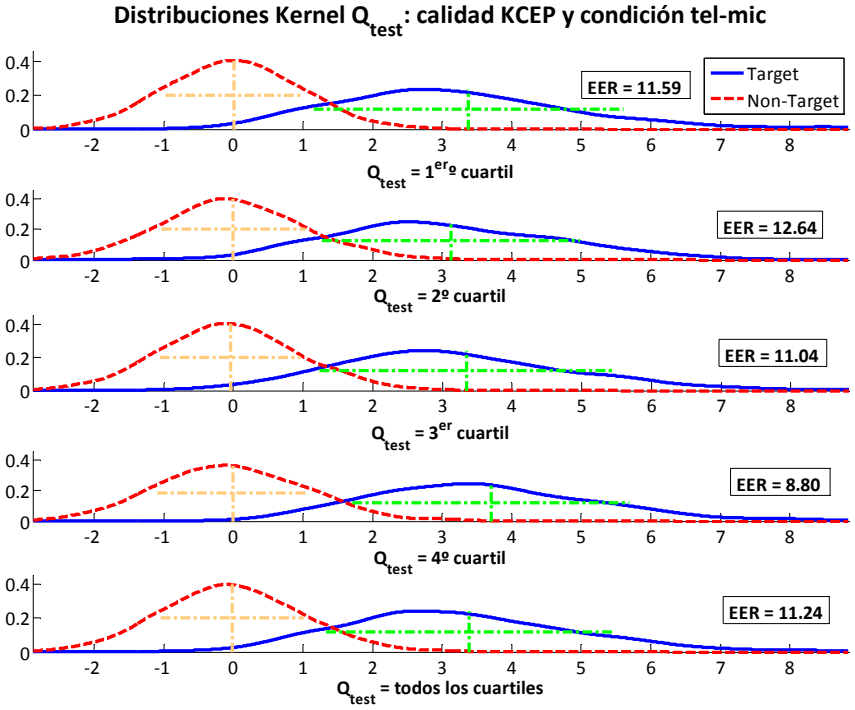


Figura 7.25. Desalineamiento para la Q_{test} KCEP: condición *tel-mic.*

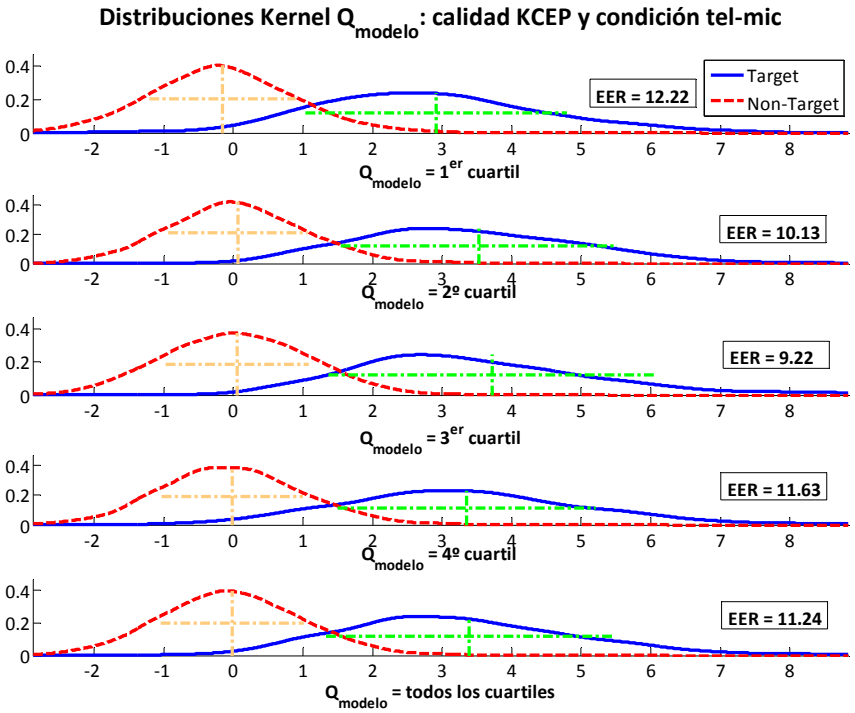
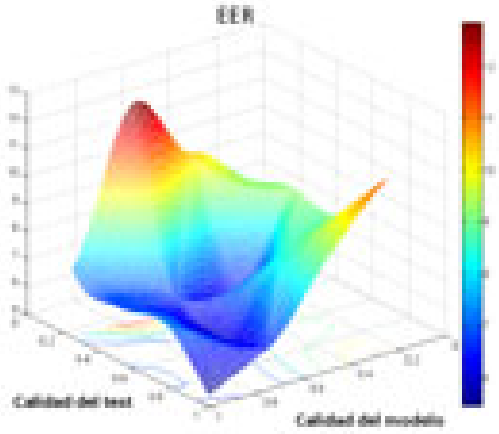


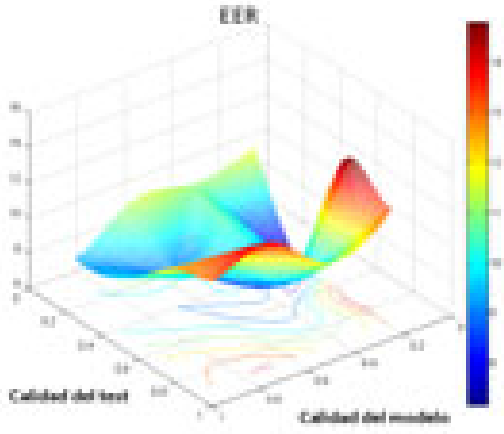
Figura 7.26. Desalineamiento para la Q_{modelo} KCEP: condición *tel-mic*.

IMPACTO DE LA VARIABILIDAD EN KLPC

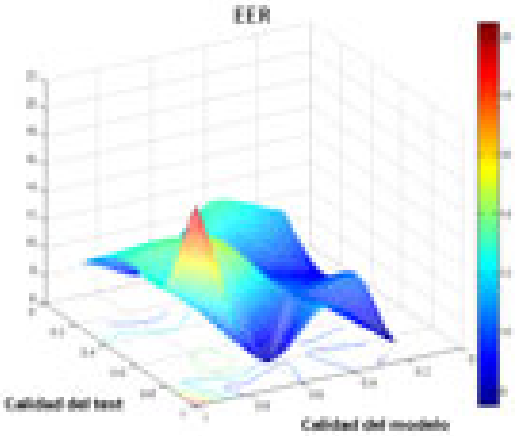
El carácter de este indicador de rendimiento es peor que el del resto de calidades salvo en la condición *tel-tel* (**Figura** 7.27), que presenta una tendencia clara de mejora con la evolución de la calidad del modelo y del test. Observando la **Figura** 7.29 y **Figura 7.30** se puede apreciar como el EER aumenta (disminuye el rendimiento del sistema) cuando se evalúan las muestras de mayor calidad, justo al contrario de lo que debería suceder, lo que le convierte en un indicador de degradación no muy bueno. Este efecto podría estar motivado por la falta de archivos con variabilidad en la calidad a evaluar (que no existieran archivos de calidad KLPC muy alta o muy baja) pero observando los histogramas del número de enfrentamientos y ficheros (de la **Figura** 7.3 a la **Figura** 7.10) se puede apreciar como no sucede, o al menos en menor medida que para la UBML o SNR cuyo rendimiento es notablemente superior a la KLPC, lo que confirma su valor como predictor de rendimiento en términos generales.



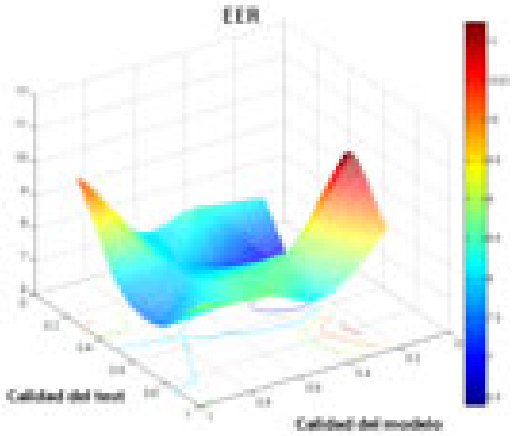
a) EER en función de la calidad del modelo y del test: condición *tel-tel.*



b) EER en función de la calidad del modelo y del test: condición *tel-mic.*



c) EER en función de la calidad del modelo y del test: condición *mic-tel.*



d) EER en función de la calidad del modelo y del test: condición *mic-mic.*

Figura 7.27. EER en función de la calidad KLPC del modelo y del test para las 4 condiciones.

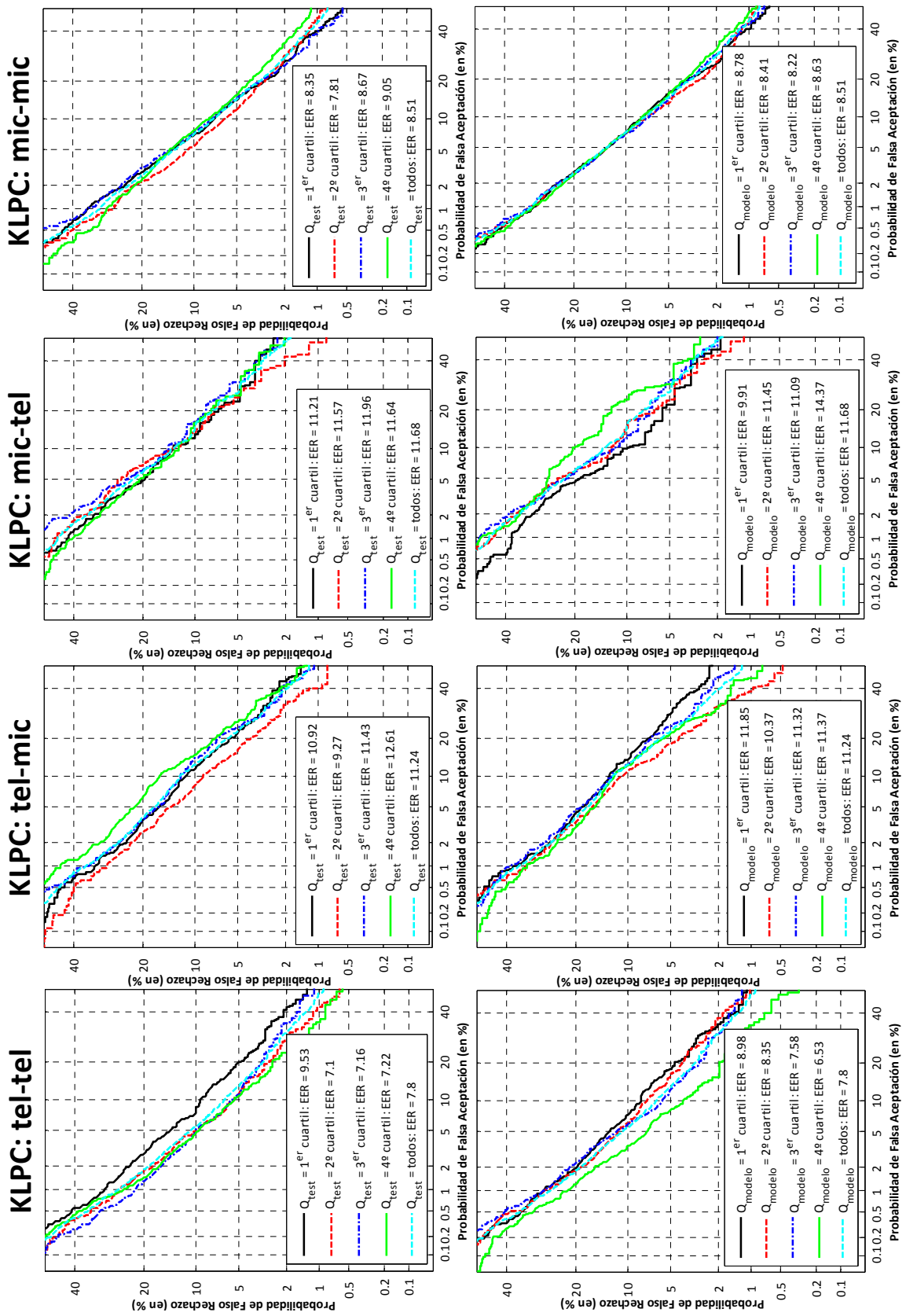


Figura 7.28. Curvas DETS para los subconjuntos dependientes de calidad KLPC.

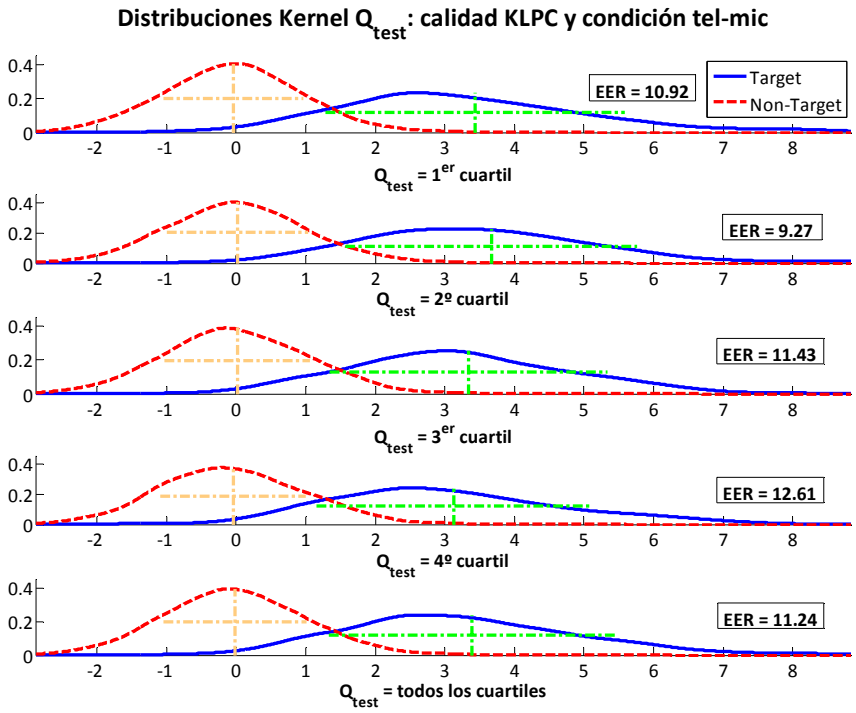


Figura 7.29. Desalineamiento para la Q_{test} KLPC: condición *tel-mic*.

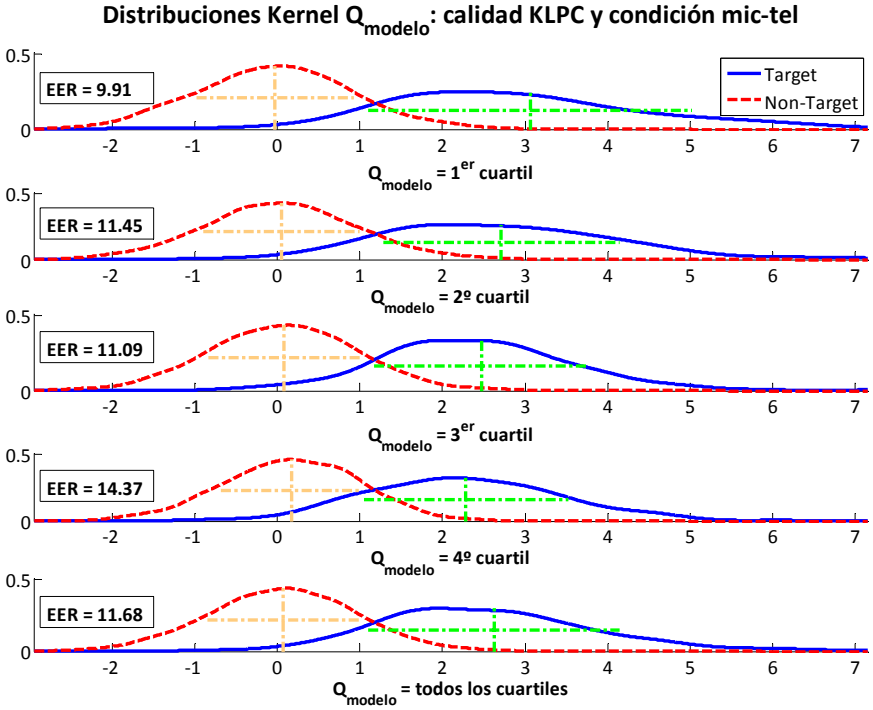


Figura 7.30. Desalineamiento para la Q_{test} KLPC: condición *mic-tel.*

Respecto a la **Figura** 7.29 y la **Figura 7.30**, se aprecia un desalineamiento en las distribuciones dependientes de la calidad KLPC aunque no muy significativo, por lo que será compensado y evaluado de igual forma que el resto de indicadores en el apartado 9.2.

CONCLUSIONES GENERALES

De forma generalizada se puede resumir que el desalineamiento es menos acusado que cuando existe variabilidad en duración, muy posiblemente debido al filtrado Wiener que presentan las condiciones *tel-mic*, *mic-tel* y *mic-mic* que junto con la T-Normalización hace que las distribuciones *non-target* se encuentre alineadas. En este caso la pérdida de rendimiento viene motivada por un desalineamiento perceptible en las distribuciones *target* (ver media representada en color verde en las gráficas de distribuciones *Kernel*) por lo que motiva la definición de nuevos algoritmos que consideren también esta información para otorgar robustez al sistema como los que se estudiarán en la capítulo siguiente.

Otra consideración a tener en cuenta es que el impacto de la variabilidad en la calidad es menos problemático que en la duración en la base de datos evaluada, ya que el peor EER obtenido es de 18.95% para el primer cuartil de la condición *mic-tel* de la calidad SNR (**Figura 7.18**), muy por encima, en cuanto a robustez se refiere, del 43.97% obtenido cuando se estudiaba la variabilidad en la longitud de las muestras. Realmente, la duración y la calidad son igual de críticas, solo que para la calidad no se ha podido observar de forma razonable ya que los archivos de la base de datos de NIST SRE 2008 presentan un rango homogéneo de calidad sin existir casi archivos de calidades extremas como ocurría en duraciones, lo cual motivó la implementación de las bases de datos *DurTelSRE06* y *DurTelSRE08*.

8 COMPENSACIÓN DE VARIABILIDAD CON INFORMACIÓN DE CALIDAD Y DURACIONES

8.1 INTRODUCCIÓN

La principal conclusión derivada del análisis de la variación en las condiciones del habla en los capítulos 6 y 7 es que la variabilidad se traduce en un desalineamiento de los *scores* que causa una degradación del rendimiento global. Aunque este problema puede solucionarse en parte con una normalización de *scores* basada en una cohorte de usuarios impostores que alinee las distribuciones *non target* (4), el típico uso de una cohorte estática para realizar T-Norm (normalización dependiente de test) o Z-Norm (normalización dependiente de modelo) no elimina completamente los efectos de la variabilidad como ya se ha analizado. De hecho, este tipo de normalizaciones tienen por objetivo alinear las distribuciones *non target* sin considerar que el rendimiento de EER viene dado también en función del alineamiento de las distribuciones *target*. Por lo tanto, es acertado pensar que una normalización que considere las distribuciones *target* ayudaría a aumentar el rendimiento del sistema disminuyendo el área de solape entre las dos distribuciones y por lo tanto la EER, pero este tipo de metodología es impracticable ya que se necesitarían muchas muestras del locutor a identificar para disponer de una cohorte representativa del mismo que garantizase unos buenos resultados. Es por ello que en este trabajo se proponen varios métodos de calibración de *scores* basados en modelos de regresión logística y modelado gaussiano, algunos de carácter novedoso (1) (2) y otros basados en trabajos existentes (6) (59) (50), que tengan presente la calidad y/o duración de las muestras de voz para la mejora del rendimiento global considerando la información existente de las distribuciones *target*. Las técnicas que en este capítulo se estudian consisten en transformar los *scores* a LLRs (*Log-Likelihood-Ratio*) (26) que pueden ser interpretados bajo un marco bayesiano (60). Estas técnicas, actualmente revisadas y aceptadas para su publicación en diferentes congresos (1) (2), son:

- **Modelado Gaussiano en una y dos dimensiones** (*One Dimension Gaussian Modelling* o 1D-GM y *Two Dimensions Gaussian Modelling* o 2D-GM).
- **Regresión Logística Lineal en una y dos dimensiones** (*One Dimension Linear Logistic Regression* o 1D-LLR y *Two Dimensions Linear Logistic Regression o* 2D-LLR).
- **Regresión Logística Bilineal con diferentes tipos de información adicional**

(*Bilinear Logistic Regression Type 1, 2, 3 y 4 o* BLR-1, BLR-2, BLR-3 y BLR-4).

Los métodos 1D-GM, 1D-LLR y BLR-1 se pueden aplicar a los subconjuntos de *scores* agrupados por calidad o duración de modelo y de test, por lo que en total se han implementado 11 métodos diferentes como parte de este proyecto fin de carrera, de los cuales la versión dependiente de las condiciones de test de 1D-GM, 2D-GM, 2D-LLR y los 4 tipos de BLR (5 contando que BLR-1 puede utilizar información de la variabilidad en test o en el modelo) son aportaciones originales de este trabajo al grupo ATVS y a la comunidad científica internacional (1) (2).

Antes de describir cada algoritmo es importante remarcar que para comprobar la eficacia de las técnicas propuestas, los parámetros de compensación deben ser extraídos de enfrentamientos en los que no intervenga ni el mismo modelo ni el mismo test que en la comparación bajo estudio. Es decir, el entrenamiento se hará de forma realista (no se dispone de ninguna información de la procedencia de la muestra de voz), bien mediante el uso de bases de datos diferentes para entrenar los modelos y probarlos o bien mediante el uso de técnicas de validación cruzada,

Por último, mencionar que aunque los métodos se expliquen basándose en el ejemplo de la duración son totalmente compatibles con cualquier tipo de calidad. Simplemente los valores i y j corresponderán a la duración del modelo y del test de forma respectiva o al cuartil bajo estudio si se trata de compensar la variabilidad de la calidad.

8.2 Modelado Gaussiano (*Gaussian Modelling* o GM)

Este algoritmo transforma los *scores* generados por el sistema dada la duración o valor de calidad del modelo y fichero de test en una relación de verosimilitud en el dominio logarítmico (*Log-Likelihood-Ratio*). De esta forma, asumiendo que las distribuciones de *scores* son gaussianas[6] para la puntuación bajo estudio x_{ij}, cuya duración o calidad del modelo es i y la duración o calidad del fichero de test j, la nueva similitud se calcula de la siguiente manera:

$$x_{ij}{}^{norm} = \log\left(\frac{P_{ij}\left(x_{ij}|T,\mu_{ij}{}^{T},\sigma_{ij}{}^{T}\right)}{P_{ij}\left(x_{ij}|NT,\mu_{ij}{}^{NT},\sigma_{ij}{}^{NT}\right)}\right)$$

donde P_{ij} son distribuciones gaussianas. Los parámetros $\mu_{ij}{}^{T}$ y $\sigma_{ij}{}^{T}$ son la media y

[6] La suposición de que las distribuciones son gaussianas es factible ya que los *scores* han sido transformados mediante técnicas de normalización como T-Norm (8).

desviación típica calculadas de las puntuaciones *target* de entrenamiento para la duración o calidad del modelo i y la duración del test j, y $\mu_{ij}{}^{NT}$ y $\sigma_{ij}{}^{NT}$ son la media y desviación típica calculadas de las puntuaciones *non target* también de entrenamiento.

Una de las versiones de una dimensión de este método (1D-GM) se ha propuesto en un estudio anterior para compensar la variabilidad en la longitud de las muestras (6), donde el valor de i correspondiente al modelo de entrenamiento variaba y el valor la duración o calidad del test j era ignorado agrupando los *scores* por segmento de test (equivalente a considerarlos todos de forma simultánea). La otra versión de una dimensión desarrollada en este trabajo consiste en realizar justo lo contrario: variar la duración o valor j de la calidad del segmento de test ignorando la duración o calidad del modelo i de la calidad (agrupando los modelos). Combinando estas dos técnicas se obtiene el método de modelado gaussiano de dos dimensiones (2D-GM), original de este proyecto.

El esquema de compensación se representa en la Figura 8.1 y la Figura 8.2 para el subconjunto de *scores* dependiente de la longitud del modelo y el subconjunto dependiente de la longitud de test respectivamente. Como se puede observar, la **Figura 8.1** muestra un enfrentamiento de duración de modelo i (D_{modelo}) y todas las duraciones de test j. En ella el conjunto de *scores* de la base de datos *DurTelSRE06* se utiliza para entrenar las medias y desviaciones $\mu_{ij}\, y\, \sigma_{ij}$ que se usarán para normalizar el *score* de igual duración de la base de datos de test *DurTelSRE08*. La **Figura** 8.2 ilustra el mismo concepto pero con la duración del test j (D_{test}).

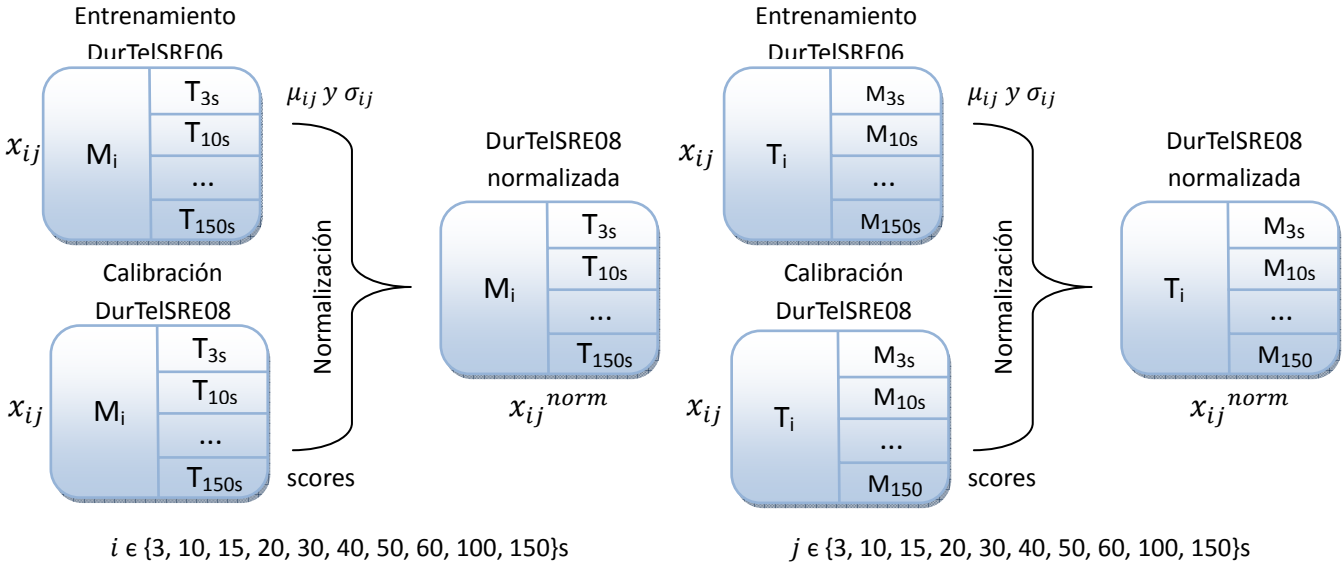


Figura 8.1. Esquema de compensación para D_{modelo} para los métodos 1D-GM y 1D-LLR.

Figura 8.2. Esquema de compensación para D_{test} para los métodos 1D-GM y 1D-LLR.

La **Figura** 8.3 y la **Figura 8.4** esquematizan el mismo método de compensación pero esta vez para la calidad del modelo (Q_{modelo}) y la calidad del test (Q_{test}) por medio de cuartiles, como se ha descrito en el apartado 7.3. En este caso, como se ha detallado en el capítulo 5.1, la base de datos de entrenamiento y de test es la misma, existiendo la posibilidad de utilizar algoritmos de validación cruzada como *Jackknife* (61) que normalicen cada *score* con la misma base de datos, ignorando aquellos enfrentamientos que contengan el mismo modelo o fichero de test de la puntuación a estudiar. También se representa el esquema relativo a la compensación en dos dimensiones en la **Figura** 8.5 y **Figura 8.6**.

Por último, destacar que para una correcta normalización es necesario disponer de la media y desviación típica propia de cada distribución dependiente de la duración del modelo y del test (según el tipo de normalización GM), por lo que en escenarios reales donde la duración de los archivos en cuestión es distinta a los valores de entrenamiento, dichos valores de medias y desviaciones se obtienen de una interpolación cúbica de los estadísticos obtenidos en condiciones similares de entrenamiento, de forma análoga a como se ha realizado en (6).

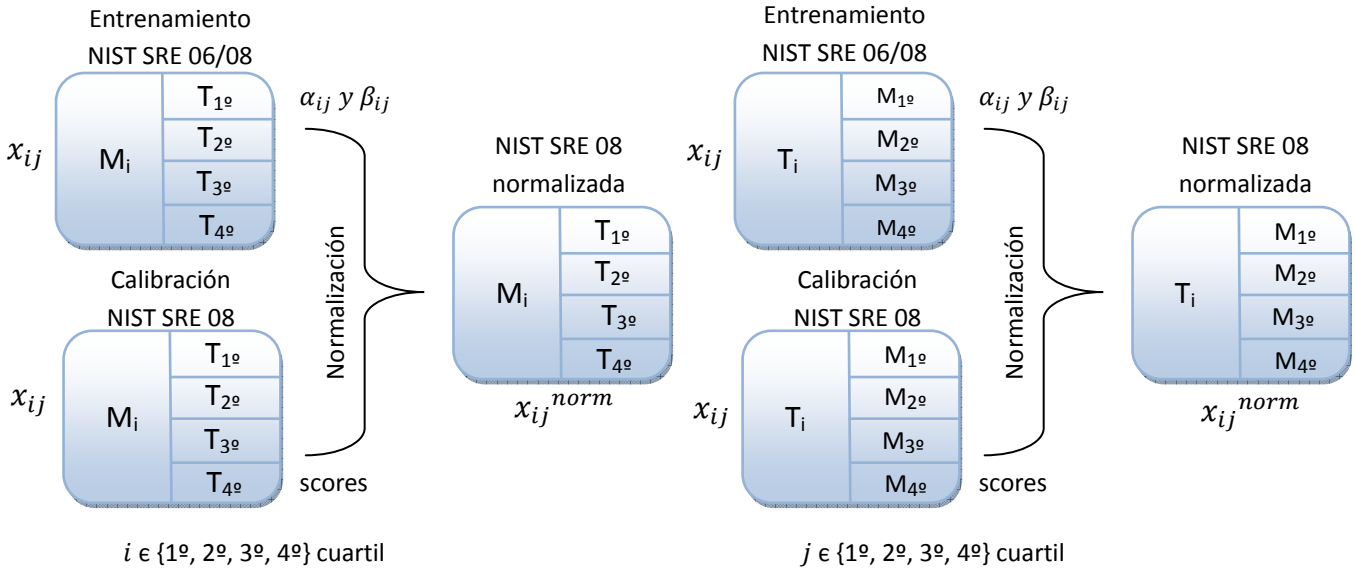
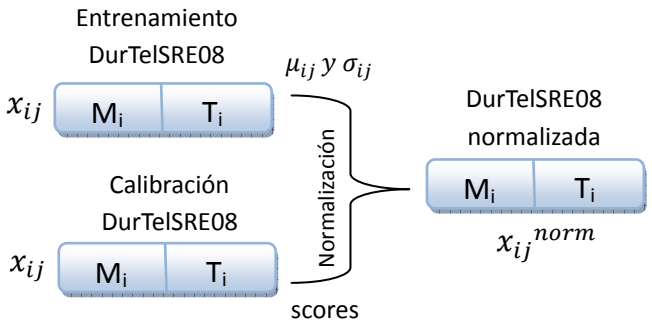


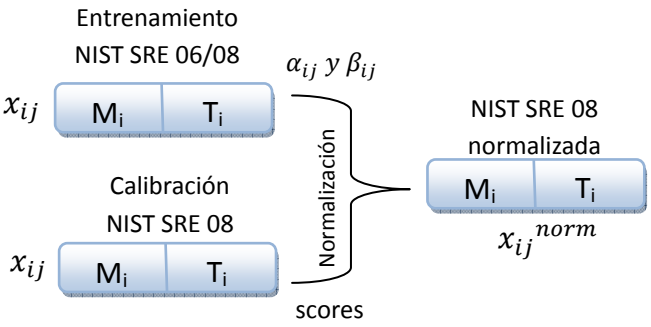
Figura 8.3. Esquema de compensación para Q_{modelo} para los métodos 1D-GM y 1D-LLR.

Figura 8.4. Esquema de compensación para Q_{test} para los métodos 1D-GM y 1D-LLR.



i, j ϵ {3, 10, 15, 20, 30, 40, 50, 60, 100, 150}s

Figura 8.5. Esquema de compensación para D_{modelo}/D_{test} para los métodos 2D-GM y 2D-LLR.



i, j ϵ {1º, 2º, 3º, 4º} cuartil

Figura 8.6. Esquema de compensación para Q_{modelo}/Q_{test} para los métodos 2D-GM y 2D-LLR.

Destacar también que para los métodos de 2 dimensiones D_{modelo} y D_{test} son equivalentes ya que en ambos casos se utilizarán ambos. Lo mismo sucede para Q_{modelo} y Q_{test}.

8.3 Regresión logística lineal (*Linear Logistic Regression* o LLR)

Esta técnica constituye una contribución original de este trabajo no habiéndose utilizado previamente para la compensación de la variabilidad en duración y en calidad. En la actualidad ha sido revisada por expertos en la materia de carácter nacional e internacional (1) (2).

El método LLR (*Linear Logistic Regression*) es un algoritmo de compensación similar al GM salvo que en vez de trabajar con funciones de densidad de probabilidad gaussianas trabaja con modelos de regresión logística. Por lo tanto, de igual modo, transforma el conjunto de puntuaciones generadas bajo unas condiciones de calidad o duración en el logaritmo de una relación de verosimilitud (62). Para entender el concepto de regresión logística es necesario mencionar la dependencia entre los parámetros de interés mediante del *Teorema de Bayes* (63) representado en las dos primeras fórmulas, donde se expresa la relación que existe ente el *score* x_{ij} bajo estudio y las probabilidades condicionadas *target* y *non target*.

$$P_{ij}(x_{ij}|T) \cdot P_{ij}(T) = P_{ij}(T|x_{ij}) \cdot P_{ij}(x_{ij})$$

$$P_{ij}(x_{ij}|NT) \cdot P_{ij}(NT) = P_{ij}(NT|x_{ij}) \cdot P_{ij}(x_{ij})$$

Relacionando ambas ecuaciones (64) se consigue llegar al siguiente modelo:

$$\frac{P_{ij}(T|x_{ij})}{P_{ij}(NT|x_{ij})} = \frac{P_{ij}(x_{ij}|T)\cdot P_{ij}(T)\cdot \cancel{P_{ij}(x_{ij})}}{P_{ij}(x_{ij}|NT)\cdot P_{ij}(NT)\cdot \cancel{P_{ij}(x_{ij})}} = LR_{ij}\cdot\frac{P_{ij}(T)}{P_{ij}(NT)} = \frac{P_{ij}(T|x_{ij})}{1-P_{ij}(T|x_{ij})} = A$$

$$P_{ij}(T|x_{ij}) = A - A\cdot P_{ij}(T|x_{ij})$$

$$P_{ij}(T|x_{ij}) = \frac{A}{1+A} = \frac{1}{1+A^{-1}} = \frac{1}{1+\left(LR_{ij}\cdot\frac{P_{ij}(T)}{P_{ij}(NT)}\right)^{-1}} = \frac{1}{1+e^{-LLR_{ij}-\log\left(\frac{P_{ij}(T)}{P_{ij}(NT)}\right)}}$$

El modelo de regresión logística consiste en añadir los pesos α_{ij} y β_{ij} que ponderen $LLR_{ij}\ y\ \log(\frac{P_{ij}(T)}{P_{ij}(NT)})$ tal que se optimice su función de coste para cada conjunto de datos de entrenamiento (*target* y *non target*):

$$P_{ij}(T|x_{ij}) = \frac{1}{1+e^{-\alpha_{ij}\cdot LLR_{ij}-\beta_{ij}'\cdot\log\left(\frac{P_{ij}(T)}{P_{ij}(NT)}\right)}}$$

$$C(\alpha_{ij},\beta_{ij}) = \sum_{m=1}^{M}\log\left(1+e^{-\alpha_{ij}\cdot LLR_{ij}-\beta_{ij}'\cdot\log\left(\frac{P_{ij}(T)}{P_{ij}(NT)}\right)}\right)$$

donde $C(\alpha_{ij},\beta_{ij})$ es la función a optimizar para cada conjunto de datos M de entrenamiento tal que se satisfaga las siguientes relaciones:

$$P_{ij}(T|x_{ij})\ \uparrow\uparrow\ si\ x_{ij}\ \epsilon\ target$$

$$P_{ij}(T|x_{ij})\ \downarrow\downarrow\ si\ x_{ij}\ \epsilon\ non\ target$$

La relación entre el *score* normalizado ${x_{ij}}^{norm}$ y los pesos α_{ij} y β_{ij} se describe a continuación:

$$P_{ij}(T|x_{ij}) = \frac{1}{1+e^{-\alpha_{ij}\cdot LLR_{ij}-\beta_{ij}'\cdot\log\left(\frac{P_{ij}(T)}{P_{ij}(NT)}\right)}} = \frac{1}{1+B}$$

$$\frac{P_{ij}(T|x_{ij})}{P_{ij}(NT|x_{ij})} = \frac{\frac{1}{1+B}}{1-\frac{1}{1+B}} = e^{\alpha_{ij}\cdot LLR_{ij}+\beta_{ij}'\cdot\log\left(\frac{P_{ij}(T)}{P_{ij}(NT)}\right)}$$

$$\log\left(\frac{P_{ij}(T|x_{ij})}{P_{ij}(NT|x_{ij})}\right) = \alpha_{ij} \cdot LLR_{ij} + \beta_{ij}{}' \cdot \log\left(\frac{P_{ij}(T)}{P_{ij}(NT)}\right)$$

$$\log\left(\frac{P_{ij}(x_{ij}|T) \cdot P_{ij}(T) \cdot \cancel{P_{ij}(x_{ij})}}{P_{ij}(x_{ij}|NT) \cdot P_{ij}(NT) \cdot \cancel{P_{ij}(x_{ij})}}\right) = \alpha_{ij} \cdot LLR_{ij} + \beta_{ij}{}' \cdot \log\left(\frac{P_{ij}(T)}{P_{ij}(NT)}\right)$$

$$\log\left(\frac{P_{ij}(x_{ij}|T)}{P_{ij}(x_{ij}|NT)}\right) = \alpha_{ij} \cdot LLR_{ij} + \beta_{ij}{}' \cdot \log\left(\frac{P_{ij}(T)}{P_{ij}(NT)}\right) - \log\left(\frac{P_{ij}(T)}{P_{ij}(NT)}\right)$$

$$\log\left(\frac{P_{ij}(x_{ij}|T)}{P_{ij}(x_{ij}|NT)}\right) = \alpha_{ij} \cdot LLR_{ij} + \beta_{ij}$$

$$x_{ij}{}^{norm} = \alpha_{ij} \cdot x_{ij} + \beta_{ij}$$

Por lo tanto, el objetivo que persigue esta técnica consiste en hallar α_{ij} y β_{ij} tal que si $x_{ij}{}^{norm}$ ↑↑ se cumpla la hipótesis *target* y si $x_{ij}{}^{norm}$ ↓↓ se cumpla la *non target* (modelo discriminativo).

Es importante remarcar los pesos se obtienen de las puntuaciones de entrenamiento (35)[7] de las comparaciones de los modelos de duración o cuartil i de calidad con los ficheros de test de duración o cuartil j de calidad. De nuevo, al igual que con GM, se obtienen tres variantes de esta técnica: si sólo se tiene en cuenta i o sólo se tiene en cuenta j el método recibe el nombre de 1D-LLR (**Figura 8.1** y la **Figura** 8.2). Si por el contrario se habla de variabilidad en modelo y en test a la vez entonces el método propuesto se denominará 2D-LLR (**Figura** 8.5 y **Figura 8.6**). Al igual que para GM este algoritmo también se puede se puede generalizar para cualquier valor de calidad o duración realizando algún tipo de interpolación cúbica de los α_{ij} y β_{ij}.

8.4 Regresión Logística Bilineal (*Bilinear Logistic Regression* o BLR)

Este método es una técnica elegante que optimiza el modelo de regresión logística condicionado por el conjunto de valores tomando información adicional del enfrentamiento. Un método similar ha sido previamente usado en (10) para fusionar

[7] El toolkit FoCal ha sido usado para el entrenamiento de la LLR. http://sites.google.com/site/nikobrummer/focal

sistemas usando información lingüística del locutor (si el locutor es nativo o no) como información complementaria, pero no información de calidad o duración como en este trabajo, por lo que también constituye una contribución original del mismo. Esta técnica, al estar basada en el modelo de regresión logística se define de forma similar:

$$x_{ij}{}^{norm} = \alpha \cdot x_{ij} + \sum_{k=1}^{K} \alpha_k \cdot \lambda_k \cdot x_{ij} + \beta$$

donde los pesos α, α_k y β son ahora pesos fijos para todos los posibles conjuntos de *scores* dependientes de calidad o duración, y λ_k definido en el rango $(0-1)$ corresponde a la información de las calidades o duraciones del modelo y/o del test exigiendo una normalización del valor respecto del máximo posible (150 segundos para duraciones y 1 para calidades). Para la realización de este trabajo se ha definido dicha información de cuatro maneras diferentes dando lugar a cuatro modalidades del algoritmo:

- **BLR-1**: $\lambda_1 = \{D_m, D_t\}$ o $\lambda_1 = \{Q_m, Q_t\}$ → donde la información adicional corresponde con la duración del modelo o del test o a su calidad, en función del tipo de normalización que se desee hacer.

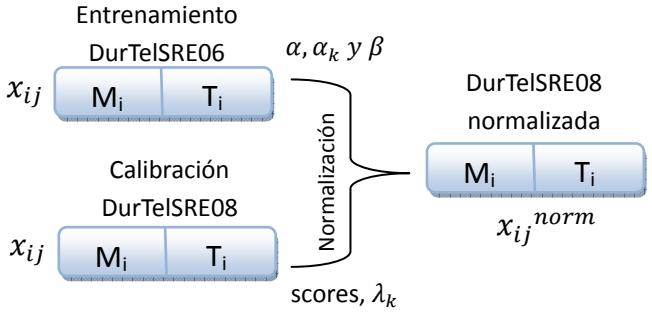
- **BLR-2**: $\lambda_1 = D_m$; $\lambda_2 = D_t$ o $\lambda_1 = Q_m$; $\lambda_2 = Q_t$ → donde se toma la información de duración o de calidad únicamente del modelo o del fichero de test.

- **BLR-3**: $\lambda_1 = |D_m - D_t|$ o $\lambda_1 = |Q_m - Q_t|$ → donde la información adicional es la diferencia entre la duración del modelo y del test o la de sus calidades. Esta configuración intenta motivar el hecho de que puntuaciones tomadas con desajustes grandes de condiciones pueden dar lugar a diferentes desalineamientos.

- **BLR-4**: $\lambda_1 = \sqrt{D_m \cdot D_t}$ o $\lambda_1 = \sqrt{Q_m \cdot Q_t}$ → donde la información complementaria corresponde con la media geométrica de las duraciones o de los valores de calidad.

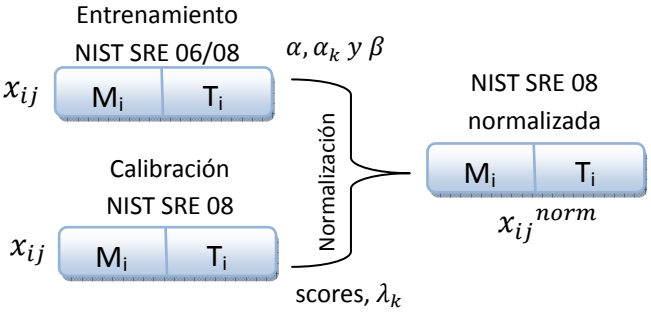
Cabe destacar que aunque sólo se han estudiado estas cuatro variantes se puede considerar como información adicional distintos tipos de calidad o la duración y la calidad a la vez.

Al igual que para los algoritmos anteriormente explicados se muestra el diagrama para la normalización mediante calidad o duración para el método BLR.



i, j ϵ {3, 10, 15, 20, 30, 40, 50, 60, 100, 150}s

Figura 8.7. Esquema de compensación para D_{modelo}/D_{test} para el método BLR.



i, j ϵ {1º, 2º, 3º, 4º} cuartil

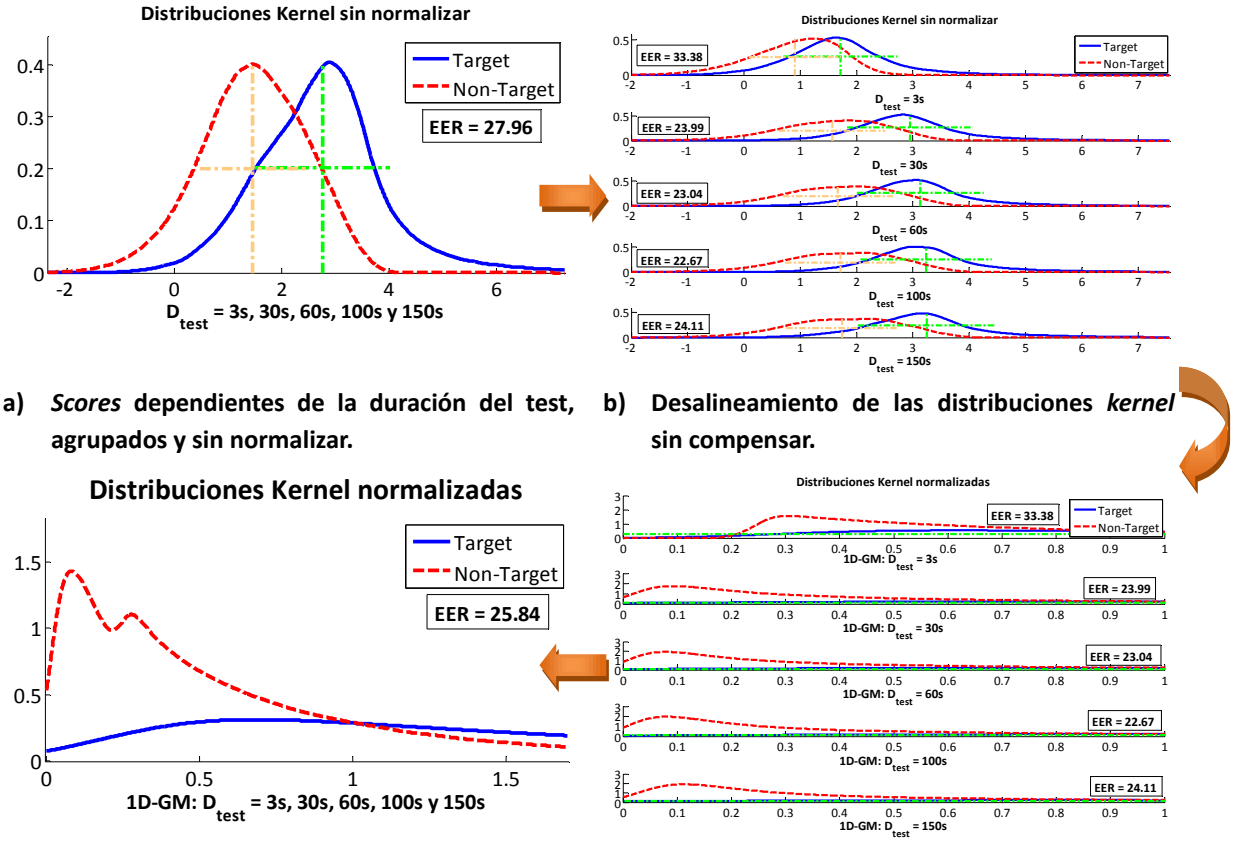
Figura 8.8. Esquema de compensación para Q_{modelo}/Q_{test} para el método BLR.

Por último, es importante mencionar que los algoritmos aquí propuestos no son comparables cuantitativamente frente las técnicas en las que se basan ya que no se dispone de un marco experimental común donde se pueda medir el rendimiento de forma equiparable, es decir, cada técnica ha sido evaluada sobre una base de datos y un sistema distinto coherente con el contexto científico en el que se implementaron.

9 RESULTADOS DE COMPENSACIÓN DE VARIABILIDAD

9.1 COMPENSACIÓN DEL IMPACTO DE LA VARIABILIDAD EN DURACIÓN

La metodología que se ha seguido para comprobar la eficacia de los métodos propuestos es compensar los subconjuntos de *scores* dependientes de duración para comparar el EER y el $MinC_{llr}$ con el del conjunto original. En primera instancia se han testeado aquellos grupos de puntuaciones cuya duración es muy elevada o muy reducida (presentan un carácter extremo) para comprobar el tipo de alineación obtenida cuando se presentan distribuciones muy poco parecidas. Este efecto puede comprobarse en la **Figura** 9.1: la figura a) representa las distribuciones *kernel* de un sistema que trabaja archivos de test de 3s, 10s, 60s, 100s y 150s sin compensar el efecto de desalineamiento que en conjunto producen como se muestra en b). Seguidamente, se normaliza cada conjunto de *scores*



a) ***Scores*** **dependientes de la duración del test, agrupados y sin normalizar.**

b) **Desalineamiento de las distribuciones** ***kernel*** **sin compensar.**

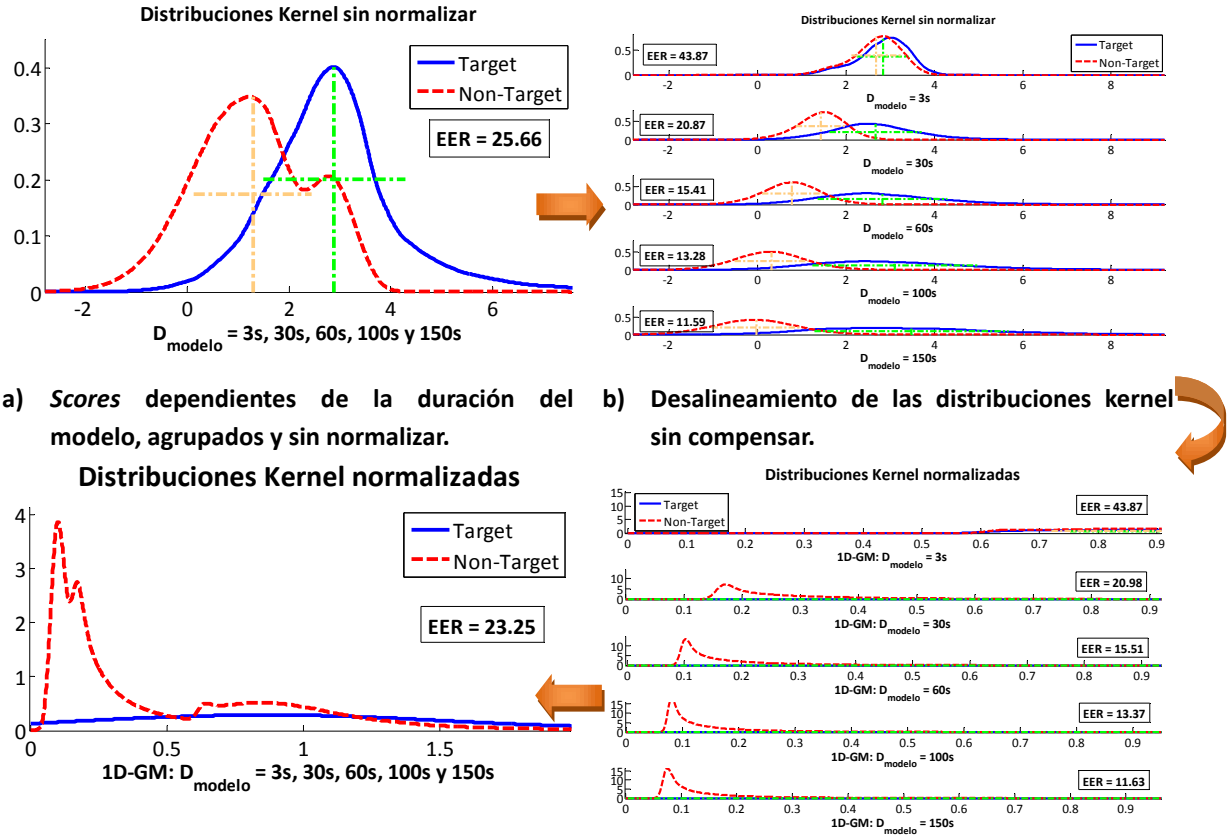
c) **Conjunto de distribuciones** ***kernel*** **agrupadas y normalizadas (*scores* como *log-LR*).**

d) **Distribuciones** ***kernel*** **compensadas y parcialmente alineadas (*scores* como *log-LR*).**

Figura 9.1. Distribuciones kernel: Normalización 1D-GM para el conjunto de *scores* dependientes de la duración del test.

dependientes de la duración de test (figura d)) para después agruparlos y ver que el sistema ha mejorado en cuanto a rendimiento (figura a)) en un 7.56%. Destacar que para ilustrar este efecto no se tiene en cuenta la duración del modelo, es decir, se considera de forma simultánea el conjunto de *scores* cuyo modelo de entrenamiento presenta variabilidad en la duración.

El mismo efecto puede comprobarse si se varía la duración de modelo y no se tiene en cuenta la del fichero de test. Al igual que antes, en la nueva figura a) se representan las distribuciones *target* y *non target* muy solapadas, que tras normalizar mediante 1D-GM se puede comprobar cómo la distribución *non target* se agrupa en torno al valor 0. Aunque en la gráfica c) parece que el solape es mayor que en a), realmente es mucho inferior, ya que al realizar la división y posterior logaritmo, como indica el método descrito en el apartado 8.2, el rango de valores que toma la distribución *target* es mucho mayor que la *non target*, rango que no se representa en la figura para poder observar el efecto de la compensación sobre la distribución *non target.* Por lo tanto, se puede concluir que, favoreciendo al rendimiento, las distribuciones *target* y *non target* presentan un mayor alineamiento debido al rango en el que se definen.



a) ***Scores*** **dependientes de la duración del modelo, agrupados y sin normalizar.**

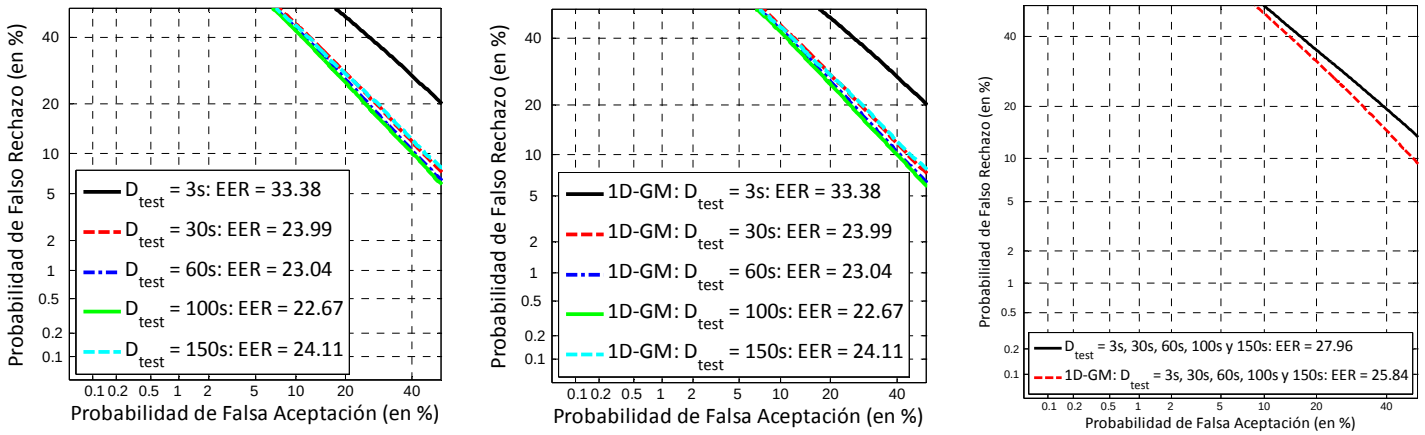
b) Desalineamiento de las distribuciones kernel sin compensar.

c) Conjunto de distribuciones kernel agrupadas y normalizadas (*scores* como log-LR).

d) Distribuciones kernel compensadas y parcialmente alineadas (*scores* como log-LR).

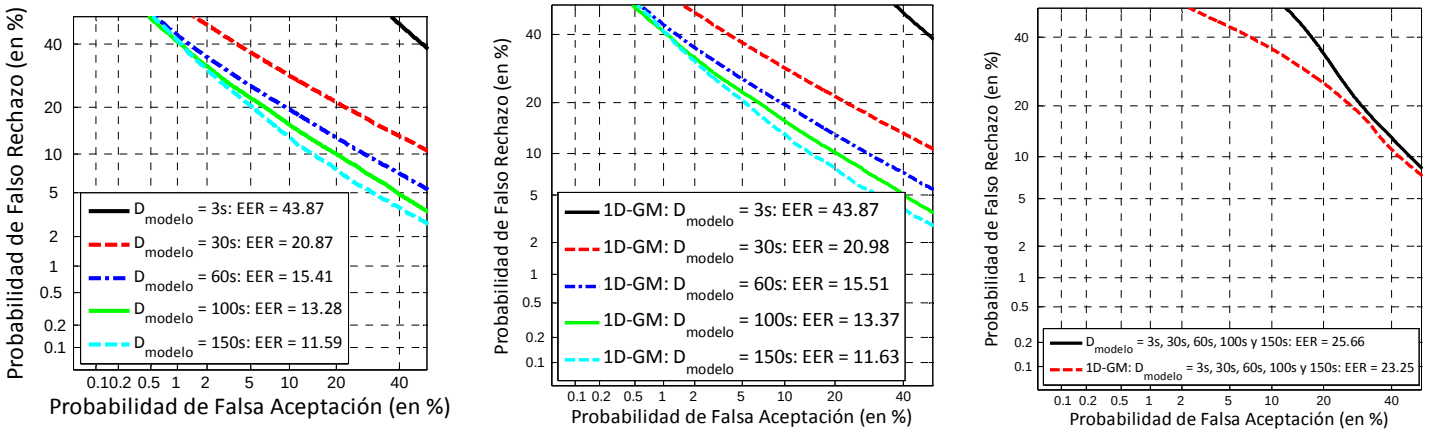
Figura 9.2. Distribuciones kernel: Normalización 1D-GM para el conjunto de *scores* dependientes *DurTelSRE08* de la duración del modelo.

Es importante destacar que la compensación llevada a cabo no mejora propiamente el rendimiento de los sistemas independientes ya que realiza una transformación para todas las puntuaciones (ver mismo EER en la **Figura** 9.3 y en la **Figura** 9.4, donde se representan las curvas DET correspondientes a la **Figura** 9.1 y la Figura 9.2), no siendo así para el conjunto evaluado de forma simultánea en el que sí se aprecia una mejora.



a) ***Scores*** **dependientes de la duración del test, sin normalizar.** **b)** ***Scores*** **dependientes de la duración del test, normalizados.** **c)** ***Scores*** **antes y después de normalizar.**

Figura 9.3. Curvas DET: Normalización 1D-GM para el conjunto de *scores DurTelSRE08* dependientes de la duración del test.



a) ***Scores*** **dependientes de la duración del modelo, sin normalizar.** **b)** ***Scores*** **dependientes de la duración del test, normalizados.** **c)** ***Scores*** **antes y después de normalizar.**

Figura 9.4. Curvas DET: Normalización 1D-GM para el conjunto de *scores DurTelSRE08* dependientes de la duración del test.

En segundo lugar, se ha medido el rendimiento de estas técnicas a través de la evaluación de distintos subconjuntos de *scores* en los que el test presenta varias duraciones de forma simultánea (por ejemplo, agrupando los subconjuntos de 3s y 10s), y el modelo puede tomar cualquiera de las disponibles en la base de datos creada (agrupando los subconjuntos de duración igual a 3s, 10s, 15s, 20s, 30s, 40s, 50s, 60s, 100s y 150s). Los mismos experimentos se han realizado para el modelo, ignorando la duración del test.

Las tablas que a continuación se exponen presentan una comparativa entre los 8 métodos estudiados para los diferentes conjuntos de datos en cuanto a EER y $MinC_{llr}$. En ellas se muestra el EER original y la mejora en rendimiento de cada método sobre el conjunto indicado en la primera columna. También se expresa la media de mejora para cada conjunto de duraciones y para cada técnica. Por último, destacar que para los algoritmos de dos dimensiones también se tiene en cuenta la duración que debería obviarse (la del modelo, en la **Tabla** 9.1 y la **Tabla 9.2** o la del test en la **Tabla** 9.3 y la **Tabla** 9.4).

	EER (%)	Mejora del EER (%)								
D_{test} (seg)	Sin compensar	1D-GM	2D-GM	1D-LLR	2D-LLR	BLR-1	BLR-2	BLR-3	BLR-4	
3, 10	31,34	2,79	7,17	2,68	7,41	2,63	6,86	4,04	-0,95	**4,08**
3, 100	31,81	9,61	16,99	9,31	17,23	9,34	15,73	4,59	0,67	**10,43**
3, 150	33,08	5,21	10,26	5,15	10,54	4,75	8,73	1,62	1,06	**5,92**
10, 100	26,39	3,47	15,67	3,33	15,79	3,57	15,49	2,29	-0,71	**7,36**
10, 150	27,41	1,97	11,98	1,95	12,28	1,89	11,67	0,22	-0,20	**5,22**
3, 10, 150	31,11	4,55	10,42	4,43	10,73	2,14	6,32	-0,16	0,04	**4,81**
3, 100, 150	31,33	10,28	19,01	10,00	19,29	8,38	13,89	6,67	-0,15	**10,92**
3, 30, 60, 100, 150	27,96	7,59	19,54	7,44	19,80	3,38	11,67	2,02	-1,04	**8,80**
Conjunto completo	26,55	4,30	16,37	4,20	16,62	1,84	12,21	0,16	0,10	**6,98**
		5,53	**14,16**	**5,39**	**14,41**	**4,21**	**11,40**	**2,38**	**-0,13**	**Media**

Tabla 9.1. Mejora en EER para los 8 métodos para el conjunto de datos *DurTelSRE08* con variabilidad en la duración del test.

	MinCllr	Mejora del MinCllr (%)								
D_{test} (seg)	Sin compensar	1D-GM	2D-GM	1D-LLR	2D-LLR	BLR-1	BLR-2	BLR-3	BLR-4	
3, 10	0,83	2,10	5,93	2,03	5,94	1,26	4,19	2,78	1,39	**3,20**
3, 100	0,83	7,04	14,62	6,81	14,56	4,42	9,06	3,99	-0,08	**7,55**
3, 150	0,86	4,60	9,09	4,46	9,05	3,39	6,14	2,48	0,18	**4,92**
10, 100	0,74	3,12	14,38	3,02	14,36	2,41	11,28	1,92	0,74	**6,40**
10, 150	0,76	2,21	10,89	2,16	10,91	1,97	9,05	0,99	0,16	**4,79**
3, 10, 150	0,83	3,91	9,14	3,80	9,14	1,63	5,29	0,30	0,00	**4,15**
3, 100, 150	0,82	7,31	16,06	7,10	15,98	3,91	9,14	4,14	0,03	**7,96**
3, 30, 60, 100, 150	0,77	6,33	17,87	6,19	17,81	1,35	9,19	1,45	1,83	**7,75**
Conjunto completo	0,75	4,05	15,46	3,95	15,41	1,05	9,67	0,46	2,06	**6,51**
		4,52	**12,60**	**4,39**	**12,57**	**2,38**	**8,11**	**2,06**	**0,70**	**Media**

Tabla 9.2. Mejora en MinCllr para los 8 métodos para el conjunto de datos *DurTelSRE08* con variabilidad en la duración del test.

	EER (%)	Mejora del EER (%)								
D_{modelo} (seg)	Sin compensar	1D-GM	2D-GM	1D-LLR	2D-LLR	BLR-1	BLR-2	BLR-3	BLR-4	
3, 10	38,90	0,16	6,86	-0,18	6,76	-0,58	2,03	2,00	1,09	**2,27**
3, 100	35,66	12,99	16,88	13,56	17,29	13,68	14,66	0,71	8,50	**12,28**
3, 150	39,64	6,11	10,13	6,43	10,37	6,6	7,73	0,5	2,66	**6,32**
10, 100	29,60	13,85	21,73	14,48	21,97	14,34	16,99	0,25	3,87	**13,44**
10, 150	31,58	6,48	15,53	6,86	15,63	6,73	9,97	-0,35	-0,06	**7,60**
3, 10, 150	36,86	3,18	9,48	3,04	9,57	3,48	5,48	0,48	-1,09	**4,20**
3, 100, 150	33,47	14,66	18,49	15,48	19,02	15,64	16,52	3,16	10,08	**14,13**
3, 30, 60, 100, 150	25,66	9,40	14,31	9,82	14,68	10,00	11,63	0,00	0,64	**8,81**
Conjunto completo	26,55	9,42	16,37	9,42	16,62	9,56	12,21	0,16	0,10	**9,23**
		8,47	**14,42**	**8,77**	**14,66**	**8,83**	**10,80**	**0,77**	**2,87**	**Media**

Tabla 9.3. Mejora en EER para los 8 métodos para el conjunto de datos DurTelSRE08 con variabilidad en la duración del modelo.

	MinCllr	Mejora del MinCllr (%)								
D_{modelo} (seg)	Sin compensar	1D-GM	2D-GM	1D-LLR	2D-LLR	BLR-1	BLR-2	BLR-3	BLR-4	
3, 10	0,94	2,21	5,08	1,91	4,98	1,67	2,30	0,90	-0,18	**2,36**
3, 100	0,82	7,64	9,30	7,76	9,45	7,18	7,64	-0,37	3,74	**6,54**
3, 150	0,89	3,27	4,48	3,36	4,58	3,10	3,47	-0,27	1,57	**2,95**
10, 100	0,74	6,93	11,33	6,41	11,24	6,50	7,40	-0,78	0,65	**6,21**
10, 150	0,79	3,02	7,54	2,57	7,36	2,75	3,71	-0,31	-0,36	**3,29**
3, 10, 150	0,90	3,84	6,67	3,33	6,54	2,49	3,16	0,24	0,71	**3,37**
3, 100, 150	0,78	8,61	10,29	8,79	10,50	8,04	8,49	-0,70	4,43	**7,31**
3, 30, 60, 100, 150	0,73	12,67	16,13	12,43	16,24	9,97	10,83	0,00	4,06	**10,29**
Conjunto completo	0,75	10,7	15,46	10,04	15,41	8,6	9,67	0,46	2,06	**9,05**
		6,54	**9,59**	**6,29**	**9,59**	**5,59**	**6,30**	**-0,09**	**1,85**	**Media**

Tabla 9.4. Mejora en MinCllr para los 8 métodos para el conjunto de datos *DurTelSRE08* con variabilidad en la duración del modelo.

A la vista de los resultados obtenidos se pueden destacar varios aspectos:

- Todos los métodos salvo el BLR-3 mejoran, en términos generales, el rendimiento en cuanto a EER y $MinC_{llr}$ se refiere. Por lo tanto, la diferencia entre las longitudes del vector de entrenamiento y el de test no es representativa como para considerarse como información complementaria de cada a ayudar en la

normalización. Realmente es lógico que suceda esto ya que enfrentamientos de longitud elevada con igual diferencia entre las duraciones del modelo y del test (150s y 140s) presentarían la misma información complementaria que aquellos de longitud pequeña (15s y 5s), por lo que hace que esta información sea poco distintiva y en conclusión de poca utilidad.

- Los métodos de una dimensión (1D-GM, 1D-LLR y BLR-1) presentan una mejora mucho menor que los de dos dimensiones (ver **Figura** 9.5) siendo más acusado cuando los *scores* dependen de la duración del fichero de test: por ejemplo el método 1D-GM, considerando el conjunto completo de *scores* dependientes de la duración de test, presenta un 4.30% de mejora en EER frente al 16.37% que ofrece 2D-GM. Este hecho se debe principalmente al uso de T-Norm, que alinea parcialmente las distribuciones *non target* haciendo que el sistema presente por sí mismo un EER más competitivo que cuando no se usa (los EER originales son mayores, en general, para los subconjuntos dependientes de la duración del modelo).

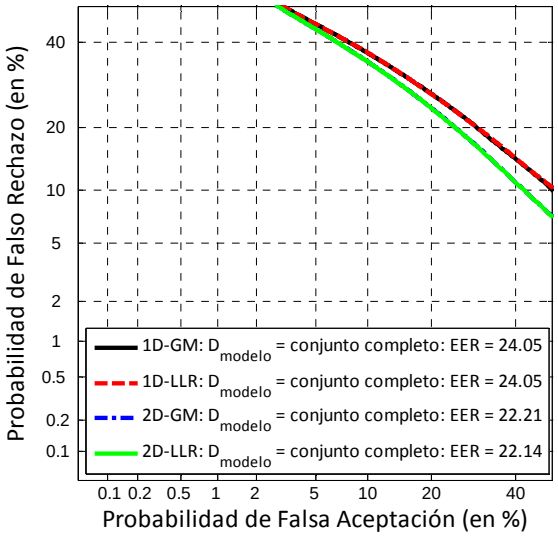
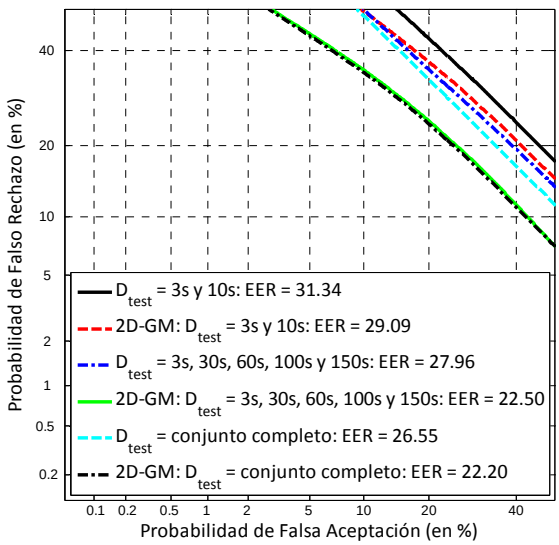


Figura 9.5. Rendimiento de los métodos 1D-GM, 1D-LLR, 2D-GM y 2D-LLT sobre el conjunto completo dependiente de la calidad del modelo

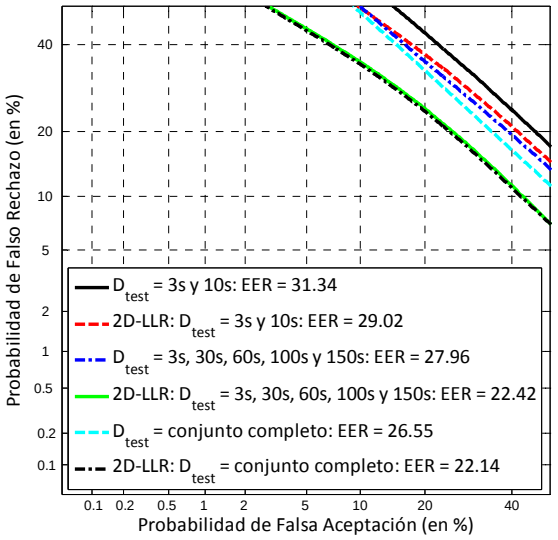
- Cuando se utilizan técnicas que consideran la duración del modelo y del test el rendimiento aumenta notablemente, siendo mayor en las técnicas 2D-GM y 2D-LLR (15.46% y 15.41% de mejora de EER para el subconjunto dependiente de la duración de test y 16.37% y 16.62% para la dependiente del modelo).

- Las mejores rendimientos dependientes de la duración del test se obtienen para 2D-GM y 2D-LLR evaluando las longitudes 3s, 30s, 60s, 100s y 150s de manera simultánea (ver **Figura** 9.6. y **Figura** 9.7)

- Aunque las técnicas BLR requieren de menos parámetros que entrenar que 2D-GM y 2D-LLR, éstas presentan peor rendimiento (ver **Figura** 9.8) y pueden diverger por no mencionar que presentan un coste computacional más elevado que en el resto de técnicas propuestas.

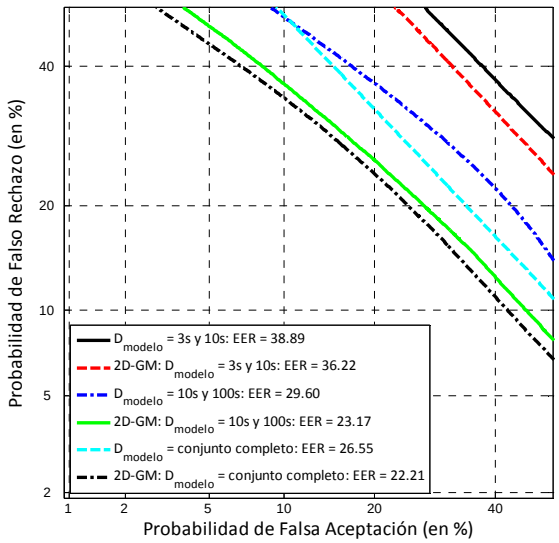


a) **Peor, mejor y global (subconjunto) dependiente de la duración de test para el método 2D-GM.**

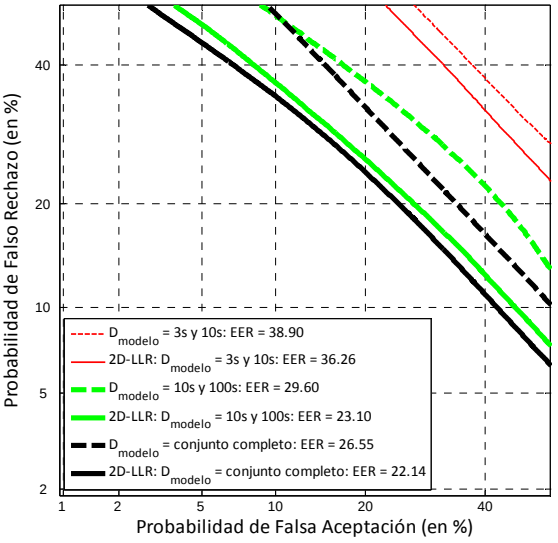


b) **Peor, mejor y global (subconjunto) dependiente de la duración de test para el método 2D-LLR.**

Figura 9.6. Rendimiento de las técnicas 2D-GM y 2D-LLR para el peor subconjunto, el mejor y el conjunto global de *scores* dependiente de la duración de test.

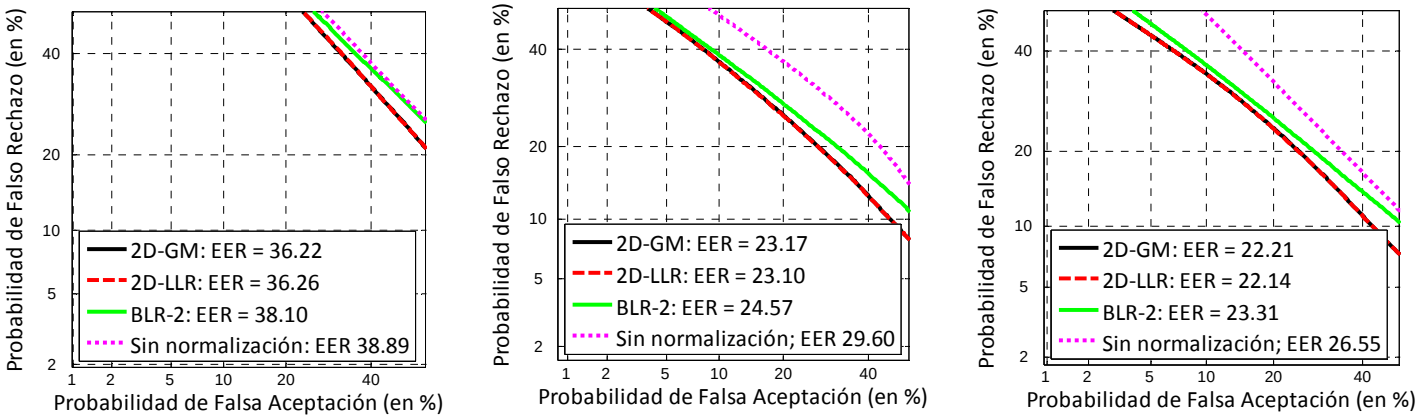


a) **Curvas DET representativas del método 2D-GM para subconjuntos dependientes de la duración del modelo.**



b) **Curvas DET representativas del método 2D-LLR para subconjuntos dependientes de la duración del modelo.**

Figura 9.7. Rendimiento de las técnicas 2D-GM y 2D-LLR para el peor subconjunto, el mejor y el conjunto global de *scores* dependiente de la duración de modelo.



a) Comparativa de los 3 mejores métodos para D_{modelo} = 3s y 10s. **b) Comparativa de los 3 mejores métodos para D_{modelo} = 10s y 100s.** **c) Comparativa de los 3 mejores métodos para D_{modelo} = conjunto completo.**

Figura 9.8. Rendimiento de los 3 mejores sistemas evaluado sobre el conjunto global de *scores*: 2D-GM, 2D-LLR y BLR-2.

9.2 COMPENSACIÓN DEL IMPACTO DE LA VARIABILIDAD EN CALIDAD

Los experimentos que en este apartado se presentan, analizan el rendimiento de los algoritmos estudiados en el capítulo 8 en cuanto a compensación de variabilidad en la calidad de las muestras de voz se refiere. A continuación, en la **Tabla** 9.5 se detalla el conjunto de experimentos llevado a cabo. Cabe destacar que la realización de cada ensayo ha sido motivada por su predecesor, adaptando las condiciones experimentales para extraer un análisis representativo del impacto de la calidad en sistemas de reconocimiento de locutor.

Antes de analizar la tabla es necesario aclarar algunos conceptos:

- La condición realista equivale a entrenar los algoritmos con una base de datos que no contenga información del modelo o fichero de test del enfrentamiento analizar, es decir, diferente a la de evaluación (escenario real). También se han realizado pruebas a posteriori (escenario optimista, en el que la base de datos de entrenamiento contiene muestras del enfrentamiento evaluado) ya que no se disponía de archivos microfónicos procedentes de la base de datos de NIST SRE 2006. Para suplir esta carencia, experimentos en esta línea se han llevado a cabo realizando una validación cruzada (*Jackknife*) en la que se entrena la misma base de datos que la que se evalúa pero sin contener al modelo y fichero de test del enfrentamiento a compensar.

- En todos los casos se analiza el subconjunto de *scores*, obteniéndose los mismos resultados para los métodos de 2 dimensiones cuando dicho conjunto depende de la calidad del modelo o del test, salvo para BLR-1 cuya información complementaria puede corresponderse con la calidad del modelo (λ_m) o la calidad test (λ_m). En el caso de las técnicas de una dimensión (1D-GM y 1D-LLR) ocurre igual que en BLR-1, por lo que se analizarán como métodos diferentes dependientes de la calidad del modelo (Q_m) o de la calidad del test (Q_t). Por lo tanto se realizarán experimentos en los cuales el objetivo es medir el rendimiento de los 11 métodos propuestos (1D-GM:Q_m, 1D-GM:Q_t, 1D-LLR:Q_m, 1D-LLR:Q_t, 2D-GM, 2D-LLR, BLR-1:λ_m, BLR-1:λ_t, BLR-2, BLR-3 y BLR-4).

Configuración de las muestras	**Muestras de entrenamiento**	**Muestras de evaluación**	**Condiciones**	**Escenario**	**Métodos**	**Calidades**
4 bins	NIST SRE 06	NIST SRE 08	*tel-tel*	Realista	Todos	KLPC KCEP UBML SNR P563
4 cuartiles	NIST SRE 06	NIST SRE 08	*tel-tel*	Realista	Todos	KLPC KCEP UBML SNR P563
4 cuartiles	NIST SRE 08	NIST SRE 08	*tel-tel* *tel-mic* *mic-tel* *mic-mic*	Optimista	Mejor funcionamiento (2D-GM, 2D-LLR, BLR-2 y BLR-4).	KLPC KCEP UBML SNR P563
Sin considerar el 25% de calidad peor	NIST SRE 08	NIST SRE 08	*tel-tel* *tel-mic* *mic-tel* *mic-mic*	Optimista	Técnicas BLR	UBML
2, 4 y 8 cuartiles	NIST SRE 08	NIST SRE 08	*tel-tel* *tel-mic* *mic-tel* *mic-mic*	Realista (*Jackknife*)	2D-LLR	UBML
4 cuartiles	NIST SRE 08	NIST SRE 08	*tel-tel* *tel-mic* *mic-tel* *mic-mic*	Realista (*Jackknife*)	BLR	UBML

Tabla 9.5. Experimentos para calidades.

EXPERIMENTOS CON 4 BINS Y 4 CUARTILES EN EL ESCENARIO REALISTA

Bins	EER (%)	Mejora del EER (%)											
06/08	Original	1D-GM: Q_m	1D-GM: Q_t	1D-LLR: Q_m	1D-LLR: Q_t	2D-GM	2D-LLR	BLR-1: λ_m	BLR-1: λ_t	BLR-2	BLR-3	BLR-4	
KLPC	7,80	1,16	-0,35	-0,31	-0,08	-3,01	-4,17	-0,35	-0,50	-0,39	-2,01	-0,70	**-0,97**
KCEP		-0,36	0,36	0,36	0,20	3,01	-0,69	0,12	0,00	-0,08	-0,08	0,00	**0,26**
UBML		0,00	3,05	0,00	1,66	3,01	2,66	-1,00	1,27	1,16	3,24	2,36	**1,58**
SNR		0,35	1,01	0,27	2,01	0,35	1,35	-0,31	1,01	2,01	1,01	0,73	**0,89**
P563		0,00	1,51	0,00	0,42	1,70	-0,35	1,00	0,66	1,00	0,00	2,32	**0,75**
		0,23	**1,12**	**0,06**	**0,84**	**1,01**	**-0,24**	**-0,11**	**0,49**	**0,74**	**0,43**	**0,94**	Media

Tabla 9.6. Mejora del EER para todos los métodos y calidades usando 4 bins (escenario realista)

Bins	$MinC_{llr}$	Mejora del $MinC_{llr}$ (%)											
06/08	Original	1D-GM: Q_m	1D-GM: Q_t	1D-LLR: Q_m	1D-LLR: Q_t	2D-GM	2D-LLR	BLR-1: λ_m	BLR-1: λ_t	BLR-2	BLR-3	BLR-4	
KLPC	7,80	-0,33	0,09	-0,44	0,31	-2,52	-2,06	-0,02	-0,04	-0,08	-0,05	-0,19	**-0,48**
KCEP		0,09	-0,07	-0,05	-0,05	2,60	-1,42	0,03	-0,07	-0,04	0,00	-0,06	**0,09**
UBML		0,30	1,94	0,23	2,00	2,60	1,99	0,37	0,31	0,66	2,00	0,99	**1,22**
SNR		-0,49	0,48	-0,53	0,84	-0,47	-1,67	0,12	0,29	1,10	-0,02	0,38	**0,00**
P563		-0,28	1,01	0,11	0,76	0,51	0,66	0,71	0,66	0,71	-0,03	1,00	**0,53**
		-0,14	**0,69**	**-0,14**	**0,77**	**0,54**	**-0,50**	**0,24**	**0,23**	**0,47**	**0,38**	**0,42**	Media

Tabla 9.7. Mejora del MinCllr para todos los métodos y calidades usando 4 bins (escenario realista).

Cuartiles	EER (%)	Mejora del EER (%)											
06/08	Original	1D-GM: Q_m	1D-GM: Q_t	1D-LLR: Q_m	1D-LLR: Q_t	2D-GM	2D-LLR	BLR-1: λ_m	BLR-1: λ_t	BLR-2	BLR-3	BLR-4	
KLPC	7,80	-0,77	-0,46	-0,35	-0,08	-2,66	-4,67	-0,35	-0,58	-0,54	-1,85	-0,66	**-1,18**
KCEP		-0,31	-0,04	-0,35	0,20	-0,35	-0,35	-0,62	-0,35	-0,66	-0,46	-0,66	**-0,36**
UBML		0,23	2,66	-0,04	1,66	2,08	2,01	-1,00	1,27	1,16	3,24	2,36	**1,42**
SNR		-1,00	1,54	-1,35	2,01	-0,35	-2,01	-0,35	1,35	1,47	1,00	1,35	**0,33**
P563		0,00	0,93	0,73	0,42	1,00	0,50	1,00	0,66	2,01	0,00	2,36	**0,87**
		-0,37	**0,93**	**-0,27**	**0,84**	**-0,06**	**-0,90**	**-0,26**	**0,47**	**0,69**	**0,39**	**0,95**	Media

Tabla 9.8. Mejora del EER para todos los métodos y calidades usando 4 cuartiles (escenario realista).

Cuartiles	$MinC_{llr}$	Mejora del $MinC_{llr}$ (%)											
06/08	Original	1D-GM: Q_m	1D-GM: Q_t	1D-LLR: Q_m	1D-LLR: Q_t	2D-GM	2D-LLR	BLR-1: λ_m	BLR-1: λ_t	BLR-2	BLR-3	BLR-4	
KLPC	7,80	-0,06	-0,51	-0,82	0,31	-2,04	-2,91	-0,04	-0,04	-0,08	-0,01	-0,19	**-0,58**
KCEP		-0,19	0,12	-0,41	-0,05	-1,07	-1,02	-0,03	0,02	0,00	0,09	-0,09	**-0,24**
UBML		0,52	0,84	0,45	2,00	1,05	0,91	0,37	0,31	0,66	2,00	0,99	**0,92**
SNR		-0,26	0,86	-0,08	0,84	0,75	0,49	0,10	0,39	0,48	0,03	0,53	**0,38**
P563		0,51	0,68	0,67	0,76	0,50	0,18	0,71	0,66	1,09	-0,03	0,99	**0,61**
		0,10	**0,40**	**-0,04**	**0,77**	**-0,16**	**-0,47**	**0,22**	**0,27**	**0,43**	**0,42**	**0,45**	Media

Tabla 9.9. Mejora del MinCllr para todos los métodos y calidades usando 4 bins (escenario realista).

El análisis de las tablas anteriores se detalla a continuación:

Motivado por la existencia de pocos archivos de calidad extrema (alta o baja), según se analizó en la **Figura** 7.1 y en la **Figura** 7.2, se ha procedido a analizar el efecto que produce el considerar grupos fijos de calidades mediante la definición de intervalos o *bins,* o grupos dinámicos mediante la definición de cuartiles con la consecución de los siguientes resultados: examinando las tablas anteriores se puede comprobar cómo el rendimiento de las técnicas, para la condición *tel-tel* que aquí se estudia, es ligeramente superior cuando se usan *bins* que cuando se usan cuartiles. Este hecho aunque no es representativo, dada la pequeña diferencia en los resultados (ambos en general inferiores al 1%), puede explicarse si se considera que el conjunto de muestras no está lo suficiente desalineado. Es decir, todas las muestras presentan una calidad de valor medio y según se había visto en el apartado 9.1, el rendimiento de los métodos de compensación es mayor cuanto más dispares sean las muestras. Por lo tanto, dada la naturaleza de éstas (todas telefónicas y de calidad relativamente similar, sin valores extremos) este estudio no es concluyente. No obstante, se considerará sólo el uso de cuartiles en los experimentos posteriores ya que permiten trabajar en condiciones más generales donde puede no existir algún tipo de muestra (los algoritmos de modelado gaussiano, basados en el cálculo de los estadísticos de estos intervalos podrían fallar al no disponer de información del *bin* que se desea analizar).

En cuanto a la mayor compensación en términos de EER, ésta se logra mediante la técnica BLR-3 (explicada en el apartado 8.4) cuyo valor es del 3.24% cuando se considera variabilidad en la calidad UBML (destacar que para los métodos BLRs los enfrentamientos no se agrupan por cuartiles o *bins*). Otras compensaciones como 1D-GM: Q_t o 2D-GM muestran una mejora por encima de la media (3.05% usando *bins* y 2.66% usando cuartiles para el método de una dimensión considerando UBML). No obstante, para la compensación de datos telefónicos, bien modelados y compensados en la actualidad la mejora es insignificante respecto a la obtenida en el apartado 9.1 cuando se hablaba de compensar el impacto de la duración (16.37% de mejora de EER para 2D-GM, **Tabla** 9.3), lo

que motiva el análisis del rendimiento en otros ámbitos donde exista un mayor disparidad (condiciones extremas) en cuanto a la calidad de las muestras, como es el caso de las condiciones que tienen en cuenta también micrófonos.

EXPERIMENTOS CON 4 CUARTILES EN EL ESCENARIO OPTIMISTA CONSIDERANDO MUESTRAS MICROFÓNICAS.

Siguiendo con la investigación, el siguiente paso ha sido medir la eficacia de las técnicas propuestas bajo condiciones de mayor variabilidad, es decir, evaluando el escenario optimista cuya base de datos presenta datos telefónicos y microfónicos.

Al repetir los experimentos usando 4 cuartiles sobre el conjunto de métodos desarrollados se ha observado que los métodos 2D-GM, 2D-LLR, BLR-2 y BLR-4 presentan, en términos generales, un rendimiento superior al resto. Estos resultados se evalúan a continuación para las condiciones *tel-tel, tel-mic, mic-tel* y *mic-mic* de la base de datos de 2008.

TT	EER (%)	Mejora del EER (%)					$MinC_{llr}$	Mejora del $MinC_{llr}$ (%)				
08/08	Original	2D-GM	2D-LLR	BLR-2	BLR-4		Original	2D-GM	2D-LLR	BLR-2	BLR-4	
KLPC	7,80	-1,66	-1,16	1,58	0,66	**-0,15**	0,29	-1,37	-0,78	0,35	0,27	**-0,38**
KCEP		-1,43	0,00	0,35	-0,35	**-0,36**		-0,90	-0,28	0,08	-0,01	**-0,28**
UBML		0,00	-0,35	1,51	2,08	**0,81**		0,64	1,20	0,78	0,95	**0,89**
SNR		2,05	1,97	1,35	1,35	**1,68**		0,98	1,79	0,77	0,51	**1,01**
P563		0,00	-0,35	1,70	2,01	**0,84**		0,54	0,75	1,10	1,02	**0,85**
		-0,21	**0,02**	**1,30**	**1,15**	**Media**		**-0,02**	**0,54**	**0,62**	**0,55**	**Media**

Tabla 9.10. Resultados de los métodos con mejor rendimiento (escenario optimista: condición *tel-tel*).

TM	EER (%)	Mejora del EER (%)					$MinC_{llr}$	Mejora del $MinC_{llr}$ (%)				
08/08	Original	2D-GM	2D-LLR	BLR-2	BLR-4		Original	2D-GM	2D-LLR	BLR-2	BLR-4	
KLPC	11,24	0,9	0,45	0,68	0,23	**0,57**	0,36	-1,1	-0,04	-0,09	-0,02	**-0,31**
KCEP		4,52	5,43	1,81	0	**2,94**		2,49	3,45	0,16	-0,04	**1,52**
UBML		6,26	8,82	6,79	6,56	**7,11**		3,92	5,89	2,49	2,65	**3,74**
SNR		6,11	7,01	6,94	6,56	**6,66**		3,14	3,54	3,38	2,87	**3,23**
P563		0	0,15	-0,23	0	**-0,02**		-0,33	0,89	-0,02	-0,01	**0,13**
		3,56	**4,37**	**3,20**	**2,67**	**Media**		**1,62**	**2,75**	**1,18**	**1,09**	**Media**

Tabla 9.11. Resultados de los métodos con mejor rendimiento (escenario optimista: condición *tel-mic*).

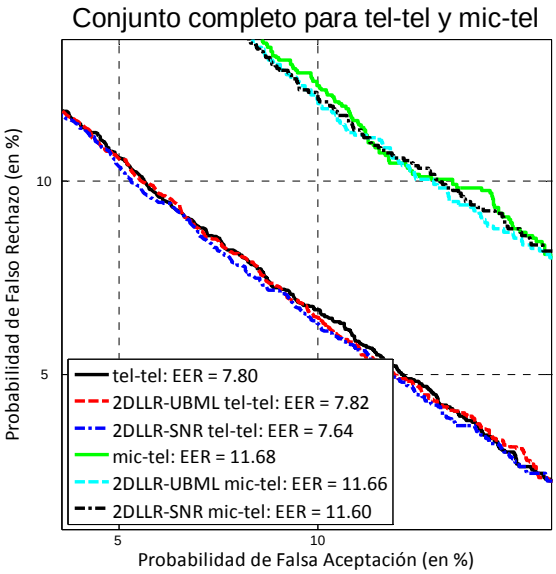
MT	EER (%)	Mejora del EER (%)					$MinC_{llr}$	Mejora del $MinC_{llr}$ (%)				
08/08	Original	2D-GM	2D-LLR	BLR-2	BLR-4		Original	2D-GM	2D-LLR	BLR-2	BLR-4	
KLPC	11,68	-2,66	-0,81	-0,81	-0,81	**-1,27**	0,41	0,67	1,29	0,02	0	**0,50**
KCEP		-2,33	-0,81	-1,85	-2,5	**-1,87**		1,07	1,86	0,64	0,76	**1,08**
UBML		0,64	0,16	-0,81	-1,05	**-0,27**		2,8	3,16	-0,38	-0,3	**1,32**
SNR		-1,61	0,72	-3,46	-3,38	**-1,93**		4,3	4,09	0,37	0,17	**2,23**
P563		0	-1,93	-1,53	-0,4	**-0,97**		2,18	2,67	0,77	0,88	**1,63**
		-1,19	**-0,53**	**-1,69**	**-1,63**	**Media**		**2,20**	**2,61**	**0,28**	**0,30**	**Media**

Tabla 9.12. Resultados de los métodos con mejor rendimiento (escenario optimista: condición *mic-tel*).

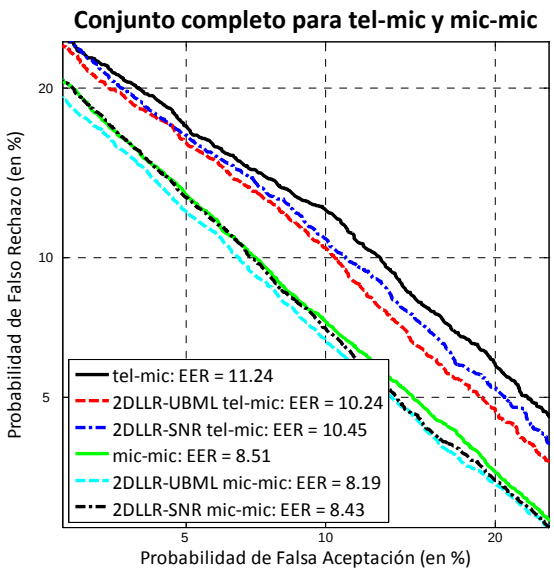
MM	EER (%)	Mejora del EER (%)					$MinC_{llr}$	Mejora del $MinC_{llr}$ (%)				
08/08	Original	2D-GM	2D-LLR	BLR-2	BLR-4		Original	2D-GM	2D-LLR	BLR-2	BLR-4	
KLPC	8,51	0,47	0	0,05	-0,1	**0,11**	0,31	-0,02	0,35	0,42	0,24	**0,25**
KCEP		0,68	0,52	0,31	0,31	**0,46**		1,76	1,96	0,28	0,24	**1,06**
UBML		3,34	3,76	1,93	2,3	**2,83**		3,98	4,03	2,48	2,69	**3,30**
SNR		0,89	0,94	1,3	1,3	**1,11**		1,51	1,59	1,19	1,16	**1,36**
P563		-0,78	-1,04	0	0	**-0,46**		0,3	0,42	0,2	0,05	**0,24**
		0,92	**0,84**	**0,72**	**0,76**	**Media**		**1,51**	**1,67**	**0,91**	**0,88**	**Media**

Tabla 9.13. Resultados de los métodos con mejor rendimiento (escenario optimista: condición *mic-mic*).

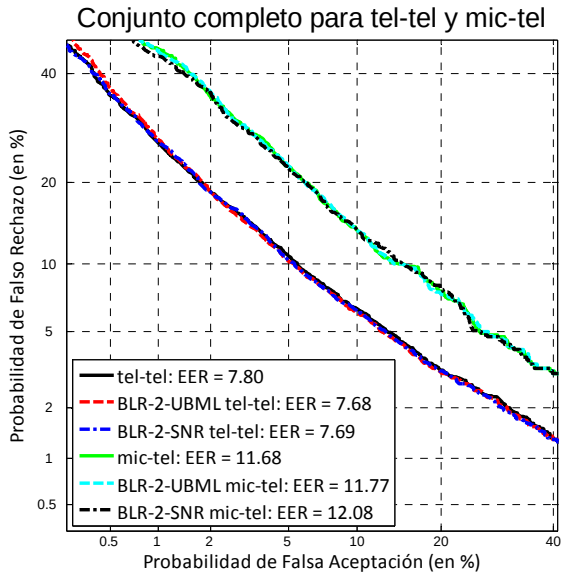
Como era de esperar, usando información con distinto tipo de calidad se aprecian mejoras más sustanciales, como se puede apreciar para la condición *tel-mic*, la cual presenta su mayor rendimiento cuando se evalúa el conjunto de *scores* dependiente de la calidad UBML (hasta 8.82% de mejora en EER para 2D-LLR, **Tabla** 9.11 y **Figura** 9.9.b). No obstante, las técnicas descritas también presentan buenos resultados con otras medidas de calidad como la SNR (hasta 7.01% de mejora en EER para 2D-LLR, **Tabla** 9.11 y **Figura** 9.9.b), hecho que se esperaba debido a la alta correlación entre ambas medidas de calidad. Por otra parte, las medidas KLPC, KCEP y P563 siguen sin ser útiles en la compensación posiblemente debido a que no son indicadores de rendimiento tan buenos como la UBML o la SNR. Otra valoración importante a realizar es que las técnicas mejoran más los enfrentamientos *mic-mic* que los *tel-tel* (ver **Figura** 9.9), hecho que se atribuye a una mayor variabilidad en la calidad de los enfrentamientos como puede verse en la **Figura** 7.2 y en la **Figura** 7.5. No obstante, aunque existe mucha variabilidad para las calidades KLPC y KCEP (**Figura** 7.5) la compensación no logra grandes resultados, posiblemente debido a que no son indicadores de degradación suficientemente útiles (ver resultados de KLPC y KCEP en la **Tabla** 9.13). Por último, destacar que los métodos usados no consiguen mejorar el rendimiento de los enfrentamientos *mic-tel* (**Tabla** 9.12 y **Figura** 9.9).



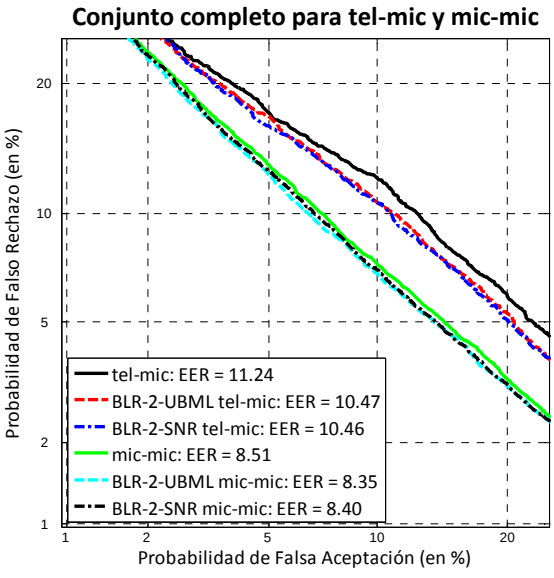
a) **Rendimiento del método 2D-LLR dependiente de UBML y de SNR para las condiciones con peor mejora (*tel-tel* y *mic-tel*).**



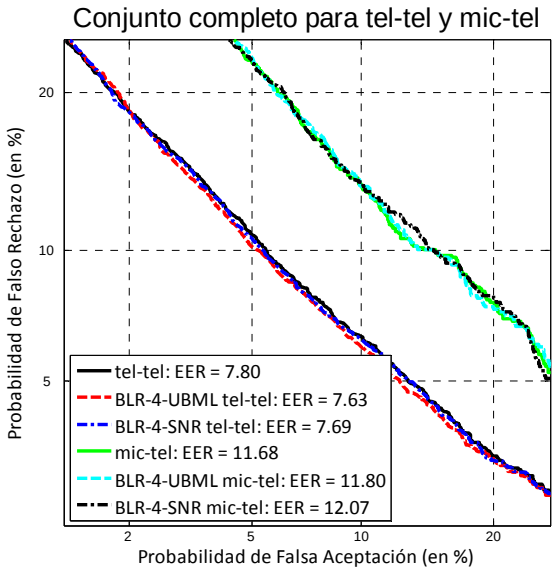
b) **Rendimiento del método 2D-LLR dependiente de UBML y de SNR para las condiciones con mayor mejora (*tel-tel* y *mic-tel*).**



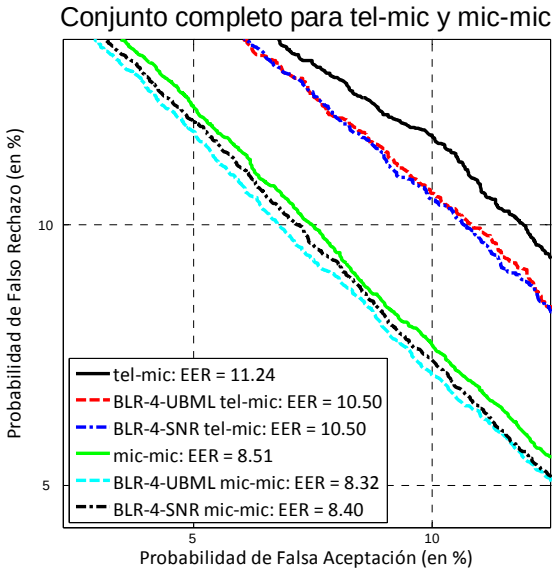
c) **Rendimiento del método BLR-2 dependiente de UBML y de SNR para las condiciones con peor mejora (*tel-tel* y *mic-tel*).**



d) **Rendimiento del método BLR-2 dependiente de UBML y de SNR para las condiciones con peor mejora (*tel-tel* y *mic-tel*).**



e) Rendimiento del método BLR-4 dependiente de UBML y de SNR para las condiciones con peor mejora (*tel-tel* y *mic-tel*).



f) Rendimiento del método BLR-4 dependiente de UBML y de SNR para las condiciones con peor mejora (*tel-tel* y *mic-tel*).

Figura 9.9. Rendimiento de las técnicas 2D-LLR, BLR-2 y BLR-4.

EXPERIMENTOS CON 4 CUARTILES SOBRE EL CONJUNTO DE *SCORES* DEPENDIENTE DE LA CALIDAD UBML IGNORANDO EL 25% DE LOS ENFRENTAMIENTOS CON PEOR CALIDAD.

En este apartado se evalúa el rendimiento de las técnicas BLR con el objetivo de ver si realmente estos métodos compensan bien las calidades bajas, siendo ideal su análisis al no requerir de clasificación de *scores* (en función del valor de las medidas de calidad bajo estudio) ya que toman directamente la información de calidad del modelo y del test sobre el conjunto de enfrentamientos sin el 25% de ellos que presentan un peor valor de calidad UBML (sería equivalente a excluir el primer cuartil).

TT-UBML	Mejora del EER (%)		Mejora del $MinC_{llr}$ (%)		
08/08	Sin excluir	25% excluido	Sin excluir	25% excluido	
BLR-1 λ_{modelo}	-0,39	-0,69	0,37	0,35	
BLR-1 λ_{test}	2,01	1,66	0,53	0,41	
BLR-2	1,51	1,51	0,78	0,75	
BLR-3	3,01	1,00	2,04	0,77	
BLR-4	2,08	2,01	0,95	0,82	
	1,64	**1,10**	**0,93**	**0,62**	**Media**

Tabla 9.14. Rendimiento para BLR excluyendo el 25% de *scores* de peor UBML (condición *tel-tel*).

TM-UBML	Mejora del EER (%)		Mejora del $MinC_{llr}$ (%)		
08/08	Sin excluir	25% excluido	Sin excluir	25% excluido	
BLR-1 λ_{modelo}	3,17	3,17	0,88	1,06	
BLR-1 λ_{test}	4,98	4,75	1,79	1,45	
BLR-2	6,79	6,94	2,49	2,71	
BLR-3	1,13	0,83	0,13	0,30	
BLR-4	6,56	6,33	2,65	2,24	
	4,53	**4,40**	**1,59**	**1,55**	**Media**

Tabla 9.15. Rendimiento para BLR excluyendo el 25% de *scores* de peor UBML (condición *tel-mic).*

MT-UBML	Mejora del EER (%)		Mejora del $MinC_{llr}$ (%)		
08/08	Sin excluir	25% excluido	Sin excluir	25% excluido	
BLR-1 λ_{modelo}	-0,32	0,08	0,14	-0,02	
BLR-1 λ_{test}	-1,53	-0,32	-0,40	-0,38	
BLR-2	-0,81	-1,61	-0,38	-0,17	
BLR-3	0,08	0,16	0,16	0,23	
BLR-4	-1,05	-0,81	-0,30	0,22	
	-0,73	**-0,50**	**-0,16**	**-0,02**	**Media**

Tabla 9.16. Rendimiento para BLR excluyendo el 25% de *scores* de peor UBML (condición *mic-tel).*

MM-UBML	Mejora del EER (%)		Mejora del $MinC_{llr}$ (%)		
08/08	Sin excluir	25% excluido	Sin excluir	25% excluido	
BLR-1 λ_{modelo}	0,99	0,37	1,40	0,77	
BLR-1 λ_{test}	0,89	1,30	1,25	1,17	
BLR-2	1,93	2,45	2,48	1,90	
BLR-3	1,25	0,05	0,95	1,05	
BLR-4	2,30	1,46	2,69	1,86	
	1,47	**1,13**	**1,75**	**1,35**	**Media**

Tabla 9.17. Rendimiento para BLR excluyendo el 25% de *scores* de peor UBML (condición *mic-mic).*

En general, suprimir del conjunto de *scores* aquellos enfrentamientos de peor calidad ayudará a mejorar el rendimiento del sistema global (a costa de no evaluarlos), pero lo que en este apartado se pretende caracterizar es la influencia de este hecho sobre la eficacia de las técnicas BLR mostradas en la tabla. Comparando los resultados "*Sin excluir*" con los "*25% excluido*" se puede observar como apenas existe diferencia en términos generales. De hecho, observando la media de mejoras (última línea de la **Tabla** 9.14, la **Tabla**

9.15 y la **Tabla** 9.17) se observa un pequeño decremento en el rendimiento cuando se excluyen datos, lo que viene a confirmar que cuanta más extrema sea la calidad mayor eficaces serán los métodos propuestos.

EXPERIMENTOS CON 2, 4 Y 8 CUARTILES PARA EL MÉTODO 2D-LLR Y CALIDAD UBML MEDIANTE VALIDACIÓN CRUZADA (*JACKKNIFE*)

En esta sección se pretende verificar en el caso realista, usando *Jackknife* sobre los datos de 2008), que la elección de 4 cuartiles para el desarrollo de este trabajo ha sido acertada. Observando la **Tabla** 9.18 se puede ver cómo la mejora para la condición *tel-mic* en términos de EER y $MinC_{llr}$ es mayor usando 4 cuartiles (7.54% y 3.75% respectivamente) frente a 2 cuartiles (6.33% y 2.93) y 8 cuartiles (7.47% y 0.74%). Notar que la configuración de 8 cuartiles es especialmente crítica para la condición *mic-mic.*

2D-LLR	EER (%)	Mejora del EER (%)			$MinC_{llr}$	Mejora del $MinC_{llr}$ (%)		
UBML	Original	2 cuartiles	4 cuartiles	8 cuartiles	Original	2 cuartiles	4 cuartiles	8 cuartiles
TEL-TEL	7.8	-1.08	-1	0.62	0.29	0.14	-0.53	-1.74
TEL-MIC	11.24	6.33	7.54	7.47	0.36	2.93	3.75	0.74
MIC-TEL	11.68	-0.48	-2.33	-7.17	0.41	-0.29	-0.71	-5.91
MIC-MIC	8.51	2.19	3.03	3.44	0.31	1.79	3.22	3.24

Tabla 9.18. Comparación del rendimiento de 2D-LLR-UBML usando 2, 4 y 8 cuartiles (condición *tel-tel*).

COMPARACIÓN DEL RENDIMIENTO USANDO ESCENARIO REALISTA (*JACKKNIFE*) Y OPTIMISTA PARA LAS CUATRO CONDICIONES

La **Tabla** 9.19 muestra el rendimiento de los algoritmos 2D-LLR y BLR-4 dependientes de la calidad UBML del conjunto de *scores* en función del escenario elegido. Como se puede ver, el uso de un marco experimental optimista para la elaboración de los experimentos sobrevalora la eficacia de los métodos implementados bajo las condiciones de variabilidad de la base de datos de NIST SRE 2008: existe una diferencia de hasta 2 puntos de mejora (de 3.75% a 5.89% en 2D-LLR para $MinC_{llr}$) utilizando la evaluación optimista en vez de la realista, lo que supone todo un acierto el haber utilizado métodos de validación cruzada.

	Mejora del EER (%)				Mejora del $MinC_{llr}$ (%)			
UBML	2D-LLR: JACKKNIFE	2D-LLR: Optimista	BLR-4: JACKKNIFE	BLR-4: Optimista	2D-LLR: JACKKNIFE	2D-LLR: Optimista	BLR-4: JACKKNIFE	BLR-4: Optimista
TEL-TEL	-1	-0,35	1.93	2.08	-0.53	1,20	0.76	0.95
TEL-MIC	7.54	8,82	6.11	6.56	3.75	5,89	2.30	2.65
MIC-TEL	-2.33	0,16	-1.53	-1.05	-0.71	3,16	-0.78	-0.3
MIC-MIC	3.03	3,76	2.14	2.30	3.22	4,03	2.53	2.68

Tabla 9.19. Comparación del rendimiento de 2D-LLR-UBML usando 2, 4 y 8 cuartiles (condición *tel-tel*).

10 CONCLUSIONES Y TRABAJO FUTURO.

10.1 CONCLUSIONES

En este trabajo se ha realizado un estudio exhaustivo del impacto de la variabilidad en la duración y calidad de los ficheros de voz que intervienen en la identificación de un sujeto en un sistema de verificación de locutor. Para ello se ha analizado el rendimiento del sistema presentado por el ATVS en las evaluaciones de 2008 (apartado 5.2) bajo las suposiciones de variabilidad en las condiciones de ensayo. A lo largo de los capítulos 6 y 7 se ha demostrado que esta variabilidad se traduce en un desalineamiento en las distribuciones *target* y non *target* dependiente del indicador de variabilidad estudiado, el cual causa un empeoramiento del rendimiento global del sistema en términos de EER y $MinC_{llr}$. Pero la variabilidad en las condiciones no sólo afecta a este rendimiento sino que supone un problema a la hora de establecer un único umbral en el sistema para que pueda trabajar de forma robusta en estas condiciones, o cuando se desea combinar o calibrar sistemas a nivel de puntuación para obtener un sistema global más robusto o con una interpretación probabilística (60). Por ello, y para reducir este efecto se han implementado 11 nuevos métodos, algunos de ellos estudiados previamente en la literatura (1D-GM dependiente del modelo y BLR haciendo uso de la información lingüística del locutor) (5) (10) (11) y otros como 1D-LLR (dependiente de la variabilidad en el modelo y dependiente de la variabilidad en el test), 2D-GM, 2D-LLR y BLR con la información complementaria en este proyecto implementada, de carácter novedoso (1) (2). Dichas técnicas consideran toda la información disponible (funciones de probabilidad *target* y *non target*) para aplicar una normalización más efectiva y complementaria a la de otros métodos existentes como T-Norm (también utilizado) o Z-Norm, que sólo tratan de alinear el conjunto de distribuciones *non target* a través de una normalización de media y desviación dependiente de test o modelo de enfrentamiento, respectivamente.

La metodología que se ha seguido para comprobar la eficacia de los métodos propuestos es compensar los subconjuntos de *scores* dependientes de duración o calidad para comparar el EER normalizado con el original.

Respecto a la compensación de duración, en primera instancia se ha evaluado el rendimiento de los sistemas que trabajan en condiciones de duración extrema, bien del modelo de entrenamiento o bien del fichero de test, obteniéndose los mejores resultados de hasta el 21.77% para el método 2D-GM y del 21.97% para el método 2D-LLR en condiciones de variabilidad extrema y del 16.37% y 16.62% para el conjunto completo (ver

Tabla 9.3, **Figura** 9.6, **Figura** 9.7 y **Figura** 9.8), ambos implementados como contribución propia de este trabajo. De esta investigación, en cuanto a compensación de duración se refiere se pueden extraer también las siguientes conclusiones:

- Todos los métodos salvo el BLR-3 mejoran, en términos generales, el rendimiento en cuanto a EER y $MinC_{llr}$. En especial los de 2 dimensiones, siendo los métodos 2D-GM y 2D-LLR óptimos en cuanto a rendimiento ofrecido y coste computacional se refiere. También son generalizables a cualquier tipo de duración interpolando los parámetros de compensación extraídos.

- Los métodos de una dimensión (1D-GM, 1D-LLR y BLR-1), independientemente del tipo de duración a compensar (modelo o test), presentan una mejora mucho menor que respecto a los de dos dimensiones (ver **Figura** 9.5) siendo más acusado cuando los *scores* dependen de la duración del fichero de test. Este hecho se debe principalmente al uso de T-Norm, que alinea parcialmente las distribuciones *non target* haciendo que el sistema presente por sí mismo un EER más competitivo que cuando no se usa.

- Cuando se utilizan técnicas que consideran la duración del modelo y del test el rendimiento aumenta notablemente, siendo mayor en las técnicas 2D-GM y 2D-LLR (15.46% y 15.41% de mejora de EER para el subconjunto dependiente de la duración de test y 16.37% y 16.62% para la dependiente del modelo).

- Aunque las técnicas BLR sean más óptimas que 2D-GM y 2D-LLR, éstas presentan peor rendimiento (ver **Figura** 9.8) y pueden diverger por no mencionar que presentan un coste operacional muy elevado, factor que las convierte en computacionalmente inviables cuando se trata manejan pruebas grandes como ha sido el caso de este proyecto.

Respecto a la compensación de la variabilidad en calidad, el enfoque seguido se presenta a continuación: la primera aproximación ha consistido en evaluar las técnicas de compensación de *scores* utilizando diferentes configuraciones de las muestras de los conjuntos de enfrentamientos dependientes de la calidad KLPC, KCEP, UBML, SNR y P.563, motivado por la existencia de pocos archivos de calidad extrema (alta o baja): se han utilizado intervalos o *bins* y cuartiles, demostrando que el rendimiento de las técnicas bajo estas suposiciones es similar, aunque se ha decidido utilizar 4 cuartiles para el resto de experimentos por su carácter generalista (métodos como 2D-GM pueden fallar al calcular los estadísticos sobre un conjunto de *scores* que no existe), habiéndose comprobado que su rendimiento es mayor que utilizando otro número de cuartiles. Los experimentos se han realizado en:

- **Escenarios realistas**: por una parte, se ha evaluado la eficacia de los algoritmos para compensar la variabilidad de la base de datos de NIST SRE 2008 *tel-tel* dependiente de las 5 calidades a través del entrenamiento de la base de datos de NIST SRE 2006. Por otra parte se ha evaluado el rendimiento de los algoritmos 2D-LLR y BLR para las condiciones *tel-tel, tel-mic- mic-tel y mic-mic* mediante validación cruzada para la evaluación de 2008 (ver **Tabla** 9.5).

- **Escenarios optimistas**: se ha evaluado la utilidad de los de algoritmos implementados para las condiciones *tel-tel, tel-mic- mic-tel y mic-mic* usando la base de datos de NIST SRE 2008.

Bajo todas estas consideraciones se ha llegado a las siguientes conclusiones:

- Todas las bases de datos experimentales utilizadas presentan una calidad de valor medio y no de carácter extremo como en duraciones, y pese a que es sencillo construir un conjunto de datos de duraciones extremas, no es trivial ni incluso prudente construir artificialmente un conjunto de datos de calidades extremas. Por lo tanto, se ha tenido que trabajar con la base de datos disponible, y por ello el rendimiento de los métodos de compensación no es muy elevado. No obstante, se aprecian buenos resultados en la condición *tel-mic* para los conjuntos de *scores* dependientes de la calidad UBML o la calidad SNR, obteniéndose un valor máximo del 8.82% de mejora en EER para 2D-LLR (**Tabla** 9.11 y **Figura** 9.9.b) y un 7.07% para cada calidad, lo que motiva el análisis del rendimiento en otros ámbitos donde exista un mayor disparidad (condiciones extremas) en cuanto a la calidad de las muestras. Por otra parte, las medidas KLPC, KCEP y P563 siguen sin ser útiles para la compensación posiblemente debido a que no son indicadores de rendimiento tan buenos como la UBML y la SNR. Por último, destacar la posibilidad de trabajar con la UBML como medida a compensar ya que su cálculo es obligatorio para establecer la similitud entre el fichero de test y el modelo usando un sistema basado en modelos de mezclas gaussianas (ver apartado 7.2.1).

A continuación se expresa en formato de tabla los objetivos, descritos en el apartado 1.2, conseguidos:

Creación de bases de datos con variabilidad en duración.	
Análisis del impacto de la variabilidad en la duración en el rendimiento del sistema.	
Análisis del impacto de la variabilidad en la calidad en el rendimiento del sistema.	
Implementación de 11 métodos diferentes de compensación.	
Análisis cuantitativo de la eficacia de los algoritmos implementados.	

Tabla 10.1. Objetivos conseguidos.

10.2 TRABAJO FUTURO

El presente proyecto ha generado las siguientes líneas de investigación futuras:

- Trabajar con otras bases de datos con suficiente variabilidad en cuanto al indicador de degradación a estudiar se refiere (duración, UBML, SNR, etc.) para medir de una forma más precisa y realista la eficacia de los algoritmos presentados en este trabajo.

- Evaluar el rendimiento de los métodos descritos bajo otras técnicas de normalización como ZT-Norm que realiza una normalización dependiente de modelo y de test.

- Evaluar el rendimiento de las técnicas estudiadas bajo aproximaciones adaptativas al locutor como KL-Tnorm o AT-Norm, o T-Normalización mediante una cohorte de modelos de impostor de igual duración o calidad que el locutor que se propone como continuación de este proyecto.

GLOSARIO

CMN
Cepstral Mean Normalization, 18, 41, 44, 52

DCT
Discrete Cosin Transform, 33

DET
Detection Error Trade-off, xi, xiv, 25, 54, 55, 56, 57, 63, 76, 77, 80, 105, 109

EER
Equal Error Rate, xii, xiii, xv, xvi, 25, 53, 54, 55, 59, 60, 63, 76, 77, 80, 83, 86, 89, 92, 93, 103, 105, 106, 107, 108, 112, 113, 114, 115, 117, 118, 119, 122, 123, 124

EM
Expectation Maximization, 38, 52

FM
Feature Mapping, 42, 44

FW
Feature Warping, 42, 44

GMM
Gaussian Mixture Model, 26, 34, 36, 37, 38, 40, 42, 43, 44, 51, 52, 55, 67

HMM
Hidden Markov Models, 34, 35, 36

IBM
International Business Machines, 7

ITU
International Telecommunication Union, 4, 65, 67, 83

JFA
Joint Factor Analysis, 42, 43, 44

KCEP
Kurtosis Cepstral, v, xiii, 4, 65, 67, 68, 69, 70, 71, 72, 73, 74, 75, 86, 87, 88, 111, 112, 113, 114, 115, 123, 124

KLPC
Kurtosis of Linear Prediction Coefficients, v, xiii, 4, 65, 67, 68, 69, 70, 71, 72, 73, 74, 75, 89, 90, 91, 92, 111, 112, 113, 114, 115, 123, 124

LPC
Linear Predictive Coefficients, 30, 33, 35, 39, 68

MAP
Maximum a Posteriori, 38, 52

MFCC
Mel-Frecuency Cepstral Coefficients, , xi, 30, 31, 33, 35, 39, 51, 68

NAP
Nuisance Attribute Projection, 43, 44, 52

NIST
National Institute of Standards and Technology, v, viii, xii, xiii, xv, 2, 3, 4, 15, 44, 45, 47, 48, 49, 50, 51, 52, 53, 55, 63, 65, 69, 70, 71, 72, 73, 74, 75, 92, 110, 111, 119, 124, 128, 129, 132, 133

Norm
Normalization, 2, 45, 46, 52, 53, 54, 93, 94, 108, 122, 123, 125

PFA
Probabilidad de falsa aceptación, 24, 25, 58

PFR
Probabilidad de falso rechazo, 24, 25, 58

SNR
Signal to Noise Ratio, v, xiii, 4, 65, 66, 67, 69, 70, 71, 72, 73, 74, 75, 76, 80, 81, 82, 86, 89, 92, 111, 112, 113, 114, 115, 116, 117, 123, 124, 125

SRAL
Sistema de Reconocimiento Automático de Locutor, 1, 29, 35

SRE
Speaker Recognition Evaluation, v, viii, xii, xiii, xv, 2, 3, 4, 15, 45, 47, 48, 49, 50, 51, 52, 55, 63, 69, 70, 71, 72, 73, 74, 75, 92, 110, 111, 119, 124

SV
SuperVector, 40

SVM
Support Vector Machine, 26, 34, 39, 40, 43, 44, 131

T-Norm
Test-Normalization, 45, 46

UBM
Universal Background Model, xi, 37, 38, 55, 69

UBML
Universal Background Model Likelihood, v, xiii, xvi, 4, 65, 67, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 86, 89, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 123, 124, 125

Z-Norm
Zero-Normalization, 45, 46, 93, 122

REFERENCIAS

1. **Sergio Pérez Gómez, Daniel Ramos Castro, Joaquín González Rodríguez y Julián Fierrez.** Modelos de Regresión Logística a Nivel de Puntuaciones para Incorporar Calidad en la Verificación de Locutor. Huesca, España : V Jornadas de Reconocimiento Biométrico de Personas, Septiembre, 2010.

2. **Sergio Perez-Gomez, Daniel Ramos, Javier Gonzalez-Dominguez and Joaquin Gonzalez-Rodriguez.** Score-level Compensation of Extreme Speech Duration Variability in Speaker Verification. Japan : INTERSPEECH, September, 2010.

3. **NIST.** Speaker Recognition Evaluation Plan. [En línea] 2008. http://www.nist.gov/speech/tests/sre/2008/index.htm.

4. **NIST**. Speaker Recognition Evaluation Plan. [En línea] 2006. http://www.nist.gov/speech/tests/sre/2006/index.htm.

5. **J. Pelecanos, U. Chaudhari, and G. Ramaswamy.** Compensation of utterance length for speaker verification. Toledo, Spain : Proc. of Odyssey 2004, 2004, págs. 161-164.

6. **A. Harriero.** Fiabilidad en sistemas forenses de reconocimiento de locutor explotando la calidad de la señal de voz. s.l. : Proyecto Fin de Carrera, Febrero, 2010.

7. **NIST speech group website.** [En línea] http://nist.gov/itl/iad/mig/sre.cfm.

8. **R. Auckenthaler, M. Carey, and H. Lloyd-Tomas.** Score normalization for text-independent speaker verification systems. s.l. : Digital Signal Processing, 2000, Vol. 10, págs. 42-54.

9. **P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel.** A study of interspeaker variability in speaker verification. s.l. : IEEE Trans. on Audio, Speech and Language Processing, 2008, Vol. 16, págs. 980–988.

10. **L. Ferrer,M. Graciarena, A. Zymnis, and E. Shriberg.** System combination using auxiliary information for speaker verification. Las vegas, Nevada, USA : Proc. of ICASSP, 2008, págs. 4853–4856.

11. **N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Scwartz, and A. Strasheim.** Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. s.l. : IEEE Transactions on Audio, Speech and Signal Processing, 2007, Vol. 15, págs.

2072–2084.

12. **J. L. Wayman, A. K. Jain, D. Maltoni and D. Maio.** *Biometric Systems: Technology, Desing and Performance Evaluation.* s.l. : Springer, 2006.

13. **A. K. Jain and A. Ross.** Introduction to Biometrics. *Handbook of Biometrics.* s.l. : Springer, 2008, 1.

14. **D. Maltoni, D. Maio, A. K. Jain and S. Prabhakar.** *Handbook of Fingerprint Recognition.* s.l. : Springer-Verlag, 2003.

15. **C. Wilson, A.R. Hicklin, M. Bone, H. Korves, P. Grother, B. Ulery, R. Micheals, M. Zoepfl, S. Otto and C. Watson.** Fingerprint Vendor Technology Evaluation 2003: Summary of Results and Analysis Report. s.l. : NIST Technical Report NISTIR 7123, June, 2004.

16. **J. Galbally, R. Cappelli, A. Lumini, G. Gonzalez-de-Rivera, D. Maltoni, J. Fierrez, J. Ortega-Garcia and D. Maio.** An Evaluation of Direct Attacks Using Fake Fingers Generated from ISO Templates. s.l. : Pattern Recognition Letters, June, 2010, Vol. 31, págs. 725-732.

17. **S. Z. Li and A. K. Jain.** Handbook of Face Recognition. s.l. : Springer-Verlag, 2005.

18. **R. Zunkel.** Hand Geometry Based Authentication. *Biometrics: Personal Identification in Networked Society.* s.l. : Kluwer Academic Publishers, 1999, págs. 87-102.

19. **J. Daugman.** Recognizing Persons by Their Iris Patterns. *Biometrics: Personal Identification in Networked Society.* s.l. : Kluwer Academic Publishers, 1999, págs. 103-122.

20. **J. P. Campbell.** Speaker Recognition: a Tutorial. s.l. : Proceedings of the IEEE, September, 1997, Vol. 85, págs. 1437-1462.

21. **A. H. Choi and C. N. Tran.** Hand Vascular Pattern Technology. *Handbook of Biometrics.* s.l. : Springer, 2008, págs. 253-270.

22. **D. J. Hurley, B. Arbab-Zavar, M. S. Nixon.** The Ear as a Biometric. *Handbook of Biometrics.* s.l. : Springer, 2008, págs. 131-150.

23. **L. Lee, T. Berger and E. Aviczer.** Reliable On-line Human Signature Verification Systems. s.l. : IEEE Transactions on Pattern Analysis and Machine Intelligence, June, 1996, Vol. 18, págs. 643-647.

24. **F. Monrose and A. Rubin.** Authentication Via Keystroke Dynamics. Zurich, Switzerland : Proceedings of Fourth ACM Conference on Computer and Communications Security, April, 1997, págs. 48-56.

25. **M. S. Nixon, J. N. Carter, D. Cunado, P. S. Huang and S. V. Stevenage.** Automatic Gait Recognition. *Biometrics, Personal Identification in Networked Society.* s.l. : Kluwer Academic Publishers, 1999, págs. 231-249.

26. **International Biometric Group.** Biometrics Revenues by Technology. [En línea] 2006. http://www.biometricgroup.com/.

27. **A. A. Ross, K. Nandakumar and A. K. Jain.** Biometrics: When Identity Matters. *Handbook of Multibiometrics.* s.l. : Springer, 2006, págs. 1-32.

28. **D. A. van Leeuwen and N. Brümmer.** An introduction to application-independent evaluation of speaker recognition systems. *Speaker Classification.* s.l. : Springer, 2007.

29. **D. A. Reynolds.** Channel robust speaker verification via feature mapping. s.l. : IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003, Vol. 2, págs. II–53–6.

30. **A. A. Ross, K. Nandakumar and A.l K. Jain.** Information Fusion in Biometrics. *Handbook of Multibiometrics.* s.l. : Springer, 2006.

31. **J. P. Campbell and D. A. Reynolds and R. B. Dunn.** Fusing high- and low-level features for speaker recognition. s.l. : Proc. of Eurospeech, 2003, págs. 2665-2668.

32. **G. Doddington.** Speaker recognition based on idiolectal differences between speakers. s.l. : Proc. of Eurospeech, 2001, págs. 2517-2520.

33. **D. A. Reynolds, J. P. Campbell, W. M. Campbell, R. B. Dunn, T. P. Gleason, D. A. Jones, T. F. Quatieri, C. B. Quillen, D. E. Sturim and P. A. Torres-Carrasquillo.** *Beyond Cepstra: Exploiting High-Level Information in Speaker Recognition.* s.l. : Proc. Workshop on Multimodal User Authentication, 2003. págs. 223-229.

34. **J. R. Deller, J. H. L. Hansen and J. L. Proakis.** *Discrete-time processing of speech signals, 2nd Ed.* s.l. : John Wiley and Sons, 1999.

35. **D. Ramos.** *Forensic evaluation of the evidence using automatic speaker recognition systems.* Madrid, Spain : s.n., 2007. Available at http://atvs.ii.uam.es.

36. **L. R. Rabiner.** A tutorial on hidden Markov models and selected applications in speech recogniton. s.l. : Proceedings of the IEEE, 1989, Vol. 77, págs. 257-286.

37. **C. M. Bishop.** Pattern Recognition and Machine Learning. s.l. : Springer, 2006.

38. **C. Cortes and V. Vapnik.** Support-Vector Networks. s.l. : Kluwer, 1995, Vol. 20, págs.

273-297.

39. **W. Wan and W. Campbell.** Support vector machines for speaker verification and identification. s.l. : Proc. of IEEE International Workshop on Neural Networks for Signal Processing, 2000, págs. 775-784.

40. **W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer and P. A. Torres-Carrasquillo.** Support vector machines for speaker and language recognition. s.l. : Computer Speech and Language, 2006, Vol. 20, págs. 210-229.

41. **S. Furui.** Cepstral analysis technique for Automatic Speaker verification. s.l. : IEEE Trans. Acoust. Speech, Signal Processing, 1981, Vol. 29, págs. 254-272.

42. **H. Hermansky and N. Morgan.** Rasta Processing of Speech. s.l. : IEEE Transactions on Speech and Audio Processing, special issue on Robust Speech Recognition, October 1994, Vol. 2, págs. 578-589.

43. **J. Pelecanos and S. Sridharan.** Feature warping for robust speaker verification. s.l. : Proc. of Odyssey, 2001, págs. 213-218.

44. **D. A. Reynolds.** Channel robust speaker verification via feature mapping. s.l. : Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),, 2003, Vol. 2, págs. 53-56.

45. **A. Solomonoff, W. M. Campbell, and I. Boardman.** Advances in channel compensation for SVM speaker recognition. s.l. : Proc. of ICASSP, 2005, págs. 629-632.

46. **R. Vogt, C. Lustri and S. Sridharan.** Factor Analysis Modelling for Speaker Verification with Short Utterances. s.l. : Proc. of Odyssey, 2008.

47. **A. A. Ross, K. Nandakumar and A. K. Jain.** Score level fusion. *Handbook of Multibiometrics.* págs. 91-141.

48. **K. P. Li and J. E. Porter.** Normalizations and selection of speech segments for speaker recognition scoring. s.l. : Proc. of ICASSP, 1988, págs. 595-598.

49. **F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier and I. Magrin-Chagnolleau.** A Tutorial on Text-Independent Speaker Verification. s.l. : Journal on Applied Signal Processing, 2004, Vol. 4, págs. 430-451.

50. **D. Sturim and D. A. Reynolds.** Speaker Adaptive Cohort Selection for TNorm in Text-independent Speaker Verification. s.l. : Proc. of ICASSP, 2005, págs. 741-744.

51. **J. Gonzalez-Rodriguez and D. Ramos.** Forensic Automatic Speaker Classification in the

Coming Paradigm Shift. *Speaker Classification.* s.l. : Springer, 2007, Vol. 4343.

52. **F. Alonso-Fernandez, J. Fierrez, J. Ortega-Garcia, J. Gonzalez-Rodriguez, H. Fronthaler, K. Kollreider and J. Bigun.** A comparative study of fingerprint image-quality estimation methods. s.l. : IEEE Trans. on Information Forensics and Security, December, 2007, Vol. 2, págs. 734–743.

53. **V. Grancharov and W. B. Hleijn.** Speech Quality Assessment. *Handbook of Speech Processing.* s.l. : Springer, 2007, 5.

54. **NIST.** *Biometric sample quality standard draft (revision 4), 6.* 2008.

55. **P. Grother and E. Tabassi.** Performance of biometric quality measures. s.l. : IEEE Trans. Pattern Anal. Mach. Intell, 2007, Vol. 29, págs. 531–543.

56. **D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez and J. Ortega-Garcia.** Using Quality Measures for Multilevel Speaker Recognition. 2005.

57. **A. Harriero, D. Ramos and J. Gonzalez-Rodriguez.** Analysis of the utility of classical and novel speech quality measures for speaker verification. s.l. : Proceedings of International Conference on Biometrics, Springer, June, 2009, Vol. 5558 of LNCS, págs. 434–442.

58. **J. Richiardi and A. Drygajlo.** Evaluation of speech quality measures for the purpose of speaker verification. s.l. : Proc. of Odyssey, 2008.

59. **F. Alonso-Fernandez.** Biometric Sample Quality and its Application to Multimodal Authentication Systems. October, 2008.

60. **N. Brümmer and J. du Preez.** Application independent evaluation of speaker detection. s.l. : Computer Speech and Language, 2006, Vol. 20, págs. 230–275.

61. **B. Efron and R. J. Tibshirani.** An introduction to the Bootstrap. s.l. : Chapman & Hall, 1993.

62. **Christopher M. Bishop.** *Pattern Recognition and Machine Learning.* s.l. : Springer, 2006.

63. **R. O. Duda, P. E. Hart and D. G. Stork.** Pattern Classification. s.l. : Wiley, 2001.

64. **S. Pigeon, P. Druyts and P. Verlinde.** Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions. s.l. : Digital Signal Processing, 2000, Vol. 10, págs. 237-248.

65. **J. R. Deller and J. H. L. Hansen and J. L. Proakis.** *Discrete-time processing of speech signals, 2nd Ed.* s.l. : John Wiley and Sons, 1999.

66. **P. Van Kerm.** Adaptative Kernel Density Estimation. 2003.

ANEXO: PUBLICACIONES

Este trabajo ha generado 2 publicaciones que han sido aceptadas en congresos de interés nacional (1) e internacional (2), ambas con revisión científica por expertos en la materia.

1. **Sergio Pérez Gómez, Daniel Ramos Castro, Joaquín González Rodríguez y Julián Fierrez.** Modelos de Regresión Logística a Nivel de Puntuaciones para Incorporar Calidad en la Verificación de Locutor. Huesca, España : V Jornadas de Reconocimiento Biométrico de Personas, Septiembre, 2010.

2. **Sergio Perez-Gomez, Daniel Ramos, Javier Gonzalez-Dominguez and Joaquin Gonzalez-Rodriguez.** Score-level Compensation of Extreme Speech Duration Variability in Speaker Verification. Japan : INTERSPEECH., September, 2010.

Modelos de Regresión Logística a Nivel de Puntuaciones para Incorporar Calidad en la Verificación de Locutor

Sergio Pérez Gómez, Daniel Ramos Castro, Joaquín González Rodríguez y Julián Fierrez*

ATVS - Grupo de Reconocimiento Biométrico
Escuela Politécnica Superior, Universidad Autónoma de Madrid
Calle Francisco Tomás y Valiente 11, 28049, Madrid, España
{sergio.perez,daniel.ramos,joaquin.gonzalez,julian.fierrez}@uam.es
http://atvs.ii.uam.es/

Resumen El presente artículo es un estudio preliminar en el que se evalúan 3 métodos propuestos para compensar la pérdida de rendimiento en los sistemas de verificación de locutor debido a la variabilidad en la calidad de las muestras de audio estudiada. Estos algoritmos están basados en otros definidos en la literatura y serán evaluados mediante la mejora en términos de EER (*Equal Error Rate*) de un sistema GMM-UBM entrenado y testeado mediante la condición *short2-short3* de la evaluación NIST SRE 2008 (*Speaker Recognition Evaluation*) [1]. La mejora relativa de EER más significativa obtenida es del 8,82 %, hallada mediante el algoritmo de *Regresión Logística Lineal de 2 Dimensiones* (2D-LLR) para la condición *tel-mic* de NIST SRE 2008, utilizando UBML (*Universal Background Model Likelihood*) como información de calidad, aunque también se han explorado resultados con la *Relación Señal a Ruido* del mismo modo. Aunque los resultados son preliminares y optimistas, puesto que los algoritmos se han entrenado con los mismos datos que la evaluación, se observa una clara y relevante tendencia de mejora.

Keywords: verificación de locutor, calidad, utilidad, rendimiento, UBML, SNR, regresión logística.

1. Introducción

La idea de que la calidad de una muestra de voz puede afectar al rendimiento de un sistema de reconocimiento automático de locutor es bastante intuitiva [2]. De hecho la medida y compensación de la calidad de una señal de audio ha sido una tarea en el que se ha invertido un gran esfuerzo en el ámbito científico biométrico en los últimos años [3]. Inicialmente este esfuerzo viene por la

* Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación (TEC2009-14719-C02-01) y la cátedra UAM-Telefónica.

necesidad de controlar la calidad de la voz en las redes telefónicas pero en la actualidad se ha transformado en la definición de medidas de calidad y algoritmos de calibración que permitan predecir el rendimiento de un sistema biométrico.

Si bien es cierto que existen técnicas como *factor analysis* que reducen de forma significativa la variabilidad introducida por el canal [4] estas técnicas dependen en gran medida de la existencia de un corpus apropiado, deseablemente con las mismas condiciones de la voz a reconocer. Sin embargo, estos modelos son dependientes de los datos utilizados para su entrenamiento. La calidad de voz está basada en el conocimiento de la señal de voz mediante la cual se puede predecir tanto el rendimiento de un sistema de verificación de locutor como un posible desalineamiento de las puntuaciones del mismo debido a cambios en dicha calidad. En este trabajo se propone el uso de la información de calidad para ajustar el desalineamiento entre las puntuaciones *target* y *non target* de un sistema de verificación de locutores, mediante modelos de regresión logística.

Esta investigación se inicia en el estudio de distintas medidas de calidad definidas y su utilización para compensar variabilidad intersesión a nivel de locución. De entre todas ellas se ha procedido a elegir 2 que definen de manera más significativa que el resto el poder discriminativo del sistema [5]: la SNR (*Signal To Noise Ratio*) y la UBML (*Universal Background Model Likelihood*). Una vez estudiado el comportamiento del sistema definido en 5.2 frente a estos indicadores de degradación de rendimiento se han diseñado 3 algoritmos basados en un modelo de regresión logística [6]: *Regresión Logística Lineal de 2 Dimensiones* (2D-LLR) y *Regresión Logística Bilineal* (BLR tipo 1 y tipo 2), que evaluarán dicho rendimiento sobre la base de datos de NIST SRE 2008 [1] que presenta un desafío de variabilidad intersesión [7].

Este artículo está organizado de la siguiente manera: la sección 2 presenta la motivación de este trabajo así como la definición de las medidas de calidad que determinan el rendimiento del sistema a mejorar. En la sección 4 se describen los métodos propuestos de compensación aplicados sobre la base de datos y el sistema descrito en la sección 5, cuyos resultados y conclusiones son ampliamente analizados en 6 y 7.

2. Tratamiento e interpretación de medidas de calidad

La calidad de una muestra biométrica viene definida por tres criterios básicos según [8], un borrador estándar de calidad según NIST [9] de dicho tipo de muestras:

- La fidelidad, que se refiere a la exactitud y precisión con la que una muestra biométrica es capturada, procesada y almacenada en el sistema.
- El carácter, entendido como la actitud o predisposición del usuario a que se capture su muestra biométrica (factores conductuales).
- La utilidad, definida como la característica para evaluar y predecir el rendimiento de un sistema de reconocimiento biométrico.

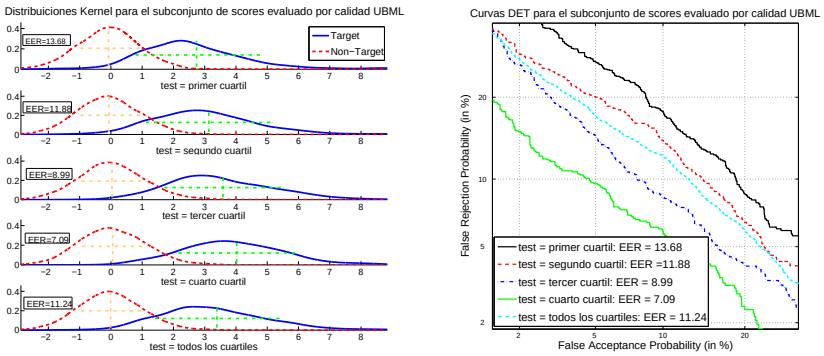


Figura 1. (Izquierda) Representación de las distribuciones *target* y *non target* en función del cuartil elegido para la calidad UBML del fichero de test mostrando su consecuente desalineamiento. Cabe destacar que la calidad del modelo no se tiene en cuenta en esta representación (los subconjuntos de scores han sido agrupados por calidad de test sin importar la del modelo). (Derecha) Curvas DET (*Detection Error Trade-off*) equivalentes a la figura anterior.

Si bien es interesante estudiar estos tres criterios que definen de forma concisa la calidad relativa de una muestra y por lo tanto el rendimiento de un sistema, este trabajo pretende evaluar mediante la *utilidad* [10] el rendimiento global del sistema bajo los diferentes algoritmos propuestos en la sección 4. Por lo tanto, es importante remarcar que los resultados obtenidos fruto de este trabajo (sección 5) dependerán en gran medida de la disponibilidad de unas buenas condiciones de fidelidad (tipo de dispositivo de adquisición, supervisión en la adquisición, tasa de compresión, etc.), del carácter del individuo (cooperación del sujeto, estado emocional, etc.) y de una base de datos apropiada, es decir, de máxima variabilidad y con un número amplio de muestras.

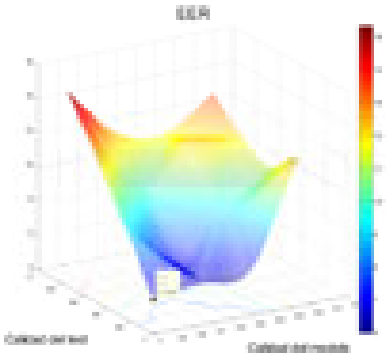


Figura 2. EER en función de la calidad del modelo y del test para el subconjunto de scores dependientes de la calidad UBML condición *tel-mic*.

Por lo tanto, y habiendo enunciado algunos hechos determinantes que influyen en la degradación de la calidad de una muestra de voz, se han seleccionado 2 tipos de indicadores por su impacto o dependencia con el rendimiento del sistema y por su coherencia con diferentes bases de datos y sistemas [5]: la *Relación Señal a Ruido* (SNR) y la *Universal Background Model Likelihood* (UBML) definida recientemente en [5].

El efecto de esta última se puede apreciar en la figura 2, en la que se observa la tendencia del EER en función de la calidad UBML del modelo y del test: a medida que la calidad el sistema aumenta la discriminación del sistema mejora. Como se especificará en 3.2 y se propone en [10] la calidad debe ser mapeada entre 0 y 1: dado que el número de ficheros con una calidad extrema es limitado (existen muy pocos ficheros de voz con muy buena o muy mala calidad) y estos son requeridos para la evaluación del rendimiento de los algoritmos propuestos se ha procedido a homogeneizar el número de ficheros ordenando los archivos por valor de calidad y tomando más o menos ficheros de una calidad determinada hasta alcanzar el porcentaje de ficheros correspondiente al 25 %, del 25 % al 50 %, del 50 % al 75 % y de éste al 100 % (4 cuartiles) de la calidad total. Por lo tanto quedarán definidos 4 bloques de calidad representados como 0,25 para el primer cuartil y 0,50, 0,75 y 1 para los restantes.

Dado que las bases de datos existentes en general no son lo suficientemente ricas en cuanto a variabilidad para implementar técnicas como ésta se recurre a otros métodos de normalización y calibración como la regresión logística, que tratan de calibrar el sistema en función de la calidad del enfrentamiento para reducir el desalineamiento de las distribuciones *target* y *non target* como se representa en la figura 1. Nótese que las distribuciones nombradas hacen referencia a un sistema automático de reconocimiento trabajando en modo verificación en el que se cumple la hipótesis de que el usuario y el fragmento de voz a evaluar son la misma persona y en el que no, de forma respectiva.

3. Medidas de calidad empleadas

3.1. Relación señal a ruido (SNR)

Como su nombre indica la SNR expresa la relación entre la potencia de la señal de voz y la potencia de ruido que la corrompe. Por lo tanto, queda definida mediante la siguiente fórmula:

$$SNR = 10 \log \left(\frac{E_{voz}}{E_{silencio}} \right) \tag{1}$$

siendo E_{voz} y $E_{silencio}$ la energía media de las zonas de voz y silencio respectivamente del fragmento de audio.

El principal problema de utilizar esta medida de calidad es que la fiabilidad de ésta dependerá de la precisión del detector de actividad de voz (VAD) siendo una no muy buena referencia de calidad si el diseño de éste no es el apropiado. No obstante es una medida ampliamente extendida y utilizada.

Siguiendo las recomendaciones de [10] y [11], para trabajar de forma homogénea con cualquier tipo de calidad es necesario expresar todo indicador de degradación en un rango entre 0 y 1 mediante una función de mapeo $Q(x)$, siendo 0 el valor mínimo de calidad y 1 el máximo [5]:

$$Q_{SNR}(x) = \frac{x}{60} \tag{2}$$

donde x corresponde al valor de SNR obtenido en un rango de $(0-60)dB$.

3.2. Similitud a un modelo de habla universal (UBML)

La UBML es una medida de calidad que trata de aproximar la similitud de una locución al modelo de habla universal utilizado para la generación del modelo estadístico de un locutor. Es una medida considerada de forma reciente en [5] que se extrae en los sistemas GMM (*Gaussian Mixture Model*) para calcular la puntuación de similitud:

$$S(O,\lambda_t) = log\left(p(O,\lambda_t)\right) - log\left(p(O,\lambda_{UBM})\right) \tag{3}$$

donde $S(O,\lambda_t)$ es el score o puntuación de similitud entre el modelo del locutor y el modelo universal y $p(O,\lambda_t)$ y $p(O,\lambda_{UBM})$ son las funciones densidad de probabilidad del modelo de usuario y universal respectivamente. Por lo tanto, la UBML queda definida mediante:

$$UBML = log\left(p(O,\lambda_{UBM})\right) \tag{4}$$

y cuya función de mapeo es:

$$Q_{UBML}(x) = \frac{x+13}{8} \tag{5}$$

donde x corresponde al valor de la $UBML$ obtenido en el rango de $(-13,-5)$.

4. Algoritmos de compensación propuestos

Los métodos que en esta sección se describen pretenden compensar el rendimiento del sistema a nivel de score dado el desalineamiento de las distribuciones (figura 1) como efecto de la variabilidad de la calidad explicado en la sección anterior. Para ello se han evaluado 3 algoritmos diferentes basados en regresión logística, ya utilizada en reconocimiento de locutor para fusión y calibración [6][12][13].

4.1. Regresión logística lineal de dos dimensiones (2D-LLR)

El método LLR (*Linear Logistic Regression*) es un algoritmo de compensación que transforma el conjunto de puntuaciones generadas bajo unas condiciones de calidad mediante un modelo de regresión lineal en el logaritmo de una relación de verosimilitud [6], el cual puede definirse de la siguiente manera:

$$x_{i,j}^{Norm} = \log \frac{P_{i,j}(x_{i,j}|T)}{P_{i,j}(x_{i,j}|NT)} = \alpha_{i,j} \cdot x_{i,j} + \beta_{i,j} \tag{6}$$

Los pesos $\alpha_{i,j}$ y $\beta_{i,j}$ se obtienen de las puntuaciones de entrenamiento [6][1] de las comparaciones de los modelos del cuartil de calidad i con los ficheros de test de calidad del cuartil j. La puntuación o score a compensar $x_{i,j}$ presenta el mismo cuartil de calidad que las puntuaciones de entrenamiento. Por lo tanto, ya que dicho algoritmo es dependiente de la calidad del modelo y del test se puede decir que la regresión logística seguida presenta dos dimensiones dando de esta manera nombre al algoritmo implementado.

Por último. destacar que este algoritmo se puede generalizar para cualquier valor de calidad realizando algún tipo de interpolación de los pesos $\alpha_{i,j}$ y $\beta_{i,j}$ (por ejemplo cúbica).

4.2. Regresión logística bilineal (BLR)

Este método está basado en [12], donde se tomaba como información complementaria información lingüística del locutor y puede definirse como sigue:

$$x_{i,j}^{Norm} = \alpha \cdot x_{i,j} + \sum_{k=1}^{K} \alpha_k \cdot \lambda_k \cdot x_{i,j} + \beta \tag{7}$$

donde α, α_k y β son ahora pesos fijos para todos los posibles conjuntos de scores dependientes de calidad, y λ_k corresponde a la información de las calidades del modelo y del test. Para la realización de este trabajo se ha definido dicha información de dos maneras diferentes dando lugar a dos algoritmos diferentes:

- BLR tipo 1: $\lambda_1 = Q_m$ y $\lambda_2 = Q_t$, donde la información complementaria corresponde a la calidad mapeada del modelo Q_m junto con la del test Q_t.
- BLR tipo 2: $\lambda_1 = \sqrt{Q_m \times Q_t}$, donde la información complementaria corresponde con la media geométrica de las calidades.

5. Experimentos

5.1. Base de datos y protocolos

El organismo norteamericano NIST (*National Institute of Standards and Technology*) [9] organiza evaluaciones bianuales abiertas de carácter competitivo en el que se elaboran bases de datos y se definen una serie de tareas o protocolos para medir de manera objetiva el rendimiento, bajo las mismas condiciones, de los sistemas presentados. La base de datos utilizada para el desarrollo de este trabajo es la de NIST SRE 2008 [1], la cual ofrece un importante desafío en cuanto a compensación de variabilidad de calidad se refiere. El protocolo seguido para la realización de estos experimentos es el *short2-short3* de la misma evaluación, que comprende archivos de audio conversacionales capturados de un

[1] El toolkit FoCal ha sido usado para el entrenamiento de la LLR. http://sites.google.com/site/nikobrummer/focal

canal telefónico o microfónico de 5 minutos de duración, de los cuales aproximadamente 2.5 minutos son de cada locutor de la conversación una vez suprimidos los silencios, y datos microfónicos en formato *interview* o entrevista en los cuales la mayoría de audio de los 3 minutos de duración corresponde al entrevistado.

El trabajo llevado a cabo presenta 4 condiciones de evaluación o escenarios diferentes:

- *tel-tel*, en el que el modelo de entrenamiento y el fichero de test han sido adquiridos de un canal telefónico.
- *tel-mic*, en el que el fichero de audio con el que se entrena el modelo del usuario a identificar ha sido adquirido a través de un canal telefónico y el fichero de test mediante un micrófono.
- *mic-tel*, en el que el modelo se ha extraído de una grabación con micrófonos y el archivo de enfrentamiento se ha capturado a través de la red telefónica.
- *mic-mic*, en la que el modelo y fichero de test han sido capturados con un dispositivo microfónico.

Nótese que las grabaciones telefónicas presentan diversos factores de degradación de la señal principalmente relacionados con la distorsión del canal de comunicación mientras que las muestras microfónicas se ven más afectados mediante otros parámetros como la fidelidad del dispositivo de captura, la distancia al micrófono, las condiciones ambientales, etc.. Tanto los ficheros de entrenamiento como los de enfrentamiento, presentan un filtrado Wiener ya que se ha demostrado que este tipo de procesado ayuda a mejorar el rendimiento de sistemas que trabajan con este tipo de muestras.

Por último, remarcar que dado el carácter preliminar del estudio los algoritmos propuestos se han entrenado con los mismos datos que la evaluación.

5.2. Sistema de desarrollo

El sistema de desarrollo sobre el cual se ha medido el rendimiento de los 3 algoritmos propuestos es el sistema presentado por el grupo ATVS en la evaluación NIST SRE de 2008. Este sistema está basado en modelo GMM (*Gaussian Mixture Model*) de 1024 mezclas y 19 parámetros MFCC (*Mel Frecuency Cepstral Coefficients*), adaptados de un UBM (*Universal Background Model*) entrenado con una gran cantidad de datos provenientes de las evaluaciones NIST hasta NIST SRE 2006. Dicho sistema también incluye compensación de canal mediante técnicas basadas en *feature warping* y *factor analysis* aplicando adaptación NAP (*Nuisance Attribute Projection*) [14] a los modelos GMM. Por último, destacar el uso de *T-Norm* para normalizar los scores [15].

6. Resultados

El cuadro 1 muestra en términos de EER la mejora de los algoritmos propuestos en función de la condición de NIST SRE 2008 evaluada. El método

2D-LLR que realiza una LLR por subconjunto de scores en función de la calidad del modelo Q_m y la calidad del test Q_t ofrece prácticamente en todos los casos una mejora en términos de EER en cuanto al EER original y al EER mejorado por los otros algoritmos. En cualquier caso, sólo sufre un empeoramiento para la condición *tel-tel* de UBML, el cual puede ser considerado despreciable por su proximidad a 0. No obstante. los métodos BLR-1 y 2 ofrecen peores resultados salvo para esta misma condición, en el que presentan una mejora por encima del resultado obtenido usando 2D-LLR siendo mejor para el segundo caso en el que la información complementaria es la media geométrica de las calidades del modelo y del test.

Los mismos resultados son representados de forma gráfica en la figura 3, en las que se han separado para los 3 algoritmos las 2 condiciones con peor rendimiento, *tel-tel* y *mic-tel* y las 2 con mejor rendimiento, *tel-mic* y *mic-mic*.

			UBML		**SNR**	
Compensación	Condición	EER	EER_{norm}	Mejora EER	EER_{norm}	Mejora EER
2D-LLR	tel-tel	7, 80	7, 82	−0, 35	7, 64	1, 97
	tel-mic	11, 24	10, 24	**8,82**	10, 45	**7,01**
	mic-tel	11, 68	11, 66	**0,16**	11, 60	**0,72**
	mic-mic	8, 51	8, 19	**3,76**	8, 43	0, 94
BLR tipo 1	tel-tel	7, 80	7, 68	1, 51	7, 69	**1,35**
	tel-mic	11, 24	10, 47	6, 79	10, 46	6, 94
	mic-tel	11, 68	11, 77	−0, 81	12, 08	−3, 46
	mic-mic	8, 51	8, 35	1, 93	8, 40	**1,3**
BLR tipo 2	tel-tel	7, 80	7, 63	**2,08**	7, 69	**1,35**
	tel-mic	11, 24	10, 50	6, 56	10, 50	6, 56
	mic-tel	11, 68	11, 80	−1, 05	12, 07	−3, 38
	mic-mic	8, 51	8, 32	2, 3	8, 40	**1,3**

Cuadro 1. Tabla resumen en la que se muestra la mejora en términos de EER del sistema en tanto por ciento de los algoritmos propuestos para las 4 condiciones y los subconjuntos dependientes de calidad.

7. Conclusiones y trabajo futuro

En este artículo se ha estudiado el rendimiento del sistema, en cuanto a términos de EER se refiere, en función de los subconjuntos de scores dependientes de las calidades UBML y SNR respaldadas en estudios anteriores como buenos indicadores de degradación. Para ello se han implementado varios algoritmos probados sobre NIST SRE 2008 con resultados prometedores.

Resumiendo los resultados obtenidos mediante esta base de datos cabe destacar que la mejora del 8,82 % evaluando los datos a partir de la calidad UBML en la condición *tel-mic* mediante el método 2D-LLR es considerable, lo cual incita a seguir realizando estudios en este sentido, ya que la fiabilidad y carácter de las muestras telefónicas son diferentes a las microfónicas produciendo un mayor desajuste en las distribuciones (ver media de las distribuciones *target* en figura 1). Otra resultado a destacar es que el método BLR-2 mejora en un 2,08 % en la condición *tel-tel*, donde el método 2D-LLR no obtiene buen rendimiento para el subconjunto de scores obtenidos a través de la calidad UBML. Dicho método es una variación elegante del método 2D-LLR aunque con mayores problemas en

cuanto a coste computacional y divergencia. Por último, en cuanto a conclusiones extraídas se refiere, destacar que para un sistema GMM obtener la calidad UBML, en la cual se obtienen mejoras ligeramente superiores a usar la SNR, no implica ningún coste adicional ya que es un hecho que debe realizarse de forma imperativa si se pretende lanzar un score como resultado final a evaluar.

Como trabajo futuro hay que remarcar que sería interesante entrenar los pesos de compensación de las regresiones logísticas mediante un conjunto de scores diferente al de test, no como en este estudio, en el que dichos parámetros contenían información de los datos a normalizar debido a que se ha utilizado la base de datos de 2008 a posteriori. Estudios en esta línea se están llevando a cabo en el ATVS. Otra línea de investigación a seguir sería evaluar el rendimiento de estos algoritmos bajo otras técnicas de normalización como ZT-Norm que trata de normalizar las distribuciones *target* y *non target* a través de los modelos de entrenamiento, o el uso de cohortes dependientes de la calidad a la hora de normalizar.

Referencias

1. NIST, "2008 speaker recognition evaluation plan: http://www.nist.gov/speech/tests/sre/2008/index.htm," 2008.
2. F. Alonso-Fernandez *et al.*, "A comparative study of fingerprint image-quality estimation methods," *IEEE Trans. on Information Forensics and Security*, vol. 2, no. 4, pp. 734–743, December 2007.
3. V.Grancharov *et al.*, *Speech Quality Assessment Chapter 5*, Springer, 2007.
4. P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.
5. Alberto Harriero *et al.*, "Analysis of the utility of classical and novel speech quality measures for speaker verification," in *Proceedings of International Conference on Biometrics*. June 2009, vol. 5558 of *LNCS*, pp. 434–442, Springer.
6. Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
7. M. A. Przybocki, A. F. Martin, and A.N. Le, "NIST speaker recognition evaluations utilizing the mixer corpora-2004, 2005, 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1951–1959, 2007.
8. NIST, "Biometric sample quality standard draft (revision 4), 6.," 2008.
9. "NIST speech group website: http://www.nist.gov/speech," .
10. Patrick Grother and Elham Tabassi, "Performance of biometric quality measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 531–543, 2007.
11. D. Garcia-Romero *et al.*, "Using quality measures for multilevel speaker recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 192–209, 2006.
12. L. Ferrer *et al.*, "System combination using auxiliary information for speaker verification," in *Proc. of ICASSP*, Las Vegas, Nevada, USA, 2008, pp. 4853–4856.
13. N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.
14. A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. of ICASSP*, 2005, pp. 629–632.
15. R. Auckenthaler *et al.*, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

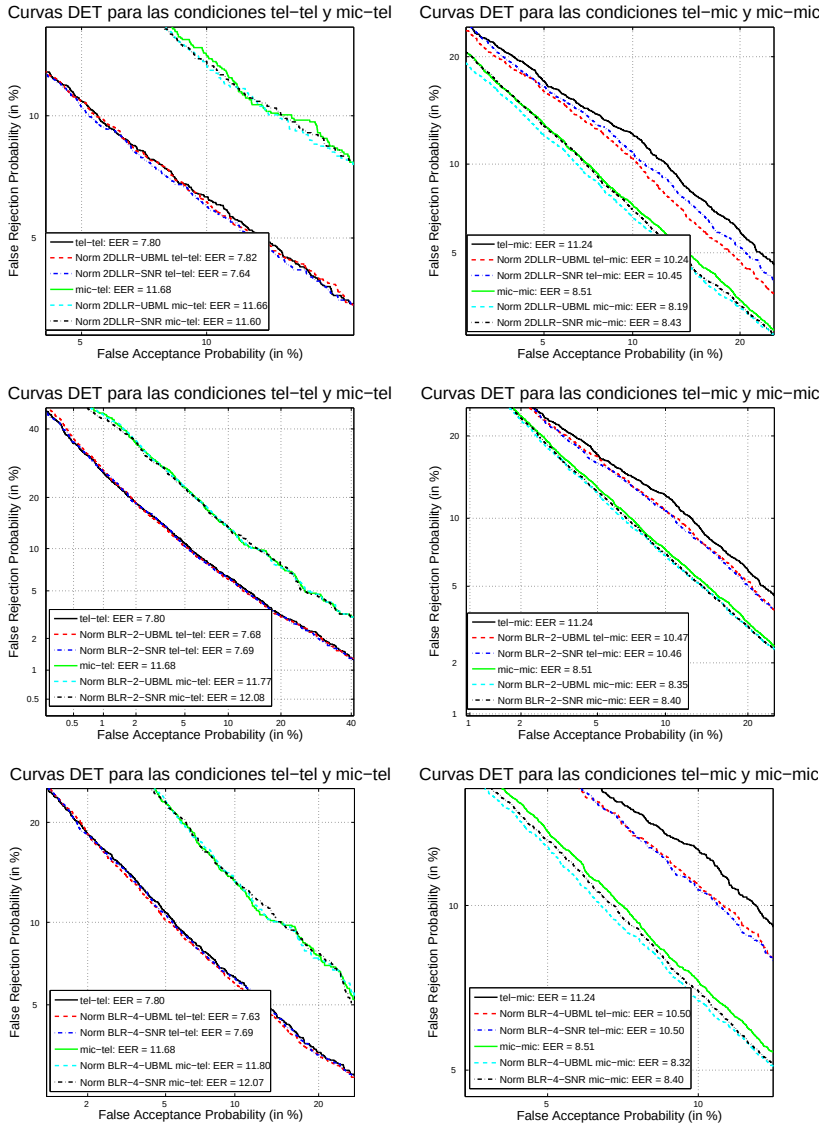


Figura 3. Rendimiento de los algoritmos mostrado en formato curva DET (*Detection Error Trade-off*). A la izquierda se muestran las curvas para las condiciones *tel-tel* y *mic-tel* que corresponden a las de peor mejora y a la derecha las de *tel-mic* y *mic-mic* con resultados prometedores. Notar que la escala en cada gráfica es diferente.

Score-level Compensation of Extreme Speech Duration Variability in Speaker Verification

Sergio Perez-Gomez, Daniel Ramos, Javier Gonzalez-Dominguez and Joaquin Gonzalez-Rodriguez

ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
sergio.perez@uam.es, daniel.ramos@uam.es, javier.gonzalez@uam.es, joaquin.gonzalez@uam.es

Abstract

In this work we aim at compensating the degrading effects of utterance length variability of speaker verification systems, which appear in many typical applications such as forensics. The paper concentrates in the score misalignments due to different utterance lengths, proposing several algorithms for its normalization. In order to test the proposed methods, we have built two corpora from NIST SRE 2006 and 2008 data to simulate high utterance length variability. Results show an improvement of the overall system performance for all the algorithms proposed, which is significant even when score normalization techniques such as T-Norm are used.

1. Introduction

Speaker verification aims at comparing two sets of speech data with the aim of discriminating among *target* comparisons where the identity of both is the same; and *non-target* comparisons where they are not. One of the biggest challenges affecting performance in automatic speaker verification is related to the variability on the conditions of the speech signal, due to multiple factors which may change in different recording sessions. In fact, compensation of such session variability has been the main technology improvement in speaker recognition in the last ten years, as testified by NIST Speaker Recognition Evaluations (SRE)[1]. One of this degrading variability factors is the length of the speech utterances involved in enrollment and testing processes [1]. However, nowadays little research has been conducted for compensating the effects of speech duration variability, or even analyzing their impact. This is mainly due to the configuration of tasks in NIST SRE, where the length of the enrollment and testing utterances present small variation in a single condition. However, there is a plethora of operational scenarios where the length of the utterance involved in the recognition process may vary, *e.g.* forensic applications [2] where the variation in the length of questioned speech (test segment) is virtually unpredictable in real casework.

A typical effect of utterance length variability is a variable score misalignment, causing a global degradation of performance when scores obtained with utterances of different duration are pooled together. While this can be partially solved with score normalization [3], the typical use of fixed cohorts for T-Norm or Z-Norm does not completely eliminate the effects of duration variability. Therefore, we propose a modified fusion/calibration stage, present in most speaker verification systems, which incorporates duration information. As we show in this paper, the proposed algorithms improve the performance of the system both when there is a previous score normalization stage or not, transforming the scores into log-likelihood-ratios which can be interpretable under a Bayesian decision framework [4]. Some of those techniques have been previously proposed in the literature [1, 5], but here we also propose several novel methods based on Gaussian models and logistic regression. In order to test the approaches proposed, we have generated two corpora from the telephone-only speech in NIST SRE 2006 and 2008 databases, namely DurTelSRE06 and DurTelSRE08. There, we have simulated variability by altering the length of the speech files.

The paper is organized as follows: Section 2 present the motivation of the work, as well as an analysis of the effects of utterance length variability. Section 3 describes the methods used for compensation. In Section 4 the experimental protocol is described, and the results showing the adequacy of the proposed techniques are presented. Finally, conclusions are drawn in Section 5.

2. Variability of utterance length in speaker verification

Figure 1 illustrates the effect of utterance length variability in the range of the speaker recognition scores. The plots show target and non-target distribution of scores generated with the Gaussian Mixture Model system based on Universal Background Modelling (GMM-UBM) presented by ATVS in NIST SRE 2008, and described in Section 4. The database and protocol has been constructed from telephone-only speech of NIST SRE 2008 short2-short3 condition, in order to simulate extreme speech length variability. This corpus has been named DurTelSRE08, and is described in detail in Section 4. Each pair of target and non-target distributions were generated with the same speaker models, enrolled with various utterance lengths; but the test segment length is different for each pair of distributions. The EER of each pair of target and non-target scores is also shown in the plot. It is shown that the range of each pair of distributions varies with utterance length. Although T-Norm has been used for score normalization, it does not completely eliminate this misalignment, mainly due to the use of a fixed cohort, which is the typical configuration. That causes that, although the EER of each set of scores independently reaches relatively low values, its value for all the scores from all durations pooled together is poor. This misalignment not only affects the overall discriminating power of the system, but also represents a problem when a single decision threshold is to be established, or when distributions have to be modelled for calibration or fusion purposes [4]. This motivates the use of the proposed algorithms, which compensate such variations in the score range due to utterance length.

This work has been funded by the Spanish Ministry of Science and Innovation under project TEC2009-14719-C02-01.

[1] http://www.itl.nist.gov/iad/mig/tests/sre/

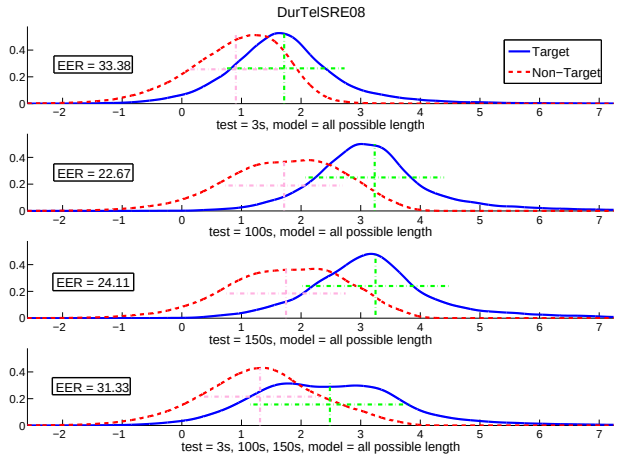


Figure 1: *Kernel distributions of scores generated with 3, 100 and 150 seconds for test segment length and all model lengths for DurTelSRE08 set-up.*

3. Compensation Methods

All the presented algorithms are supervised, meaning that the compensation models should be previously trained with a separate development set of scores, different from the test scores to compensate. For this work we have used scores from two databases built for the paper from NIST SRE data, and described in Section 4, namely DurTelSRE06 for training the algorithms and DurTelSRE08 for testing their performance.

3.1. Length-conditional Gaussian Modelling (GM)

This algorithm transforms the scores generated with given utterance lengths into a log-likelihood ratio, whose distributions are assigned to the training scores corresponding to such lengths. In this way, assuming that score distributions are Gaussian[2] and T and NT respectively correspond to the target and non-target hypotheses, for a given score $x_{i,j}$ having training speech length i and testing speech length j, a new score is calculated as follows:

$$x_{i,j}^{Norm} = \log \frac{P_{i,j}(x_{i,j}|T, \mu_{i,j}^{T}, \sigma_{i,j}^{T})}{P_{i,j}(x_{i,j}|NT, \mu_{i,j}^{NT}, \sigma_{i,j}^{NT})} \quad (1)$$

where, $P_{i,j}$ are Gaussian distributions. The parameters $\mu_{i,j}^{T}$ and $\sigma_{i,j}^{T}$ are respectively the mean and standard deviation computed from target training scores for model speech duration i and test segment speech duration j; and $\mu_{i,j}^{NT}$ and $\sigma_{i,j}^{NT}$ are the corresponding values for non-target training scores. Given that the variability to compensate will be focused on the test segment rather than the training model, a one-dimensional (1D-GM) version of this method has been used as proposed in [1], where the value of j for the testing scores varied, but the value of i was ignored, and all the possible scores from different enrolled model durations were pooled together. Moreover, we also propose the use of the training model duration i, leading to the two-dimensional approach in Equation 1 (2D-GM).

As it can be seen, the possible values of i and j will belong to a discrete set, and therefore it will be difficult to represent all possible continuous utterance lengths in the training scores. For instance, in our experimental set-up (described in Section 4), this discrete set was $\{3, 10, 15, 20, 30, 40, 50, 60, 100, 150\}$, expressed in seconds, both for model enrollment speech and test segments. In more general scenarios, where the testing lengths may have any possible value, possible strategies are the binning of the scores with respect to their utterance lengths, or also some kind of interpolation such as the one performed in [1] using cubic splines.

3.2. Length-conditional Linear Logistic Regression (LLR)

In this method, we handle the training and testing data in a similar way as for length-conditional Gaussian method, but the model used instead of Gaussian densities is linear logistic regression [6], which can be defined in the following way:

$$x_{i,j}^{Norm} = \log \frac{P_{i,j}(x_{i,j}|T)}{P_{i,j}(x_{i,j}|NT)} = \alpha_{i,j} \cdot x_{i,j} + \beta_{i,j} \quad (2)$$

The weights $\alpha_{i,j}$ and $\beta_{i,j}$ are obtained from the training scores [6][3] from comparisons of models enrolled with speech utterances having length i and test segments of length j. The score to compensate $x_{i,j}$ presents the same utterance length conditions as the training scores. Again, we will refer to as 2-dimensional LLR, or 2D-LLR, when the training and testing speech data are respectively restricted to utterance lengths i and j; and 1-dimensional LLR or 1D-LLR when the length of the speech used for test segments j is restricted but model enrollments of all lengths are pooled together. As it happened for GM models, in scenarios with a potentially continuous speech variability, binning of the scores with respect to their utterance lengths, or interpolation of $\alpha_{i,j}$ and $\beta_{i,j}$ using e.g. cubic splines are a solution for achieving generality.

3.3. Bilinear Logistic Regression (BLR)

This method is an elegant technique that optimizes a logistic regression model conditioned to a set of values which represent some form of information additional to the score. A similar method has been previously used in [5] to system fusion using linguistic information. Here we incorporate the information of the utterance length, both for the enrolled model and for the test segment. The method is an extension of LLR, and can be stated as follows:

$$x_{i,j}^{Norm} = \alpha \cdot x_{i,j} + \sum_{k=1}^{K} \alpha_k \cdot \lambda_k \cdot x_{i,j} + \beta \quad (3)$$

where α, α_k and β are now weights which are fixed for all possible values of utterance lengths, and a number of λ_k values in the $[0, 1]$ range are derived from the utterance lengths of model and test segment speech, divided by the maximum length (150 seconds in our case). If such normalized durations are respectively denoted as d_m and d_t, the selection of the λ_k values determine the information used for compensating utterance length, which leads to 3 approaches of BLR used in this paper: BLR-1 when $\lambda_1 = d_t$; BLR-2 when $\lambda_1 = d_t$ and $\lambda_2 = d_m$; and BLR-3 when $\lambda_1 = |d_t - d_m|$.

[2]The assumption of Gaussianity is supported by the use of score normalization techniques such as T-Norm [3].

[3]The FoCal toolkit has been used for LLR training. http://sites.google.com/site/nikobrummer/focal

4. Experiments

4.1. Database and protocols

In NIST SRE protocols the variability in speech length within a given condition is limited, making it unappropriated to test the proposed algorithms. Therefore, two databases and protocols have been generated. For both of these corpora, telephone speech data in NIST SRE 2006 1conv4w-1conv4w condition and the telephone-only speech of NIST SRE 2008 short2-short3 condition have been used. For each utterance in each of these databases, we have truncated them to generate speech segments with lengths 3, 10, 15, 20, 30, 40, 50, 60, 100 and 150 seconds (after silence removal). Then, all the possible comparisons among fragments obtained from model enrollment speech and from test segments were used for generating scores, obtaining roughly 4 millions of scores from NIST SRE 2006 and 3 millions of scores from NIST SRE 2008. We have respectively called such sets of scores DurTelSRE06, used for training the models in the proposed algorithms; and DurTelSRE08, which will be compensated in order to assess their performance.

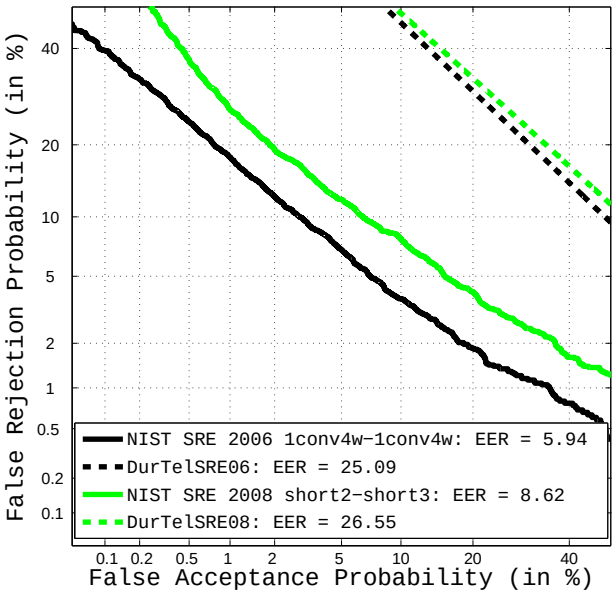


Figure 2: *ATVS system used in NIST SRE 2008. Comparison among DurTelSRE06 and DurTelSRE08 corpora, and the telephone-only subsets of NIST SRE 2006 1conv4w 1conv4w and NIST SRE 2008 short2-short3.*

In this paper, the GMM-UBM based system presented by ATVS in NIST SRE 2008 is used for all experiments. Session variability compensation is achieved by the use of factor analysis techniques, applying Nuisance Attribute Projection (NAP) for the GMM models and using feature-level channel factors for the test segments [7]. T-Norm has been also used for score normalization [3], but no Z-Norm is applied. This allows to see the effects of the proposed algorithms whith score normalization and when score normalization is not applied or is ineffective.

In order to illustrate the performance expected for corpuses DurTelSRE06 and DurTelSRE08, in Figure 2 the DET curves obtained with the system used in this paper and such high-variability corpora are compared to the standard NIST SRE 2006 1conv4w-1conv4w and NIST SRE 2008 short2-short3 telephone-only datasets. It is shown that duration variability implies a dramatic performance decrease, even if the other session variability factors are controlled.

4.2. Results

Results in terms of EER and its relative improvement using the proposed compensation algorithms are presented in Table 4.1 for DurTelSRE08. We compare different subsets of the scores presenting different durations for the speech in the test segments (*Test segment length* in the table). For instance, $10, 100$ in the *Test segment length* column indicates that only scores generated with test segments with 10 and 100 seconds are selected. This is done in order to highlight the adequacy of the proposed methods in conditions of extreme duration variability. The performance for the whole DurTelSRE08 corpus is shown in the *All Durations* row. First, it is seen that all the proposed methods obtain a performance improvement in terms of EER. 1D-methods and BLR-1 only uses duration information of the test segment, whose effects have been partially compensated by T-Norm, but even though the improvement in performance is significant (*e.g.*, $4, 30\%$ for all durations using 1D-GM). This indicates that fixed-cohort normalization schemes are not enough to compensate for duration variability effects. Moreover, when compensation methods are applied using information about the length of the speech used for model enrollment, the improvement in performance is even bigger (*e.g.*, $16, 62\%$ for all durations using 2D-LLR). This indicates that, when score normalization is not present or is uneffective, the proposed methods compensate the remaining duration effects. This is important, since score normalization techniques may not be useful when the cohort is not adapted to the data to use in operational conditions (a situation which is typical in forensics). Although BLR solution is more elegant and workable in terms of model parameters, its performance is worse than its corresponding GM and LLR approaches. It seems that, in this case, the bigger number of parameters in GM and LLR allows the model to better adapt to the problem than the more constrained BLR. Interestingly enough, the proposed LLR and GM schemes, both 1D and 2D, are superior than BLR in terms of computational cost, and also presents a better behavior in terms of their convergence to an optimal solution.

5. Conclusions

This work has discussed the problem of the degradation of speaker verification systems due to the variability of the length of the available speech, which causes misalignments of score distributions. We have compared several novel methods for compensating such effects at the score level, including some novel methods based on Gaussian modelling and Logistic Regression, which outperform previous proposals found in the literature. Results show that using the proposed techniques, significant EER improvements are obtained for different duration variability situations. Moreover, it is shown that even score normalization techniques such as T-Norm are not able to completely eliminate degrading duration effects due to misalignments, and the proposed methods help to compensate them. Future work includes a more comprehensive study of the effects of the proposed algorithms under different score normalization techniques, including adaptive schemes.

	EER (%)	EER improvement (%)						
Test segment length (sec.)		1D-GM	2D-GM	1D-LLR	2D-LLR	BLR 1	BLR 2	BLR 3
3, 10	31.34	2.79	7.17	2.68	7.41	2.63	6.86	4.04
3, 100	31.81	9.61	16.99	9.31	17.23	9.34	15.73	4.59
3, 150	33.08	5.21	10.26	5.15	10.54	4.75	8.73	1.62
10, 100	26.39	3.47	15.67	3.33	15.79	3.57	15.49	2.29
10, 150	27.41	1.97	11.98	1.95	12.28	1.89	11.67	0.22
3, 10, 150	31.11	4.55	10.42	4.43	10.73	2.14	6.32	−0.16
3, 100, 150	31.33	10.28	19.01	10.00	19.29	8.38	13.89	6.67
3, 30, 60, 100, 150	27.96	7.59	19.54	7.44	19.80	3.38	11.67	2.02
All Durations	26.55	4.30	16.37	4.20	16.62	1.84	12.21	0.16

Table 1: *EER improvement in DurTelSRE08 when utterance length compensation is applied. 1D indicates compensation using duration of the test segment (already partially compensated by T-Norm). 2D indicates compensation applied using both durations of the test segment and the speech used to enroll models.*

6. References

[1] Jason Pelecanos, Upendra Chaudhari, and Ganesh Ramaswamy, "Compensation of utterance length for speaker verification," in *Proc. of Odyssey 2004*, Toledo, Spain, 2004, pp. 161–164.

[2] J. Gonzalez-Rodriguez, Phil Rose, D. Ramos, Doroteo 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, no. 7, pp. 2072–2084, 2007.

[3] R. Auckenthaler, M. Carey, and H. Lloyd-Tomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[4] N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[5] L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg, "System combination using auxiliary information for speaker verification," in *Proc. of ICASSP 2008*, Las Vegas, Nevada, USA, 2008, pp. 4853–4856.

[6] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[7] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1969–1978, 2007.

PRESUPUESTO

EJECUCIÓN MATERIAL	
Compra de ordenador personal (software incluido)	2.000 €
Alquiler de impresora láser durante 30 meses	250 €
Material de oficina	300 €
TOTAL EJECUCIÓN MATERIAL	2.550 €

GASTOS GENERALES	
18 % sobre Ejecución Material	459 €

BENEFICIO INDUSTRIAL	
6 % sobre Ejecución Material	153 €

HONORARIOS PROYECTO	
1060 horas a 15 € / hora	15.900 €

MATERIAL FUNGIBLE	
Gastos de impresión	144 €
Encuadernación	10 €
TOTAL MATERIAL FUNGIBLE	

PRESUPUESTO TOTAL	
Presupuesto total sin I.V.A.	19.216 €
Presupuesto total con I.V.A. (18%)	22.675 €

Madrid, Julio de 2010

El Ingeniero Jefe de Proyecto

Fdo.: Sergio Pérez Gómez

Ingeniero Superior de Telecomunicación

PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un Análisis y compensación de variabilidad de la señal de voz en sistemas automáticos de verificación de locutor utilizando información de duración y calidad. 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.