Tesis Doctoral

NORMALIZACIÓN ESTADÍSTICA PARA FUSIÓN BIOMÉTRICA MULTIMODAL

Autor: Pascual Ejarque Monserrate

Director: Francisco Javier Hernando Pericás

Grupo de Procesado de la Voz

Departamento de Teoría de la Señal y Comunicaciones

Universitat Politècnica de Catalunya

Barcelona, septiembre de 2010

A mi familia

Resumen

Los sistemas de reconocimiento biométrico utilizan ciertas características humanas como la voz, los rasgos faciales, la huella dactilar, el iris o la geometría de la mano para identificar a un individuo o verificar su identidad. Dichos sistemas se han desarrollado de forma individual para cada una de estas modalidades biométricas hasta llegar a obtener unos niveles notables de rendimiento.

Los sistemas biométricos multimodales combinan diversas modalidades en un sistema de reconocimiento único. La fusión multimodal permite mejorar los resultados obtenidos por una sola característica biométrica y hacen el sistema más robusto a ruidos e interferencias y más resistente a posibles ataques. La fusión se puede realizar a nivel de las señales adquiridas por los distintos sensores, de los parámetros obtenidos para cada modalidad, de las puntuaciones proporcionadas por expertos unimodales o de la decisión tomada por dichos expertos.

En la fusión a nivel de parámetros o puntuaciones es necesario homogeneizar las características provenientes de las diferentes modalidades biométricas de manera previa al proceso de fusión. A este proceso de homogeneización se le denomina normalización y se ha demostrado determinante en la obtención de buenos resultados de reconocimiento en los sistemas multimodales.

En esta tesis, se presentan diversos métodos de normalización que modifican la estadística de parámetros o puntuaciones. En primer lugar, se propone la normalización de la media y la varianza de las puntuaciones unimodales por medio de transformaciones afines que tienen en cuenta las estadísticas separadas de las puntuaciones de clientes e impostores. En este ámbito se presenta la normalización conjunta de medias, que iguala las medias de las puntuaciones de clientes e impostores para todas las modalidades biométricas.

También se han propuesto técnicas que minimizan la suma de las varianzas de las puntuaciones multimodales de clientes e impostores. Estas técnicas han obtenido buenos resultados en un sistema bimodal de fusión de puntuaciones de espectro de voz e imágenes faciales y se ha demostrado que una reducción de las varianzas multimodales puede comportar un mejor resultado de reconocimiento.

Por otro lado, se ha utilizado la ecualización de histograma, un método ampliamente utilizado en el tratamiento de imágenes, como técnica de normalización. Para ello, se han ecualizado los histogramas de las características unimodales sobre diversas funciones de referencia. En primer lugar, se ha utilizado el histograma de las puntuaciones de una de las modalidades biométricas como referencia en el proceso de ecualización. Esta técnica se ha mostrado especialmente efectiva al combinarla con métodos de fusión basados en la ponderación de las puntuaciones unimodales.

En una segunda aproximación, se han ecualizado las características biométricas a funciones previamente establecidas, en concreto, a una gaussiana y a una doble gaussiana. La ecualización a gaussiana ha obtenido buenos resultados como normalización en sistemas de fusión de parámetros. La ecualización de doble gaussiana se ha diseñado específicamente para la normalización de puntuaciones. Las dos gaussianas representan los lóbulos de las puntuaciones de clientes e impostores que se pueden observar en los histogramas unimodales. Se han probado diferentes variantes para determinar las varianzas de dichas gaussianas.

Las técnicas de normalización estadística presentadas en esta tesis se han probado utilizando diferentes estrategias y técnicas para la fusión, tanto para bases de datos quiméricas como para una base de datos multimodal. Además, la fusión se ha realizado a diferentes niveles, en concreto, a nivel de puntuaciones para diferentes escenarios multimodales incluyendo características de espectro voz, prosodia y caras, y a los niveles de parámetros, puntuaciones y decisión en el entorno del proyecto Agatha.

Agradecimientos

La elaboración de esta tesis ha sido un proceso largo, con horas y atención “robadas” a otras personas y proyectos, tanto profesionales como personales.

Por ello, quiero agradecer su paciencia y su apoyo a mi mujer, Sandra, porque sin su complicidad y su aportación extra en todo lo demás, este trabajo nunca hubiera sido posible. También a mis hijos, Eric y Alex, por dormir bien por las noches y dejar que papá se dedique a sus quehaceres y, sobretodo, por sus sonrisas y sus inquietudes, que nos llenan de energía y cambian nuestro humor en un instante.

También agradecer al resto de mi familia, padres, abuelos, tíos, hermanos, tanto naturales como políticos, y a los amigos, su apoyo y sus ánimos, y que hayan escuchado con paciencia mis charlas sobre la bondad de la biometría y lo anticuado que se va a quedar eso de llevar llaves.

Y a los compañeros y responsables de las empresas donde he trabajado durante este tiempo, Ediciones B, Grupo Zeta, National Geographic y Tàctic, por permitirme la flexibilidad suficiente para compaginar una vida profesional activa con todas las tareas que han requerido la elaboración de esta tesis.

También tengo mucho que agradecer, por supuesto, a aquellos que han colaborado directamente en los trabajos que han dado lugar a esta tesis. En primer lugar, a los doctores A. Tefas y R. Paredes, así como a todos los colaboradores del ITI, por proporcionar las características faciales utilizadas en este trabajo. Y a Mireia, Jordi, Andrey, David H., David G., … y sobretodo, a Javier, el director de la tesis, por su entusiasmo en cada nuevo método y cada resultado satisfactorio y su constancia en animarme siempre a seguir hacia delante.

Muchas gracias a todos.

Sumario

Listado de figuras

Listado de tablas

1 Introducción

En la mayoría de los procesos que involucran personas, conocer su identidad resulta casi tan importante como las acciones que llevan a cabo. Por ejemplo, en los procesos de comunicación, el emisor, el receptor o incluso el transmisor del mensaje son, en la mayoría de los casos, tan importantes como el propio mensaje.

Además, la identificación de personas resulta especialmente crítica en sistemas de seguridad restringida o en identificaciones forenses, en que un error de identificación puede acarrear graves consecuencias para bienes o personas.

Por todo ello, se han desarrollado técnicas que permiten el reconocimiento a partir de características diferenciadas y propias de cada persona como pueden ser la voz, la cara, la huella dactilar, el iris, la verificación de firma, la geometría de la mano, etc. A estas características se las denomina modalidades biométricas o biometrías (Bolle et al., 2004; Jain, 1986; Rabiner et al., 1993).

Las aplicaciones derivadas de estas tecnologías son diversas, por ejemplo, en los ámbitos de control de acceso físico o lógico para salvaguardar bienes o información, en aeropuertos, sistema sanitario, sistema financiero, entorno del vehículo propio, despacho de trabajo, ordenador personal, etc., en el ámbito del control de presencia para el seguimiento de asistencia y tiempo de trabajo, o en el ámbito de justicia y orden público para facilitar la administración e ingresos en prisiones, la identificación en escenas del crimen y otras aplicaciones forenses.

Dichas aplicaciones permiten la verificación o la identificación de los diferentes individuos. En el caso de la verificación se determina si las características biométricas corresponden con quien el individuo dice ser. En el caso de la identificación, el sistema debe reconocer a qué persona corresponde una serie de características biométricas.

Las tecnologías relacionadas con cada una de las modalidades biométricas se han desarrollado de forma independiente y han avanzado hasta alcanzar niveles notables de rendimiento. Sin embargo, todas ellas tienen en la actualidad unos límites en su capacidad de reconocimiento y pueden ser atacadas mediante engaños (*spoof atacks*) (Jain et al., 2004a).

Además, cada uno de estos sistemas de reconocimiento puede tener inconvenientes. Por ejemplo, pueden ser sensibles a cambios ambientales o del entorno, como a cambios de la iluminación o de la orientación de la imagen en el caso del reconocimiento de caras o a ambientes acústicos ruidosos en el caso de la voz. En otros casos, las técnicas de obtención de las características son intrusivas, como en el caso del reconocimiento mediante iris, o se asocian a aplicaciones forenses, como en el caso de las huellas dactilares o el ADN (Bolle et al., 2004).

Biometría multimodal

Los sistemas de reconocimiento biométrico multimodal combinan dos o más de las modalidades biométricas mencionadas anteriormente para obtener mejores y más robustos resultados de reconocimiento que utilizando sistemas biométricos unimodales (una única biometría) (Bolle et al., 2004).

La utilización de estos sistemas puede resolver algunos de los inconvenientes antes mencionados. Por ejemplo, mejora los resultados obtenidos por una sola biometría, por lo que se pueden obtener resultados similares a los conseguidos por una biometría unimodal que se considere intrusiva mediante la combinación de la información de modalidades biométricas mejor aceptadas por los usuarios de las aplicaciones.

Además, dado que la información del sistema de reconocimiento proviene de más de una biometría, un sistema multimodal permite un reconocimiento positivo en un entorno poco adecuado para alguna de las modalidades involucradas y hace que el sistema sea más resistente a posibles ataques (Jain et al., 2004a; Kittler et al., 2002).

También se consigue, mediante la utilización de biometría multimodal, aumentar la universalidad de los sistemas de reconocimiento, dado que dichos sistemas se pueden adaptar para funcionar a partir de un número mínimo de modalidades. De esta manera, aunque un usuario no posea las características de una de las modalidades involucradas en el sistema, podrá seguir utilizándolo haciendo uso del resto de su información biométrica.

Por otro lado, estas técnicas se han mostrado efectivas en el aprovechamiento de la información proporcionada por características biométricas como la edad, la raza o el género. Así, se pueden mejorar los resultados obtenidos por otras modalidades biométricas más efectivas mediante su fusión con dichas características (Jain et al., 1999; Gutta et al., 2000; Jain et al., 2004b).

En un sistema de reconocimiento multimodal, la información se puede integrar a diversos niveles: a nivel de sensor, donde se combinan las señales obtenidas a partir de las características biométricas, a nivel de los parámetros de cada una de las modalidades biométricas unimodales, a nivel de las puntuaciones de reconocimiento proporcionadas por sistemas independientes para cada una de las modalidades biométricas y a nivel de la decisión tomada por cada uno de estos sistemas (Baker et al., 2002; Daugman, 1999).

Tanto en la fusión a nivel de parámetros como en la fusión a nivel de las puntuaciones de reconocimiento, la normalización de las características biométricas es un proceso importante para su homogeneización de forma previa al proceso de fusión propiamente dicho. Tras el proceso de normalización, los parámetros o las puntuaciones se combinan para obtener un vector de parámetros o una puntuación multimodal.

Motivación y objetivos

El principal objetivo de esta tesis es proponer nuevas técnicas o mejorar las ya existentes para la normalización de características biométricas en la fusión a nivel de parámetros y en la fusión a nivel de puntuaciones de reconocimiento.

En la fusión a nivel de puntuaciones de reconocimiento diversos investigadores han desarrollado técnicas basadas en la estadística global de las puntuaciones de las diversas modalidades biométricas. Este es el caso, por ejemplo, de una de las técnicas de normalización más utilizada, la normalización *z-score*. En esta tesis, pretendemos desarrollar técnicas de normalización de puntuaciones biométricas que aprovechen la información proporcionada por las estadísticas separadas de clientes e impostores para, así, reducir las varianzas separadas y mejorar los resultados de reconocimiento.

Por otro lado, y también en el marco de la fusión de puntuaciones de reconocimiento, existen diversos métodos de normalización que se han desarrollado en el ámbito de las biometrías unimodales y han sido escasamente aplicados a la fusión multimodal. Este es el caso de las normalizaciones *Z-Norm* y *T-Norm* para las aplicaciones de voz o de la ecualización de histograma (*HEQ*) en las caras. Queremos adaptar y mejorar algunas de estas técnicas para su aplicación a la fusión multimodal.

Otro de los objetivos de esta tesis es el diseño de *SVM*s tanto para la fusión a nivel de parámetros como para la fusión a nivel de puntuaciones de reconocimiento. Los *kernels* más usualmente utilizados son los que se basan en funciones polinomiales o en *rbf (radial basis function)*. En esta tesis queremos determinar cuales son los *kernels* más adecuados para cada una de las fusiones a realizar.

En el entorno de la fusión a nivel de parámetros se determinará el efecto de la reducción de parámetros sobre los vectores unimodales, para determinar si dicha reducción es relevante en el resultado de reconocimiento multimodal.

Queremos también estudiar, en la fusión a los niveles de fusión de parámetros, puntuaciones y decisión, cuales son las posibilidades para fusionar modalidades biométricas con una evolución temporal, como la voz, con otras, como las imágenes faciales, en que se obtienen las observaciones en un momento dado. La dificultad principal se encuentra, en este caso, en la fusión a nivel de parámetros, dado que en las biometrías con evolución temporal se obtienen numerosos vectores de parámetros por ocurrencia mientras que en el resto de modalidades biométricas únicamente se obtiene un vector de parámetros por cada ocurrencia.

Por último, y dado que la mayoría de los trabajos realizados hasta el momento involucran únicamente dos modalidades biométricas, queremos explorar el efecto de introducir tres o más biometrías en un sistema multimodal. Está demostrado que, en casi todos los casos, la fusión de dos modalidades biométricas mejora el resultado de reconocimiento de cada una de ellas por separado. Sin embargo, no está claro que añadir biometrías adicionales continúe mejorando los sistemas multimodales. Por ello, es objetivo de esta tesis realizar fusión de, al menos, tres informaciones biométricas independientes.

Estructura

Esta tesis se divide en ocho capítulos.

El capítulo 1 contiene una breve introducción a los sistemas de reconocimiento biométrico y a la biometría multimodal, los objetivos principales de la tesis y la estructura de sus contenidos.

En el capítulo 2 se revisa el estado del arte del reconocimiento biométrico. En primer lugar se detallan las principales propiedades de los sistemas de reconocimiento biométrico, para dar paso a la revisión de las modalidades biométricas de más extendida utilización. Posteriormente, se describen las dos principales variantes del reconocimiento biométrico, la identificación y la verificación, así como el proceso de reconocimiento para cada uno de estos casos. En este apartado también se describen las diferentes aplicaciones de las tecnologías biométricas y se discute sobre el derecho a la privacidad de los datos biométricos.

El capítulo 3 se dedica al estado del arte del reconocimiento multimodal indicando, en primer lugar, las ventajas del reconocimiento multimodal frente al reconocimiento unimodal. Posteriormente, se describen los procesos de normalización y fusión, destacando la importancia de realizar una homogenización de las características como paso previo a la fusión. A partir de este punto, se revisan los diferentes niveles de fusión multimodal detallando las características y técnicas de las fusiones a nivel de parámetros, puntuaciones y decisión. Finalmente, se detalla la manera en que los resultados multimodales deben ser evaluados.

En el capítulo 4 se presentan las técnicas de normalización de media y varianza basadas en las estadísticas separadas de clientes e impostores desarrolladas en esta tesis. En primer lugar, se presenta la normalización conjunta de medias, una técnica de normalización de puntuaciones en que se normalizan las medias de las puntuaciones de clientes e impostores de forma separada. Posteriormente, se presentan las técnicas de minimización de varianzas, que, a partir de una normalización de las puntuaciones unimodales, están diseñadas para minimizar la suma de las desviaciones estándar o de las varianzas multimodales de clientes e impostores.

En el capítulo 5 se describe el proceso de ecualización para dar paso a la presentación de diferentes técnicas de normalización desarrolladas en esta tesis que se basan en esta técnica. Las técnicas presentados son la ecualización al histograma de una biometría multimodal, la ecualización a gaussiana, en que se utiliza una gaussiana como referencia y la ecualización de doble gaussiana, donde se tienen en cuenta las estadísticas separadas de clientes e impostores para generar la función de referencia mediante la suma de dos gaussianas.

Los capítulos 6 y 7 muestras los resultados obtenidos en esta tesis, respectivamente, en la fusión de puntuaciones de voz y caras y mediante fusión multinivel en el entorno del proyecto Agatha. En el capítulo 8 se presentan las conclusiones de la tesis y se apuntan algunas posibles líneas de trabajo futuras.

2 El reconocimiento biométrico

La biometría es la ciencia que estudia el reconocimiento de personas por medio de sus características físicas o de comportamiento. El reconocimiento biométrico permite realizar la verificación o la identificación de una persona mediante sistemas automáticos.

Se define como modalidad biométrica cualquiera de las características humanas que permiten el reconocimiento (Bolle et al., 2004; Ross et al., 2006; Sanderson, 2008). Algunas de las modalidades biométricas que más se han estudiado y que han demostrado su fiabilidad y usabilidad son el habla para reconocimiento de locutor, la imagen fija de la cara, la huella dactilar y el iris (Jain, 1996; Rabiner et al., 1993).

Los sistemas biométricos comparan las características de un individuo con modelos obtenidos a partir de muestras anteriores obtenidas del mismo individuo y habitualmente proporcionan buenos resultados de reconocimiento. Además, las tasas de error se pueden adaptar a las necesidades de cada aplicación.

Existen dos formas de realizar la comparación entre muestras. Cuando la comparación se hace de un individuo a varios, el proceso se denomina identificación, mientras que cuando se realiza de uno a uno se denomina verificación (Campbell, 1997).

Los sistemas de reconocimiento biométrico suelen componerse de una etapa de adquisición de la señal, en que se convierten las características observables en señales, de una fase de parametrización que extrae los parámetros más relevantes de las señales recogidas, de la clasificación de las señales por comparación con los modelos de cada usuario y de la evaluación de los resultados obtenidos para determinar el resultado y la relevancia del proceso de reconocimiento que se ha llevado a cabo (Bolle et al., 2004).

Cabe destacar el gran número de aplicaciones existentes relacionadas con el reconocimiento biométrico, incluyendo el control de acceso, bien sea físico, para acceder a lugares de acceso

restringido, o lógico, para acceder a datos que requieran cierto nivel de seguridad, aplicaciones de justicia y orden público, para identificación o descarte de sospechosos, sistemas de control de tiempo de trabajo y presencia y sistemas biométricos móviles que permiten llevar la seguridad de las anteriores aplicaciones a cualquier ubicación física.

La puesta en marcha de estos sistemas implica el almacenamiento de datos o patrones biométricos, por lo que uno de los aspectos más sensibles en el uso de las tecnologías biométricas es el relacionado con la privacidad ya que algunos usuarios pueden considerar que la utilización de estas tecnologías reduce su derecho a controlar la información sobre sí mismo (Woodward, 1997).

2.1 Propiedades del reconocimiento biométrico

Existen tres formas principales de reconocimiento de un individuo que se utilizan en aplicaciones de seguridad (Wayman et al., 2005; Miller, 1994):

- Algo que el usuario conoce: un código secreto, una fecha concreta, una frase, en definitiva, una contraseña.
- Algo que el usuario posee: una llave, una tarjeta de acceso, una tarjeta de memoria, etc.
- Algo que el usuario es: biometría.

En el primer caso, la clave que el usuario conoce puede ser olvidada o "robada". En el segundo caso, el objeto que permite el acceso pude ser robado o perdido, con lo que se deben cambiar cerraduras, anular las tarjetas, etc. En el tercer caso, el usuario no tiene que memorizar nada y las características biométricas no se pueden perder o robar y es complejo falsificar las características biométricas ya que requiere más experiencia, tiempo, dinero y tecnología que en el caso de cualquier otro sistema de seguridad.

Las propiedades deseables para cualquier sistema de reconocimiento biométrico incluyen las cinco propiedades descritas en (Clarke, 1994).

- Universalidad: Se dice que una característica biométrica es universal cuando se encuentra en todos los individuos de la población.
- Unicidad: La unicidad implica que las características biométricas consideradas sean diferentes para cada uno de los individuos y diferenciables de las de cualquier otro individuo, de manera que cada individuo sea distinguible del resto a partir de dichas características.
- Permanencia: La permanencia hace referencia a la estacionalidad de las características biométricas. Es deseable que éstas se mantengan a lo largo del tiempo, de manera que no

varíen con la edad del individuo ni con situaciones transitorias como enfermedad, estado del humor, utilización de accesorios, etc.

- Cuantificación: Las características biométricas deben ser mensurables y cuantificables de forma objetiva. De esta manera, los valores obtenidos para las diferentes características serán comparables entre sí y permitirán discriminar entre los diferentes individuos.
- Aceptación: Ciertas modalidades biométricas pueden ser poco aceptadas por los potenciales usuarios, bien sea por ser consideradas intrusivas o bien por ser relacionadas con aplicaciones criminales y forenses o de alta seguridad.

La adecuada combinación de estas características determina la efectividad de un sistema biométrico, aunque ninguna modalidad biométrica cumple con ninguna de estas características de forma absoluta.

Otras características que deben tenerse en cuenta en el diseño de un sistema de reconocimiento biométrico son:

- Fiabilidad: Los resultados proporcionados por los sistemas de reconocimiento biométrico deben proporcionar unos resultados de reconocimiento acordes con la seguridad requerida por la aplicación.
- Facilidad de uso: Esta característica se refiere a la facilidad en la adquisición, la medición y el almacenamiento de los datos, así como a la facilidad y rapidez en la obtención del resultado de reconocimiento.
- Prevención de ataques: Un sistema de reconocimiento debe ser robusto a los posibles ataques malintencionados por parte de individuos que pretendan acceder a lugares o datos a los que no tienen derecho de acceso.
- Coste: El coste depende del hardware, la instalación, la facilidad de uso, el mantenimiento, la base de datos, etc. Es importante tener en cuenta el coste, principalmente en aplicaciones de seguridad media.

2.2 Modalidades biométricas

Se puede considerar modalidad biométrica a cualquier característica humana que permita distinguir entre diferentes individuos. Una de las formas más comunes de clasificar las modalidades biométricas es distinguir entre modalidades biométricas fisiológicas y conductuales.

- Biometrías fisiológicas: Son aquellas relacionadas con las características físicas de los individuos, como pueden ser las características faciales (Turk et al., 1991; Wechsler et al.,

2006), las huellas dactilares (Lee et al., 1994; Maltoni et al., 2003), la geometría de la mano (Zunkel, 1999), el iris (Wildes, 1997), la retina, la geometría de la oreja, el ADN o la altura.

- Biometrías conductuales: Están relacionadas con el comportamiento o la forma en que alguna acción se realiza. Algunas de estas modalidades biométricas pueden ser la forma de escribir o firmar, la frecuencia de pulsación en un teclado, la gestualidad o la forma de caminar.

En el caso de la voz (Furui, 1997), algunos autores consideran que es una biometría fisiológica mientras que otros consideran que es una biometría conductual, dado que en la producción de las ondas sonoras influyen tanto características del tracto vocal del individuo como características conductuales en la forma de hablar.

También algunos autores distinguen como "biometrías blandas" (soft biometrics) aquellas que cumplen en menor medida con la característica de la unicidad, como la altura, la gestualidad, etc., y que se utilizan principalmente como información complementaria para mejorar el resultado obtenido por otras modalidades que permiten en mayor medida entre diferentes individuos por sí solas, como voz, cara, huella dactilar, iris, etc.

En este apartado vamos a revisar las modalidades biométricas más utilizadas en los actuales sistemas de reconocimiento de personas del estado del arte y que poseen un alto nivel de universalidad y unicidad además de haber demostrado su fiabilidad: el habla para reconocimiento de locutor, la imagen fija de la cara, la huella dactilar y el iris.

Reconocimiento de locutor

El reconocimiento de locutor permite el reconocimiento de los diferentes individuos mediante las diferentes características de su voz (Furui, 1997). Este es un método habitualmente utilizado por las personas para reconocer a otras personas incluso con sólo unas pocas palabras, aunque en este caso, el cerebro humano se apoya en el contexto o en el escenario de la comunicación para realizar esta función.

El reconocimiento de locutor no se puede considerar del todo universal porque, por ejemplo, los niños de corta edad no tienen la capacidad del habla y existen personas adultas con disfunciones del aparato fonador.

Tampoco se puede considerar que la unicidad sea absoluta dado que existen personas con características vocales muy similares o prácticamente idénticas, aunque la fiabilidad de esta biometría es alta.

En cuanto a la permanencia, la voz varía de forma significativa con la edad, sobretodo en el paso de niño a la edad adulta, y con enfermedades que afectan al aparato fonador, como ronqueras, catarros o resfriados.

Las características del reconocimiento de locutor sí se pueden considerar cuantificables dado que se obtienen a partir de transformaciones de la medición de la onda de presión generada por el habla.

La aceptación de los sistemas de reconocimiento de locutor es buena dado que la interfase entre el individuo y el sistema suele ser un micrófono por lo que la adquisición no se considera invasiva y, además, esta tecnología no se asocia con aplicaciones forenses ni a aplicaciones de alta seguridad o seguimiento (Maltoni et al., 2003; Bolle et al., 2004).

Dado que la adquisición de la señal es sencilla, se considera una biometría fácil de utilizar. Sin embargo, esta misma característica hace que la señal pueda ser captada para intentar acceder de forma fraudulenta al sistema.

En cuanto a los factores que pueden hacer disminuir el rendimiento de los sistemas de reconocimiento de locutor, los principales son el ruido ambiental y las distorsiones introducidas por los micrófonos que se utilizan para captar la señal de voz.

Reconocimiento mediante caras

El reconocimiento mediante caras permite identificar a los diferentes individuos mediante sus características faciales (Turk et al., 1991; Wechsler et al., 2006). También en este caso se utiliza una característica que es habitual para el reconocimiento entre humanos, seguramente, la más utilizada para esta función (Maltoni et al., 2003).

Esta modalidad biométrica se puede considerar universal ya que todos los individuos tienen un rostro a partir del cual identificarlos.

En cuanto a la unicidad, las características faciales son diferentes para cada individuo aunque en el caso de hermanos genéticamente idénticos el parecido es muy grande y son difíciles de distinguir, aunque sus familiares y conocidos más allegados suelen ser capaces de hacerlo.

Las características faciales, al igual que en caso de la voz, varían con la edad del individuo. De la misma manera estas características pueden verse modificadas por la utilización de complementos como gafas o por variaciones en el estilo o peinado.

En cuanto a la cuantificación, la señal obtenida mediante los distintos sensores como cámaras fotográficas o de grabación de imagen, son digitales o pueden ser digitalizas y descompuestas en píxeles. De cada píxel se puede obtener la intensidad de la luz en el caso de señales en blanco y negro o el nivel de cada uno de los colores básicos, rojo, verde y azul (RGB: red, green, blue) para su tratamiento.

Esta biometría se considera fiable y fácil de utilizar, dado que la señal biométrica se puede captar mediante una simple cámara fotográfica o videocámara. Además es aceptada por los individuos, que no suelen considerar una violación de su privacidad la captación de una imagen de su cara (Zhao et al., 2003). Al igual que en el caso de la voz, la facilidad para captar la cara de un individuo hace que esta biométrica sea más sensible a ataques malintencionados.

Variaciones en la intensidad de la luz, en la situación relativa del origen de la luz que pueda producir sombras, rotaciones de la cabeza o problemas en la obtención de la señal como desenfoques pueden influir en la calidad del sistema de reconocimiento mediante caras (Maltoni et al., 2003; Jain et al., 1994).

Reconocimiento mediante huellas dactilares

Las huellas dactilares han sido utilizadas como prueba forense para la identificación de personas desde hace más de un siglo. La situación de arcos, presillas internas, presillas externas y verticilos en las huellas dactilares de un individuo permiten su reconocimiento (Lee et al., 1994; Maltoni et al., 2003).

Todos los individuos tienen huellas dactilares excepto en los casos en que por accidente o de forma intencionada las huellas se han "borrado" bien sea por quemaduras, por la aplicación de ácidos, etc.

Las huellas dactilares son diferentes para todos los humanos. Incluso gemelos idénticos tienen huellas dactilares diferentes dado que en su formación influyen tanto características genéticas como características ambientales dentro del útero materno.

Sin embargo, las huellas dactilares permanecen inalterables a lo largo de la vida a menos que se produzcan cortes, lesiones o enfermedades que afecten a las capas más profundas de la piel.

Las huellas dactilares pueden ser escaneadas y, por lo tanto, se puede obtener una imagen digitalizada de ellas y cuantificar las características relevantes para la identificación.

En este caso, las aplicaciones basadas en huellas dactilares suelen ser más fiables que las basadas en la voz o en la cara, aunque esta modalidad se considera más intrusiva que en los casos anteriores y se relaciona en mayor medida con aplicaciones forenses.

También es una biometría de fácil utilización dado que para adquirir la señal únicamente es necesario un escáner. Sin embargo, en este caso, la señal es más difícil de obtener por quien quiera atacar una aplicación basada en esta biometría.

Reconocimiento mediante iris

El iris es la membrana coloreada y circular del ojo. En su centro se encuentra la pupila, de color negro. La zona blanca que se encuentra alrededor se denomina esclerótica.

Todos los individuos que poseen glóbulos oculares poseen iris.

Una propiedad que el iris comparte con las huellas dactilares es la morfología aleatoria de su estructura. La textura del iris por sí misma es estocástica o posiblemente caótica y es única para cada individuo (Wildes, 1997).

Además, el iris de cada individuo no se ve alterado a lo largo de su vida a menos que el individuo se vea afectado por determinadas y poco frecuentes enfermedades oculares.

En cuanto a la cuantificación de la señal, al igual que en el caso de las caras y las huellas dactilares, la imagen que se obtiene del iris puede ser digitalizada y las características relevantes cuantificadas.

Además el patrón del iris es difícil de duplicar o falsificar, por lo que la prevención de ataques es muy alta para esta modalidad. La complejidad de estos patrones y los sensores utilizados para la adquisición de la señal hacen, por otro lado, que la facilidad de utilización de esta biometría sea menor que en los casos anteriores.

Todas estas características hacen que esta biometría sea muy fiable aunque la aceptación por parte de los posibles usuarios sea menor que en los casos anteriores, principalmente porque la captación de la señal con las tecnologías actuales es bastante intrusiva, aunque se están desarrollando nuevos sensores para reducir al máximo este inconveniente (Maltoni et al., 2003).

La siguiente tabla resume las principales características de las modalidades biométricas comentadas:

	Voz	**Cara**	**Huellas dactilares**	**Iris**
Permanencia	Media	Media	Alta	Alta
Fiabilidad	Media	Media	Alta	Alta
Facilidad de uso	Alta	Alta	Alta	Media
Prevención de ataques	Media	Media	Alta	Alta
Aceptación	Alta	Alta	Media	Media

Tabla 2-1: Resumen de características de las modalidades biométricas: voz, caras, huellas dactilares e iris.

2.3 Identificación y verificación

Las dos principales variantes del reconocimiento biométrico son la identificación y la verificación de personas (Campbell, 1997).

La identificación consiste en reconocer a qué persona corresponden una serie de características biométricas. Para llevarla a cabo se deben comparar dichas características con los modelos de cada usuario realizados previamente a partir de características recogidas para este fin, proceso que se denomina entrenamiento. En el caso de realizar una identificación en un grupo cerrado de usuarios, se decide que las características biométricas corresponden a aquel usuario cuyo modelo mejor se corresponda con dichas señales. En el caso de realizarla sobre un grupo abierto, además se debe cumplir con un mínimo de similitud; en caso contrario, se determina que las características corresponden a un usuario no identificado.

En el caso de la verificación, se determina si las características biométricas corresponden con quien el individuo dice ser. Para ello, se comparan dichas características con el modelo del individuo que se dice ser. Si la similitud es suficiente, se considera que el individuo es un cliente, es decir, que es quien dice ser. En caso contrario, se considera que el individuo es un impostor. Algunos sistemas, denominados sistemas de verificación abiertos, en contraposición a los anteriores denominados sistemas de verificación cerrados, también consideran dar una respuesta no concluyente.

Tanto en el caso de la identificación como en el de la verificación, algunos sistemas, principalmente en el reconocimiento de locutor, incluyen el entrenamiento de un modelo universal (*UBM: Universal Background Model*) que se utiliza para relativizar la similitud de las características biométricas con los modelos de los individuos.

2.4 El proceso de reconocimiento automático

El proceso de reconocimiento automático comprende desde la captación y grabación de las características o señales biométricas hasta la decisión final sobre el reconocimiento, bien sea una identificación o una verificación.

Las principales fases de los sistemas de reconocimiento automático del estado del arte son la adquisición de la señal biométrica, la obtención de parámetros relevantes, la clasificación de dichos parámetros y la evaluación de los resultados obtenidos (Bolle et al., 2004).

2.4.1 Adquisición de la señal biométrica

Las diferentes características biométricas deben ser captadas y convertidas en señales para que puedan ser procesadas por los sistemas de reconocimiento. Para ello es necesaria la utilización de sensores que conviertan propiedades físicas en señales.

Por ejemplo, en el caso de la voz, el aire expulsado por los pulmones y que se convierte en voz al pasar por el tracto vocal, genera oscilaciones de la presión del aire denominadas ondas sonoras, que son convertidas en ondas mecánicas en el oído humano. Estas oscilaciones de la presión del aire pueden ser captadas por medio de unos sensores específicos denominado micrófonos y que las convierten en señales eléctricas que por lo general son de mayor amplitud cuanto mayor es la presión sonora.

En el caso del reconocimiento facial se debe captar la luz que se refleja en la cara del individuo a reconocer. Para ello se puede utilizar una cámara fotográfica o una cámara de grabación de vídeo. En la actualidad, estos dispositivos convierten la luz recibida en señales eléctricas que son almacenadas de forma digital. Las imágenes quedan divididas en píxeles y para cada píxel se almacena la intensidad luminosa para imágenes en blanco y negro o los tres valores de intensidad luminosa que definen un color en cualquier modelo como, por ejemplo, el RGB.

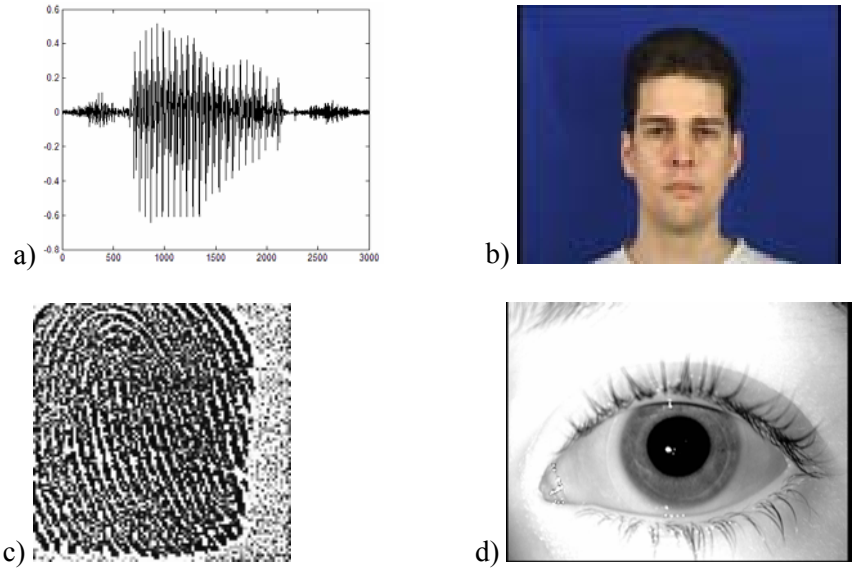


Figura 2-1: Representación de señales biométricas: fragmento de señal de voz de la base de datos POLYCOST (a), imagen facial frontal de la base de datos XM2VTS (b), huella dactilar de la base de datos BIOSEC (c) e iris de la base de datos BIOSEC (d).

Tanto para la adquisición de las huellas dactilares como del patrón del iris es necesario captar una imagen por medio de un sensor adecuado, que puede ser un escáner en ambos casos, una

capacitancia CMOS o un sensor termal o ultrasónico en el caso de la huella dactilar, o incluso una cámara fotográfica con el zoom adecuado en el caso del iris. Una vez obtenidas las imágenes que contienen las características biométricas se pueden almacenar de la misma manera que en el caso de las caras.

2.4.2 Parametrización

Por lo general, los valores de las señales biométricas no permiten reconocer de forma directa un individuo. Así, por ejemplo, no es posible reconocer a una persona a partir de la amplitud de la señal de voz y es muy costoso el reconocimiento a partir de los valores de la intensidad lumínica de una imagen.

Por este motivo, a partir de la señal biométrica, es necesario obtener una serie de características relevantes o parámetros que los sistemas automáticos utilizan para el reconocimiento de los individuos. A este proceso se le denomina parametrización. La elección de los parámetros es determinante para el rendimiento de un sistema de reconocimiento biométrico.

Los sistemas del estado del arte actual en reconocimiento de locutor utilizan características espectrales de la señal de voz, como pueden ser los parámetros LDA, los *cepstrum* o los parámetros de filtrado de frecuencias (*FF: Frequency Filtering*) (Nadeu et al., 1995). En trabajos recientes se han cumplimentado dichos parámetros espectrales con otras características como las prosódicas o el *jitter* y el *shimmer*.

Los parámetros más utilizados en el caso del reconocimiento de caras son los *eigenfaces* y los *fisherfaces*. Los sistemas de reconocimiento de caras, principalmente el clasificador de vecino más cercano (*Nearest Neighbor classifier*), son costosos computacionalmente y requieren una gran capacidad de almacenamiento si se utilizan en el espacio de la imagen completa. Por ello, es conveniente aplicar esquemas de reducción de la dimensionalidad (Belhumeur et al., 1997).

Una técnica comúnmente utilizada para la reducción de la dimensionalidad es el análisis de componentes principales (*PCA: principal component analysis*), que realiza una proyección lineal que minimiza la dispersión de las muestras proyectadas. Los parámetros obtenidos tras la aplicación de PCA a la imagen se denominan *eigenfaces*.

Por otro lado, se debe tener en cuenta que, bajo condiciones ideales, la variación dentro de una misma clase se encuentra en un subespacio lineal del espacio de la imagen. Se pude realizar una reducción de las dimensiones de una imagen usando una proyección lineal manteniendo la discriminación lineal. Por medio de una transformación lineal discriminante (Jonsson et al., 2002; Belhumeur et al., 1997) se puede realizar una reducción lineal. Los parámetros obtenidos se denominan *fisherfaces*.

Otra técnica de parametrización de caras es la basada en el algoritmo NMFaces (Zafeiriou et al., 2005a). Los sistemas de reconocimiento facial se basan en el principio de que una cara se puede representar como un conjunto de partes distribuidas de forma dispersa: ojos, nariz, boca, etc. La factorización no negativa de matrices (*NMF: Non-negative Matrix Factorization*) es utilizada en (Zafeiriou et al., 2005a, Zafeiriou et al., 2005b) para conseguir que la representación distribuida de parámetros localizados representen las partes constituyentes de la cara en las imágenes faciales. La factorización no negativa de matrices es una técnica de reconocimiento facial basada en la apariencia que utiliza las técnicas convencionales de análisis de componentes.

Las huellas dactilares son una impresión visible de las crestas papilares de los dedos. El reconocimiento de personas es posible por la identificación dentro de las huellas de diferentes patrones (*pattern*) o detalles, denominados minucias (*minutia*), que se encuentran en ellas. Los patrones identificables son la espiral o círculo (*whorl*), el arco (*arch*) y el lazo (*loop*). Las minucias que se pueden encontrar son el final de una cresta (*ridge ending*), la bifurcación (*bifurcation*) y la cresta corta o punto (*short ridge or dot*). Los parámetros utilizados en el reconocimiento mediante huellas dactilares se obtienen a partir de la identificación de estas características en una huella dactilar (Lee et al., 1994).

En el caso del reconocimiento de iris, las transformaciones de Gabor o mediante *wavelets* son las técnicas más utilizadas para extraer los parámetros a partir de una imagen del iris (Wildes, 1997; Boles et al., 1998). El filtro de Gabor es un filtro lineal cuya respuesta impulsional es una función armónica multiplicada por una función gaussiana. Son funciones casi pasa banda que obtienen óptimas resoluciones tanto en los dominios espaciales como frecuenciales. La transformación mediante *wavelets* es una herramienta muy potente para discriminación de texturas. Es una operación lineal que descompone la señal en componentes a diferentes escalas. Esta transformación se basa en la convolución de la señal con una ventana de longitud variable en función de la componente espectral deseada.

2.4.3 Reconocimiento: entrenamiento y clasificación

Se denomina clasificación al proceso de reconocimiento, identificación o verificación, a partir de los parámetros obtenidos de la ocurrencia de una modalidad biométrica. De hecho, el reconocimiento consiste en decidir a que clase pertenece una señal biométrica. En el caso de la identificación se definen tantas clases como individuos a identificar mientras que, en el caso de la verificación, se definen dos clases, cliente e impostor. En ambos casos, para los sistemas abiertos se define una clase más para el resultado indeterminado.

Para realizar la clasificación se deben comparar los parámetros de las señales biométricas con los modelos de cada clase obtenidos en la fase que se denomina de entrenamiento. En esta fase de

entrenamiento, a partir de señales específicamente obtenidas para ello se realiza el proceso de parametrización y se crea un modelo para cada usuario.

En el caso de la identificación, los parámetros de las señales biométricas se comparan con los modelos de todos los individuos que componen el sistema. El resultado de esta comparación es una puntuación (*score)* para cada individuo, es decir, para cada clase. Así, en el proceso de clasificación se decide que la señal biométrica corresponde al individuo cuyo modelo ha obtenido mayor puntuación en la comparativa con la señal biométrica. Si se trata de un sistema de identificación abierto, se establece un umbral (*threshold*) y, si ninguna puntuación supera este valor, se decide que el individuo que ha generado al señal biométrica no pertenece al conjunto de individuos del sistema.

En los sistemas de verificación se puede realizar el mismo procedimiento que en los sistemas de identificación y decidir que el individuo es un cliente si el modelo de quien dice ser es el que obtiene la mayor puntuación, o decidir que es un impostor en caso contrario.

Sin embargo, este procedimiento es computacionalmente costoso y se suele optar por comparar los parámetros de la señal biométrica únicamente con el modelo del individuo que se dice ser. En este caso, se compara la puntuación obtenida con un umbral. Si la puntuación es mayor, se decide que el individuo es un cliente y, en caso contrario, que es un impostor.

En los sistemas abiertos de verificación se definen dos umbrales. Si la puntuación está por encima de ambos se decide "cliente", si está por debajo de ambos se decide "impostor" y si el valor es intermedio se decide que el resultado no es concluyente.

Algunos sistemas utilizan además, para la obtención de las puntuaciones, un modelo universal (*UBM: Universal Background Model*) para relativizar o ponderar los resultados obtenidos por la comparación de la señal biométrica con los modelos de los individuos. Este modelo universal se entrena mediante la utilización de señales ajenas a las implicadas en el sistema de reconocimiento.

La figura 2-2 muestra el esquema de un sistema de reconocimiento.

Existen numerosas técnicas para la creación de los modelos y la clasificación de las señales biométricas, algunos específicos de ciertos parámetros o modalidades biométricas y otros que pueden ser utilizados de forma más general.

Por ejemplo, en el caso del reconocimiento de locutor, si se obtiene un vector de parámetros para cada trama de longitud fija de voz, como puede ser el caso de los parámetros LDA, cepstrum o FF, la longitud del vector de parámetros es variable para cada ocurrencia. Para resolver esta dificultad, los sistemas del estado del arte utilizan modelos ocultos de Markov (*HMM: Hidden Harkov Models*) o modelos de mezcla de gaussianas (*GMM: Gaussian Mixture Models*) (Rabiner, 1989). En este caso, la comparación entre las señales biométricas y los modelos de cada individuo se

implementa calculando la probabilidad de que una señal biométrica pertenezca a cada uno de los individuos según la regla de máxima probabilidad (*maximum likelihood rule*).

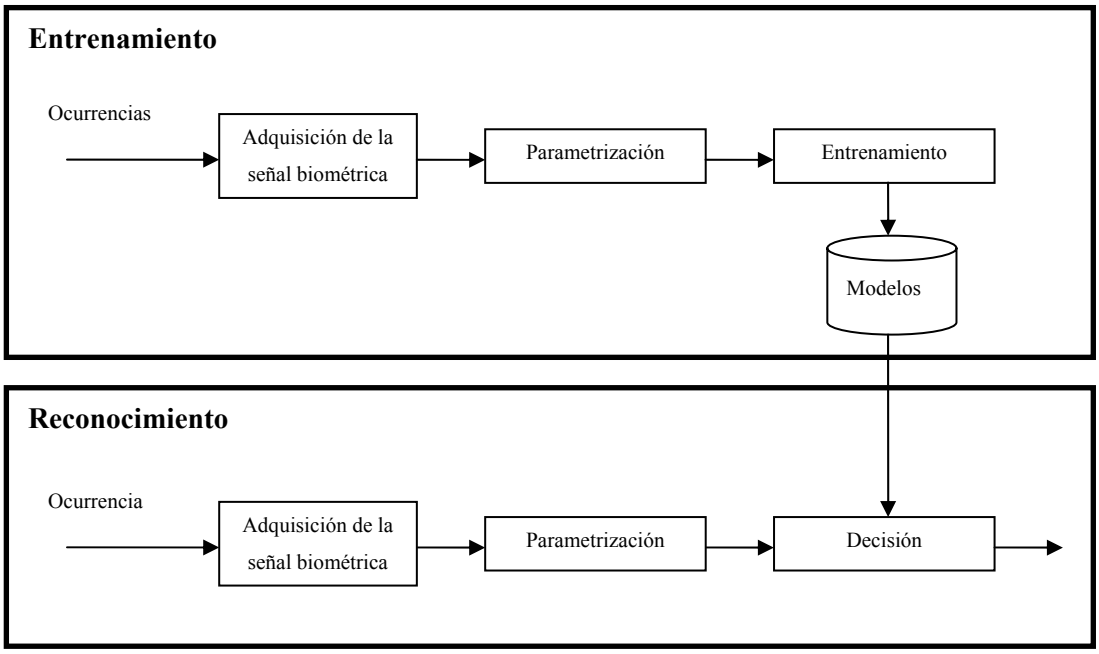


Figura 2-2: Esquema de un sistema de reconocimiento.

Otra forma de solucionar la problemática del número de parámetros variable para cada señal de voz es entrenar una GMM para cada trama de voz, por medio de una adaptación de máximo a posteriori (*MAP: maximum a posteriori*), y promediar las medias de las gaussianas obtenidas para cada trama. Así, el número de parámetros obtenido para cada ocurrencia de voz pasa a ser constante e igual al número de gaussianas del sistema multiplicado por el número de componentes de cada gaussiana. A estos vectores de parámetros se les denomina supervectores GMM (*GMM-SV: GMM supervectors*) (Campbell et al., 2006).

Para el caso del reconocimiento de locutor mediante supervectores GMM, se pueden utilizar las numerosas técnicas de clasificación que existen para sistemas con vectores de parámetros de longitud fija.

También en el reconocimiento de caras, huellas dactilares o iris, la longitud de los parámetros utilizados suele ser fija, por lo que se pueden utilizar dichas técnicas de clasificación.

Algunas de ellas son el clasificador Bayesiano (Langley et al., 1992), los árboles de decisión, las redes neuronales (*Neural Networks*) como el *Multilayer Perceptron (MLP)* (Bishop, 1995) y las *Support Vector Machines (SVMs)* (Cristianini et al., 2000; Burges et al., 1996), y, también en este caso, los modelos ocultos de Harkov y los sistemas de mezcla de gaussianas.

2.4.4 Evaluación

Para la evaluación de los sistemas de reconocimiento se han definido una serie de medidas basadas en el número de aciertos y errores que comete el sistema sobre un conjunto de señales independiente de las utilizadas para el entrenamiento.

En un sistema de verificación los indicadores que establecen el número de errores cometidos por un sistema son (Bimbot et al., 2004; Wayman el al., 2005):

- Falso rechazo (FR: False rejection): El sistema rechaza a un cliente que se ha identificado correctamente.
- Falsa aceptación (FA: False acceptance): El sistema acepta a un impostor que ha suplantado la identidad de un usuario del sistema.

A partir de estos indicadores se pueden definir las tasas de error correspondientes, que son las que se suelen utilizar habitualmente para comparar los sistemas de verificación y que son:

- Tasa de falso rechazo (FRR: False rejection rate): Número de falsos rechazos dividido por el número de pruebas de clientes.
- Tasa de falsa aceptación (FAR: False acceptance rate): Número de falsas aceptación dividido por el número de pruebas de impostores.

Estas medidas varían para cada sistema biométrico en función del umbral que se establezca para determinar si una puntuación obtenida por el sistema corresponde a un cliente o a un impostor. Si el umbral que se establece para el sistema biométrico es alto se aceptan pocos errores de clientes, por lo que FAR será bajo pero FRR será alto y el sistema se habrá definido como de alta seguridad. En caso contrario, cuando se establece un umbral bajo, el valor de FRR será bajo y FAR será mayor y el sistema será de menor seguridad pero producirá menos rechazos de clientes.

Por este motivo, una medida muy utilizada para comparar sistemas de verificación es la tasa de error equivalente o tasa de igual error (*EER: Equal Error Rate*), que es la tasa de error en el punto en que se intersecan las curvas de FAR y FRR en función del umbral. Otra medida utilizada es la mínima tasa de error total (*HTER: Half Total Error Rate*), que se define como el valor mínimo de la semisuma de FAR y FRR.

Otra opción para evaluar los sistemas es fijar FAR o FRR a una cantidad fija y comparar los valores de la otra tasa de error para el umbral establecido.

Si se quiere evaluar la relación entre las dos tasas de error existen diferentes representaciones gráficas de dichas tasas que permiten realizar este tipo de comparaciones. Las más utilizadas son el diagrama FAR-FRR, la curva ROC y la curva DET. El diagrama FAR-FRR muestra la evolución de ambas tasas en función del umbral. La curva ROC (*Receiver Operating Characteristic*)

representa la relación entre FAR y FRR (Bimbot et al., 2004). La curva DET (*Detection Error Trade-offs*) es una representación no lineal de la curva ROC y es la representación de la relación entre FAR y FRR más utilizada actualmente (Martin et al., 1997; Bimbot et al., 2004). La figura 2-3 muestra un ejemplo de estas tres representaciones según las puntuaciones del sistema basado en espectro de voz del apartado 6.2.

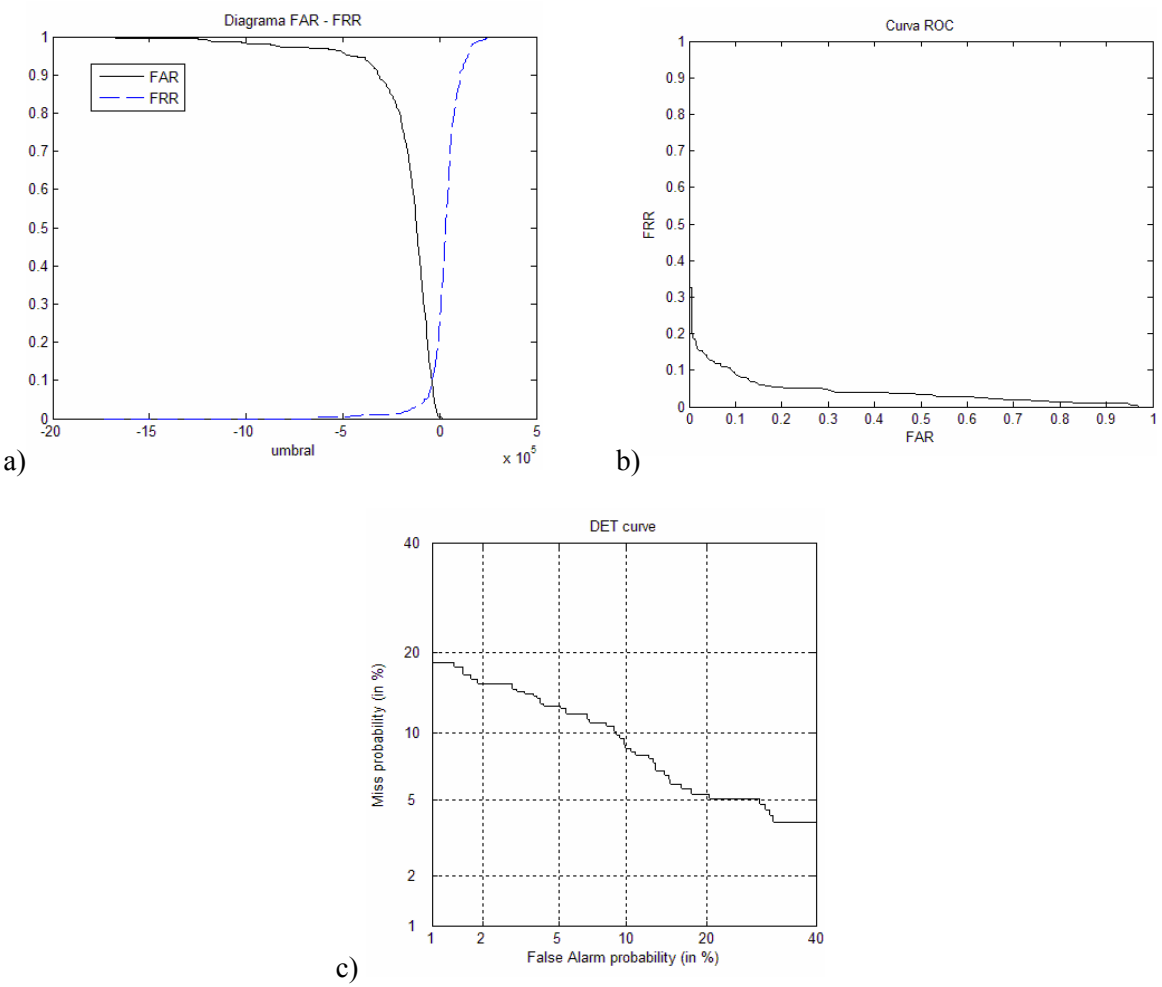


Figura 2-3: Representaciones gráficas de la relación entre FAR y FRR para un sistema de reconocimiento de locutor: diagrama FAR-FRR (a), curva ROC (b) y curva DET (c).

En algunos sistemas, se establece un coste para cada una de las tasas de error y se define una función de coste de detección (*DCF: Detection Cost Function*) (NIST website). Los parámetros incluidos en esta función son, además de FAR y FRR, los costes asociados a la falsa aceptación y al falso rechazo, C_{FA} y C_{FR} respectivamente, y las probabilidades a priori de las ocurrencias de cliente e impostor, P_C y $P_I=1-P_C$.

$$DCF = C_{FR} \cdot FRR \cdot P_C + C_{FA} \cdot FAR \cdot P_I \qquad (2.1)$$

Por otro lado, en un sistema de identificación, se utiliza como medida para la verificación del sistema la tasa de error de identificación (*FIR: False Identification Rate*) que se define como el número de errores de identificación dividido por el número total de pruebas de identificación.

En el caso en que un sistema de identificación se utilice como preclasificador para un sistema más complejo, también se utiliza como medida de evaluación el número de individuos que se deben seleccionar en cada ocurrencia para reducir la tasa de error de identificación hasta un cierto valor. Así, el sistema preclasificador selecciona los *N* individuos cuyos modelos obtienen las mayores puntuaciones y sólo se considera que hay un error cuando ninguno de los individuos seleccionados se corresponde con el usuario a identificar. De esta manera, manteniendo una degradación controlada del sistema, se simplifica la identificación mediante el sistema más complejo, que sólo precisa comparar la señal biométrica con los modelos de *N* individuos y no con los modelos del total de la base de datos.

Otras medidas de interés en los sistemas biométricos son:

- Error de adquisición (*FTA: Failure To Acquire*): Mide los errores en la adquisición de señales biométricas a procesar.
- Error de entrenamiento (*FTE: Failure To Enroll*): Mide el número de usuarios cuyo modelo no puede ser creado, por ejemplo, debido a limitaciones físicas del individuo.

2.5 Aplicaciones

La utilización de tecnologías de reconocimiento biométrico para la verificación o identificación de individuos está en aumento en diversas áreas y aplicaciones relacionadas con control de acceso, sistemas de embarque, control de fronteras, identificación civil, seguridad en redes y accesos electrónicos, etc.

Algunas de las principales aplicaciones de los sistemas biométricos, se detallan en los siguientes apartados.

Control de acceso físico

El control de acceso físico mediante biometría permite acceder a áreas restringidas o personales mediante la utilización de características biométricas. Este tipo de sistemas se utilizan en empresas, industrias, hospitales, instalaciones policíacas y militares, control de fronteras y aeropuertos, cerraduras biométricas, etc. debido a sus niveles de seguridad y eficacia.

Control de fronteras y aeropuertos

Las tecnologías biométricas proporcionan soluciones de identificación y seguridad para el control de fronteras y aeropuertos. El reconocimiento mediante caras, iris o huella dactilar, por ejemplo, son tecnologías en vías de expansión en aplicaciones de estas características.

En la última década, los gobiernos del mundo occidental han revisado la gestión de la seguridad en fronteras y aeropuertos y han incluido las tecnologías biométricas que proporcionan fiabilidad y la posibilidad de autenticar individuos con gran precisión incluso utilizando distintas bases de datos. Esto es importante dada la dificultad y complejidad de consolidar las múltiples plataformas que pueden existir en un estado y a nivel internacional.

Cerraduras biométricas

Siempre ha existido la necesidad de mantener seguro lo que es propio y evitar el acceso de personas no autorizadas, tanto a nuestros bienes como a nuestras viviendas, propiedades o negocios. Para resolver esta problemática, existen aplicaciones biométricas que permiten la apertura de accesos o puertas únicamente a aquellas personas autorizadas. Estas aplicaciones se denominan cerraduras biométricas.

Control de acceso lógico

El control de acceso lógico hace referencia a los controles electrónicos que limitan el acceso a ficheros de datos o programas, de manera que sólo los usuarios permitidos puedan acceder a ellos.

Esto es posible mediante la utilización de módulos de hardware y algoritmos que implementan la autenticación de los usuarios. Estos sistemas pueden llegar a obtener tasas de reconocimiento muy altas en función de las modalidades biométricas utilizadas por lo que son utilizados tanto por empresas y particulares como por servicios militares y gubernamentales.

Sistema sanitario

Uno de los ámbitos en que el control de acceso lógico mediante sistemas biométricos tiene un mayor auge es el de la sanidad, en que diferentes grupos de individuos pueden tener acceso a diferentes niveles de información de los pacientes y es fundamental que personal no autorizado no acceda a información crítica o confidencial.

La identificación biométrica, tanto de pacientes como de personal sanitario, puede evitar errores en la manipulación de historiales y medicación y garantiza la seguridad y el acceso a la información.

Aplicaciones financieras

También en los sistemas financieros tienen su utilidad las aplicaciones biométricas. Un ejemplo claro es la realización de operaciones en cajeros electrónicos, en que la utilización de

características biométricas en lugar de tarjetas de crédito asegura la identidad de la persona que utiliza el cajero, además de evitar los problemas derivados del robo o la pérdida de las tarjetas.

También en el caso de acceso electrónico a operaciones bancarias o de compras a través de la red, la utilización de características biométricas hace las transacciones más seguras y robustas ante el robo de usuarios y contraseñas o números de tarjetas de crédito.

Cliente y residencial

La biometría es una tecnología que ha evolucionado rápidamente de la utilización exclusiva por parte de gobiernos o grandes empresas hacia su aplicación en el mundo del consumo.

Las ventajas en este ámbito son amplias e incluyen entre otras aplicaciones el control de acceso a nuestros dispositivos con información personal como, por ejemplo, el teléfono móvil o la agenda electrónica, o la posibilidad de tener un producto según las preferencias o necesidades del individuo una vez que ha sido identificado.

Justicia y orden público

En el ámbito de la justicia y el orden público, las aplicaciones basadas en sistemas biométricos incluyen la administración e ingresos en prisiones, la identificación en escenas del crimen y otras aplicaciones forenses y programas nacionales o estatales de identificación de individuos.

Para ello, las tecnologías biométricas tienen la capacidad de reconocer huellas dactilares, iris, voz, caras, geometría de la mano, ADN, etc.

Tiempo de trabajo y control de presencia

Las aplicaciones biométricas pueden también utilizarse para identificar los lugares donde se encuentran individuos o grupos de individuos, de manera que se pueda controlar el tiempo de trabajo y el lugar donde se encuentran los individuos. Esto es útil para el control de personal en empresas y organismos gubernamentales así como para control del orden público y seguimiento o control de presos con permisos o libertad condicional.

En el ámbito empresarial, estas tecnologías permiten saber quien está haciendo qué y cuando y permiten a los responsables de amplios grupos de individuos saber quien está disponible en cada momento y lugar. De esta manera, el seguimiento de tiempo y control de presencia se simplifica enormemente.

Sistemas biométricos móviles

Para una gran cantidad de aplicaciones, tanto militares, de transporte público, control de aduanas, legales o comerciales, los sistemas biométricos móviles suponen una gran ventaja para las empresas, que deben proveer soluciones flexibles y prácticas.

Los gobiernos están desarrollando soluciones biométricas móviles para la expedición de pasaportes, tarjetas de seguridad social, registro de votaciones, etc. que permitan reducir los errores de identificación que puedan ocurrir con la utilización de tarjetas o carnés.

Las aplicaciones de justicia y orden público también utilizan soluciones biométricas móviles dado que aceleran los procesos de identificación de individuos en el lugar donde se produzca un incidente o altercado, ahorrando tiempo y recursos e identificando amenazas en tiempo real.

Sistemas de control de presencia de empleados, control de acceso y seguridad pueden también ser aplicaciones biométricas móviles utilizadas por empresas y negocios para aumentar su productividad y reducir costes.

2.6 Privacidad

La privacidad de los individuos es uno de los aspectos más sensibles en la utilización de sistemas biométricos. La privacidad se puede definir como el derecho del individuo a controlar la información sobre sí mismo, es decir, la capacidad que tiene la persona de limitar la utilización de su información personal.

La utilización de sistemas biométricos implicar captar y almacenar señales o modelos que representan la información personal del individuo y que se relacionan con su identidad. Por ello, los usuarios de estos sistemas pueden considerar que su inclusión en este tipo de aplicaciones menoscaba su derecho a la privacidad.

Por poner algunos ejemplos, el reconocimiento de huellas dactilares se relaciona con la implicación en delitos y la utilización de modalidades biométricas en general con la posibilidad de crear un *Big Brother* global que controle el comportamiento de los individuos.

Las siguientes características hacen que un sistema pueda resultar menos invasivo con la privacidad de los individuos (IBG BioPrivacy Initiative website).

- Información a los usuarios sobre la utilización y finalidad del sistema.
- Adscripción voluntaria de los individuos.
- Los sistemas de verificación son menos invasivos con la privacidad que los de identificación.
- Almacenamiento de los datos biométricos en dispositivos propiedad del usuario en lugar de en bases de datos.

- Almacenamiento de modelos y no de señales biométricas como imágenes o registros de voz.
- Encriptación de los datos almacenados.

De todas formas, no hay que perder de vista que los sistemas biométricos, utilizados de forma adecuada, sirven para proteger la seguridad y la privacidad de las personas y las entidades a las que pertenecen las personas (Woodward, 1997).

Para lograr su correcta utilización, la Asociación Internacional de la Industria de la Biometría (International Biometric Industry Association website) ha establecido una serie de principios sobre privacidad que deben seguirse para que se respete este derecho. Entre ellos podemos destacar:

- Recomendación de establecer mecanismos de seguridad para que los datos biométricos no sean utilizados de forma malintencionada.
- En el sector privado, desarrollar políticas y procedimientos para informar a los usuarios sobre como los datos son recogidos, almacenados y utilizados, y para preservar el derecho de los individuos a que sus datos no se distribuyan más allá de lo estipulado.
- En el sector público, establecer estándares legales que definan claramente y limiten las condiciones bajo las que las fuerzas de seguridad pueden adquirir, acceder, almacenar y utilizar datos biométricos.
- En ambos sectores, adoptar controles técnicos y de gestión para proteger la confidencialidad y la integridad de las bases de datos biométricas.

3 El reconocimiento multimodal

El reconocimiento multimodal consiste en la utilización de diversas (más de una) características o modalidades biométricas en un único sistema de reconocimiento de personas.

De esta manera, en los sistemas de reconocimiento multimodal se combinan, por ejemplo, características de voz y caras para desarrollar sistemas de reconocimiento más eficaces, robustos y seguros.

Para ello, es necesario obtener las características biométricas de las diferentes modalidades, normalizarlas para que dichas características sean comparables y realizar un proceso de fusión que las unifique en un único sistema de reconocimiento que permita tomar una decisión sobre la identidad de un individuo. En contraposición a los sistemas multimodales, se denominan sistemas unimodales a aquellos que hacen uso de una sola información biométrica.

Las posibles características a fusionar son el resultado de los diferentes procesos involucrados en el reconocimiento biométrico, es decir, los sistemas multimodales pueden fusionar señales biométricas, parámetros, puntuaciones obtenidas por la comparación de los parámetros con los modelos de los usuarios o las decisiones de los diferentes sistemas unimodales.

En cuanto al origen de las características a fusionar, algunos autores hablan de sistemas multibiométricos de forma genérica, distinguiendo los sistemas multisensoriales, que en el proceso de fusión utilizan múltiples sensores para la misma característica, multialgorítmicos, que usan múltiples algoritmos para las mismas señales, multiinstancia, múltiples instancias de la misma biometría, y multimodales, cuando se utilizan múltiples rasgos o modalidades biométricos (Ross, 2006). Otros autores, utilizan el término multimodal de forma genérica para designar cualquiera de estos tipos de fusión (Wayman, 2006). En esta tesis, se va a utilizar el término multimodal según esta última tendencia, dado que es el uso que se le da a este vocablo dentro de nuestro grupo de investigación.

En este apartado se revisarán las ventajas del reconocimiento multimodal, los diferentes niveles de fusión biométrica, las técnicas de normalización y fusión de las diferentes características biométricas y la manera de evaluar el rendimiento de los sistemas de reconocimiento multimodal.

3.1 Ventajas del reconocimiento multimodal

La utilización de diversas modalidades biométricas en un solo sistema de reconocimiento permite, en primer lugar, mejorar los resultados obtenidos por una sola modalidad o, lo que es equivalente, mejora la fiabilidad del sistema. Esto permite, por ejemplo, aumentar la seguridad de un sistema de reconocimiento manteniendo una buena tasa de falso rechazo (Bolle et al., 2004; Ross et al., 2006; Vacca, 2007; Sanderson, 2008).

Otra característica biométrica que mejora mediante la utilización de más de una modalidad es la universalidad ya que, aunque un individuo no posea alguna de las modalidades biométricas utilizadas por el sistema de reconocimiento, difícilmente no poseerá ninguna de las características que formen parte de un sistema multimodal. En este caso, disminuye el error de entrenamiento (FTE).

Otra ventaja de los sistemas multimodales es que aumenta la robustez del sistema en el caso de que alguna de las características biométricas se vea afectada por ruidos, distorsiones o por su propia variabilidad. Incluso en el caso extremo de que una modalidad no pudiera ser utilizada porque el individuo no poseyera la característica biométrica correspondiente o por problemas en los sensores, los sistemas multimodales se pueden adaptar para que funcionen únicamente con las modalidades disponibles.

También la resistencia a ataques malintencionados (*spoof attacks*) mejora con la utilización de diversas modalidades biométricas dado que es más difícil simular diversas características o engañar a diversos sistemas que a uno solo.

3.2 Normalización y fusión

Los sistemas de fusión multimodales combinan las características de diferentes sistemas biométricos unimodales y obtienen un resultado global de reconocimiento. Para conseguirlo es necesario realizar un proceso de fusión de las diferentes características que convierta la información multimodal en una decisión sobre la identidad de una persona.

Sin embargo, en la mayoría de los casos, el origen de las características biométricas es diverso, dado que pertenecen a diferentes modalidades y, por lo tanto, en el proceso de obtención se utilizan

los sensores correspondientes a cada biometría y los algoritmos de obtención de características más conveniente para cada sistema unimodal de reconocimiento.

Esto implica que las diferentes características involucradas en la fusión multimodal no son homogéneas y, en consecuencia, en la mayoría de los casos, no es posible utilizarlos en los procesos de fusión sin realizar un proceso de homogeneización de dichas características. A este proceso de homogeneización se le denomina ***normalización***. Por lo tanto, el proceso de fusión multimodal es en realidad un proceso en dos pasos, la normalización y la fusión propiamente dicha.

En consecuencia, en un sistema de fusión multimodal, cualquiera que sea la técnica a aplicar para la fusión de las modalidades biométricas, es necesario realizar una normalización previa de los resultados obtenidos por los sistemas unimodales que los prepare para los procesos de fusión (Wan et al., 2005), aunque en algunos casos esta normalización se realiza de forma implícita en la obtención de los parámetros o puntuaciones.

Diversos trabajos muestran ejemplos del efecto de la normalización sobre las puntuaciones de las diferentes modalidades biométricas. En (Jain, 2005) se muestras los histogramas de la figura 3-1 para las puntuaciones de clientes e impostores de sistemas de reconocimiento de caras, huellas dactilares y geometría de la mano.

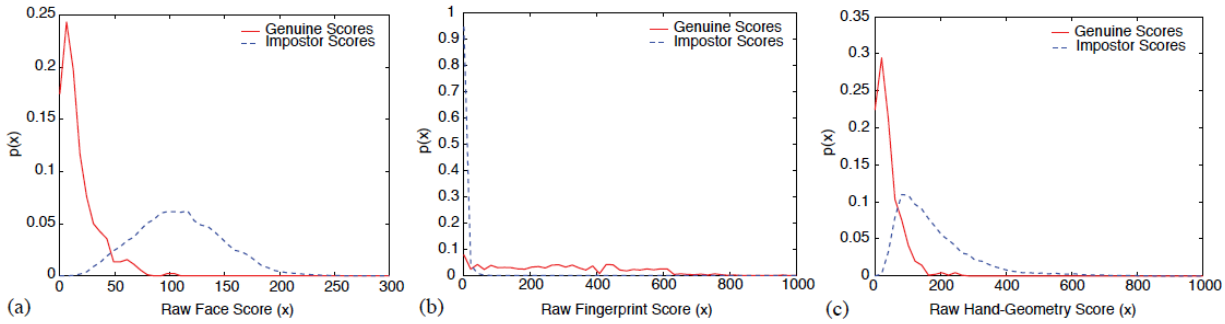


Figura 3-1: Histogramas de las puntuaciones para los sistemas de reconocimiento de caras (a), huellas dactilares (b) y geometría de la mano (c).

Se puede observar que el rango de las puntuaciones es muy diferente para las caras y para las huellas dactilares o la geometría de las manos. Además, en el caso de las caras y la geometría de la mano, las puntuaciones de mayor valor corresponden a los impostores mientras que en las huellas dactilares corresponden a los clientes.

En la figura 3-2, obtenida de la misma fuente, se pueden observar los histogramas de estas puntuaciones normalizadas con la técnica convencional *min-max*, que mapea de forma lineal las

puntuaciones de cada modalidad entre 0 y 1. En este caso, las puntuaciones se encuentran entre 0 y 1 en todos los casos y las puntuaciones de mayor valor corresponden a las de los clientes.

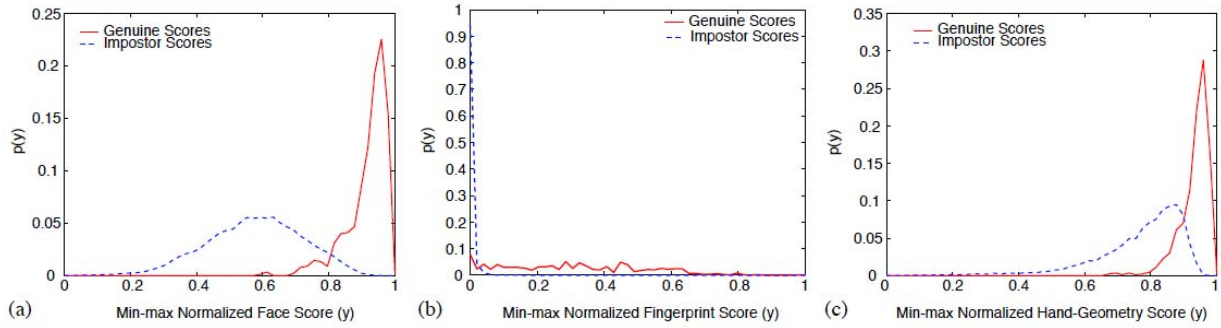


Figura 3-2: Histogramas de las puntuaciones normalizadas mediante *min-max* para los sistemas de reconocimiento de caras (a), huellas dactilares (b) y geometría de la mano (c).

La importancia de la normalización de las características unimodales no es exclusiva de la fusión de puntuaciones, sino que es aplicable a los diferentes niveles de fusión que se presentan en el próximo apartado y se amplían en los siguientes, detallando además las diferentes técnicas de normalización y fusión del estado del arte.

3.3 Niveles de fusión multimodal

Existen cuatro niveles de fusión en los que realizar la fusión multimodal a partir de diferentes características unimodales: a nivel de sensor en el caso en que se combinan las señales obtenidas a partir de las características biométricas, a nivel de los parámetros de cada una de las modalidades biométricas unimodales, a nivel de los resultados de reconocimiento proporcionados por sistemas independientes para cada una de las biometrías y a nivel de la decisión tomada por cada uno de estos sistemas (Baker, 2002; Daugman, 1999).

La figura 3-3 muestra el esquema de reconocimiento de dos o más sistemas biométricos unimodales y los niveles a los que se puede diseñar un sistema de reconocimiento multimodal.

En la fusión a nivel de sensor, las señales obtenidas por diferentes sensores se unifican en una representación conjunta de la característica biométrica. Es el caso de la representación facial en 3D a partir de fotografías de la cara desde diferentes ángulos. Este tipo de fusión no suele involucrar diferentes modalidades biométricas, por lo que no será objeto de estudio en esta tesis.

En la fusión de biometrías a nivel de parámetros se combinan los parámetros de las diferentes modalidades biométricas y se utilizan como si procedieran de una sola fuente para entrenar y probar un nuevo sistema de reconocimiento.

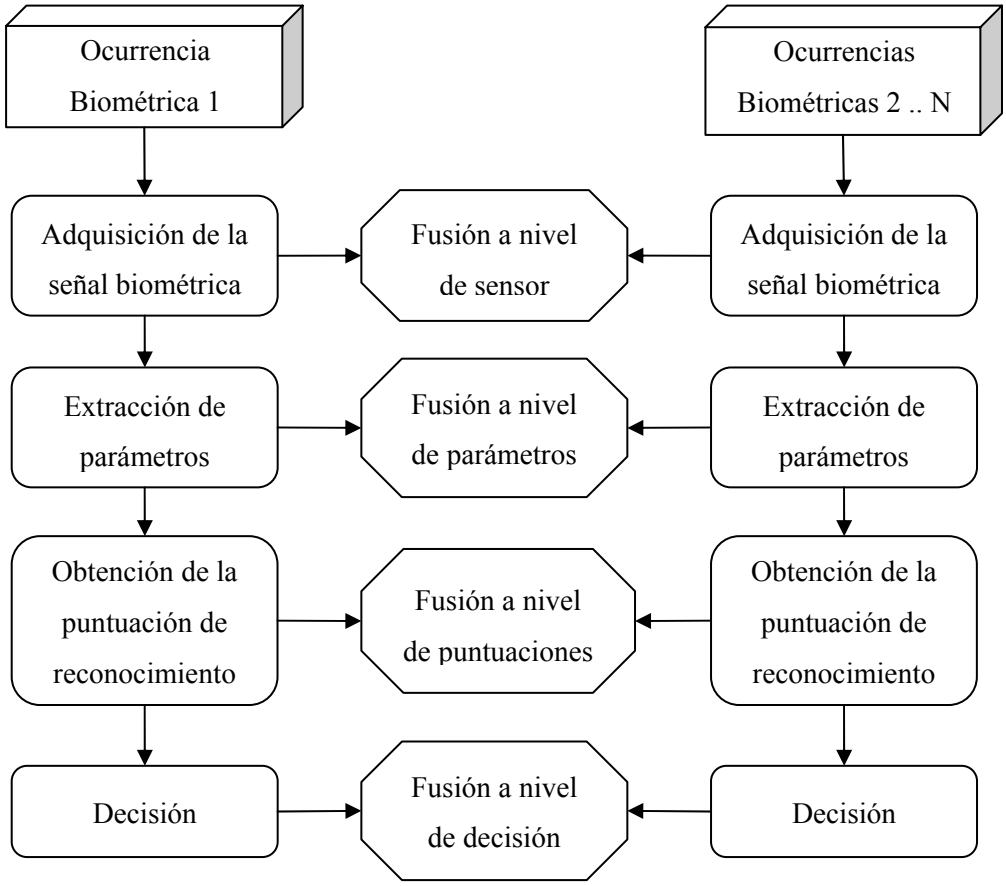


Figura 3-3: Niveles de fusión multimodal de biometrías

Este tipo de fusión reúne toda la información relevante de cada uno de los sistemas de reconocimiento unimodales. Esto garantiza que no existe pérdida de información en el paso hacia la biometría multimodal. Sin embargo, esto también se puede convertir en un punto débil, si la información llega a ser de una dimensión tan grande que sea difícil de tratar. Otro posible inconveniente de este tipo de fusión es que las diferentes informaciones paramétricas pueden ser incoherentes.

En la fusión a nivel de puntuaciones de reconocimiento se toma una decisión a partir de los resultados obtenidos por sistemas independientes para cada biometría. Las dos aproximaciones más habituales para realizar está fusión consisten en combinar dichos resultados para obtener un valor de confianza único y en diseñar sistemas de clasificación a partir de los vectores formados por estos resultados de reconocimiento.

En este caso se reduce la dimensión del vector multimodal, al utilizar un solo valor por sistema de reconocimiento, lo que simplifica el sistema de fusión. Además, es probable que este resultado contenga la mayor parte de la información relevante que pueda proporcionar cada biometría dado que es con este valor con el que se toma la decisión en el sistema unimodal.

Otro punto a considerar es que la fusión a nivel de puntuaciones permite añadir nuevos expertos de una manera más simple que la fusión a nivel de parámetros y permite fusionar con mayor simplicidad biometrías con diferente sincronía temporal como, por ejemplo, el audio y el vídeo con las huellas dactilares.

En la fusión a nivel de decisión se utilizan las decisiones de sistemas biométricos independientes para obtener una decisión final sobre la identidad de la persona reconocida.

En este tipo de fusión se pierde una cantidad significativa de información respecto al caso anterior dado que se convierten las puntuaciones de reconocimiento en decisiones antes de la fusión (Kittler et al., 2003). En este caso, el número de sistemas de reconocimiento debe ser mayor o igual que el número de modalidades biométricas. Esto es razonable para verificación pero puede llegar a ser inviable en el caso de identificación de personas.

La mayor parte de los trabajos presentados en fusión biométrica están orientados a la fusión a nivel de puntuaciones de reconocimiento. Esto es debido a diversos factores. En primer lugar, en la fusión a nivel de parámetros es necesario acceder a dichos parámetros mientras que en los otros dos tipos de fusión sólo es necesario acceder al resultado obtenido por un sistema de reconocimiento cerrado. Además, la fusión a nivel de decisión aprovecha menos la información proporcionada por dichos sistemas que la fusión a nivel de puntuaciones de reconocimiento.

Diversos trabajos comparan la fusión a diferentes niveles. Por ejemplo, en (Campbell et al., 2003a) se fusionan una gran cantidad de parámetros de alto nivel incluyendo modelos de pronunciación, dinámica de la prosodia, características de la entonación, cadenas de fonemas e interacciones conversacionales mediante diferentes técnicas a diferentes niveles de fusión. El perceptrón y una *GMM* produjeron menores errores de reconocimiento que otras fusiones como *MLP, radial basis function, k-nearest neighbor, binary tree* y *Support Vector Machines*.

En (Fox et al., 2003a) se compara también la fusión a diferentes niveles basándose en parámetros audio-visuales dinámicos.

De todas formas, se debe tener en cuenta que los diferentes niveles de fusión no son excluyentes, de manera que se pueden diseñar sistemas que combinen la fusión a diferentes niveles a partir de la información proporcionada por diferentes modalidades biométricas. Por ejemplo, en (Nefian et al., 2003) se entrenan *Coupled Hidden Markov Models* (*CHMM*) mediante la secuencia temporal de observaciones de voz y de la forma de la boca y, posteriormente, la probabilidad obtenida por cada persona de la base de datos se combina a nivel de puntuaciones de reconocimiento con la

información proveniente de un sistema de reconocimiento de caras mediante un *Embedded Hidden Markov Model* (*EHMM*).

En los siguientes apartados revisaremos la fusión a nivel de parámetros, de puntuaciones y de decisión, incluyendo las técnicas de normalización y fusión del estado del arte.

3.4 Fusión a nivel de parámetros

En la fusión a nivel de parámetros (*features*), los parámetros obtenidos por los diferentes sistemas de extracción de características biométricas se unen para crear un único vector de parámetros.

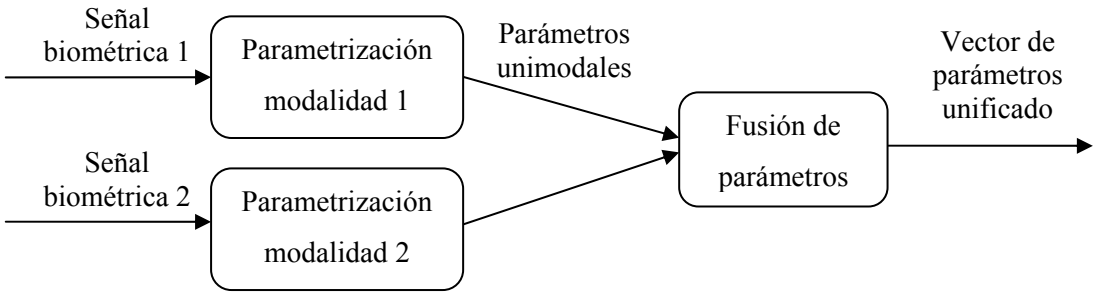


Figura 3-4: Esquema de fusión a nivel de parámetros.

Los nuevos vectores serán de una dimensión mayor a los obtenidos por cada una de las biometrías independientemente y el sistema de reconocimiento desarrollado a partir de ellos tendrá, probablemente, la capacidad de discriminar mejor entre las diferentes personas.

Sin embargo, la dimensión del vector obtenido puede ser muy grande o puede existir correlación entre los parámetros de las distintas modalidades biométricas. Por ello, en muchas ocasiones, es conveniente utilizar técnicas de reducción de parámetros para obtener aquellos más representativos.

Para realizar la fusión a nivel de parámetros se pueden utilizar los sistemas de reconocimiento habituales para las biometrías unimodales, como el reconocimiento de patrones (*Pattern Recognition*) (Theodoridis et al., 2003), los modelos ocultos de Markov (*HMM*) y los sistemas de mezcla de gaussianas (*GMM*) (Rabiner, 1989).

Por poner algunos ejemplos, en (Feng et al., 2004) se utilizan estas técnicas para realizar la fusión de los parámetros obtenidos a partir de caras y huellas dactilares. En (Rattani et al., 2009) se ha utilizado una selección de parámetros SIFT de caras e iris para crear un supervector con el que realizar la fusión. En (Bengio, 2003), se modela la función de probabilidad conjunta de pares de

secuencias asíncronas de voz y vídeo describiendo el mismo evento a partir de los parámetros de estas biometrías.

La fusión a nivel de parámetros permite, además, aprovechar el sincronismo de las biometrías involucradas como, por ejemplo, en la identificación o verificación de una persona mediante la voz, la cara y el movimiento de los labios durante la pronunciación de una frase, como en (Chen, 2001; Chaudhari et al., 2003a; Chaudhari et al., 2003b; Potamianos et al., 2004) o aprovechar la correlación entre los parámetros faciales y del habla como en (Shah et al., 2009). En (Fu, 2003) y (Nefian et al., 2003) se utilizan *Coupled Hidden Markov Models (CHMM)* para la identificación de locutores a partir de información audio-visual.

Además, también se utilizan otras técnicas como, por ejemplo, las *SVM*s, para realizar la fusión a este nivel. Tanto en (Gutschoven et al., 2000) como en (Hatch et al., 2005) se utiliza esta técnica de clasificación para la combinación de conjuntos de parámetros en aplicaciones de reconocimiento de personas.

3.4.1 Normalización de parámetros

El número de características obtenidas en los procesos de parametrización de los sistemas unimodales puede ser mayor que el que los sistemas de modelización y reconocimiento pueden soportar para tener un buen rendimiento. Por ello, en algunos casos, es necesario aplicar sistemas de reducción de parámetros que permitan disminuir la complejidad de los sistemas de reconocimiento.

Esto también sucede en el caso de la fusión a nivel de parámetros, en que se unen los parámetros procedentes de diversas modalidades biométricas en un vector multimodal.

Algunas de las técnicas de reducción de parámetros más utilizadas son el análisis de componentes principales (PCA: Principal Component Analysis) y el análisis discriminante lineal (LDA: Linear Discriminant Análisis).

El análisis de componentes principales realiza una transformación lineal de los parámetros y las características obtenidas resultan independientes entre sí. Para ello, se utiliza la matriz de correlaciones de los parámetros. La elección de los factores se realiza de tal forma que el primero recoja la mayor proporción posible de la variabilidad original; el segundo factor debe recoger la máxima variabilidad posible no recogida por el primero, y así sucesivamente. Del total de factores se elegirán aquéllos que recojan el porcentaje de variabilidad que se considere suficiente. A éstos se les denominará componentes principales.

En el caso del análisis discriminante lineal se asume que las funciones de densidad de probabilidad de los parámetros son normales y se aplica la estadística Bayesiana. El discriminante lineal de

Fisher está relacionado con la técnica anterior aunque no asume distribuciones normales ni covarianzas iguales entre las clases. En ambos casos, el objetivo es obtener un subespacio que contenga la mayor parte de la variabilidad de los parámetros iniciales.

Un ejemplo de la utilización de estas técnicas es la obtención de los parámetros denominados eigenfaces y fisherfaces a partir de las imágenes de caras, en que se aplica respectivamente PCA y discriminante lineal de Fisher sobre los datos obtenidos de las caras (Belhumeur et al., 1997).

Todas estas técnicas de reducción de parámetros, de hecho, producen nuevos conjuntos de parámetros, que se pueden considerar normalizados ya que, por ejemplo, en el caso del PCA, como ya hemos dicho, los parámetros resultantes son independientes.

En otros casos, se utilizan transformaciones espectrales como la transformada de Fourier o la transformación cepstrum en el caso de la voz o transformaciones mediante wavelets. Todas estas técnicas permiten también limitar el número de parámetros a la vez que los preparan para los procesos de fusión.

Sin embargo, también existen técnicas que permiten la normalización de estos parámetros, como la sustracción de medias cepstrales (*CMS: cepstral mean substraction*) que reduce el ruido estacionario debido al canal o el modelado de parámetros (*feature warping*), que consiste en el mapeo de la distribución de los parámetros cepstrales a una distribución normal dentro de una ventana que se desplaza en el tiempo.

Sin embargo, en el caso de la fusión multimodal, en que los parámetros se obtienen mediante distintas técnicas de extracción, puede resultar conveniente aplicar de forma adicional una normalización que uniformice las características de todos los parámetros. Por ejemplo, en (Wang et al., 2009) se utiliza la técnica *z-score* para la normalización de parámetros antes de la fusión. Para realizar este proceso de normalización se pueden utilizar las mismas técnicas que en el caso de la normalización de puntuaciones y que se presentan en el apartado 3.5.1.

3.4.2 Fusión de parámetros

Las técnicas más habitualmente utilizadas dentro de la clasificación entrenada son el clasificador Bayesiano (Langley et al., 1992), los árboles de decisión, las redes neuronales (*Neural Networks*) como el *Multilayer Perceptron (MLP)* (Bishop, 1995), los modelos ocultos de Markov (*HMM*), los modelos de mezclas de gaussianas (*GMM*) (Rabiner, 1989) y las *Support Vector Machines (SVMs)* (Cristianini et al., 2000).

Todas estas técnicas pueden ser utilizadas para realizar la fusión a nivel de parámetros. Por ejemplo, en (Bengio, 2003) se utilizan modelos ocultos de Markov asíncronos a partir de secuencias asíncronas de voz y vídeo y en (Fu, 2003) y (Nefian et al., 2003) se utilizan *Coupled*

Hidden Markov Models (CHMM) para la identificación de locutores a partir de información audio-visual.

En (Gutschoven et al., 2000) y en (Hatch et al., 2005) se clasifican conjuntos de parámetros en aplicaciones de reconocimiento de personas mediante *SVMs*. El objetivo de las *SVM*s es proporcionar una manera computacionalmente eficiente de aprender hiperplanos capaces de clasificar los datos en un espacio de parámetros con un gran número de dimensiones (Cristianini et al., 2000; Vapnik, 2000; kernel machines website).

Para ello se diseñan hiperplanos clasificadores que maximizan la distancia entre dicho hiperplano y los vectores de parámetros más próximos. La utilización de un kernel permite la transformación de los parámetros a un espacio diferente y, por lo general, de mayor dimensión que el espacio de los parámetros origen (Vapnik, 2000; Burges, 1998; Cristianini et al., 2001; kernel machines website). De esta manera, se puede diseñar el hiperplano clasificador en un espacio en que los datos estén mejor separados en función de su clase (Awad et al., 1998; Hearst, 1998).

3.5 Fusión a nivel de puntuaciones

En la fusión a nivel de puntuaciones (*scores*) cada sistema biométrico unimodal aporta una medida de confianza que se ha obtenido por comparación de los parámetros biométricos con los modelos de las diferentes personas. La fusión a nivel de puntuaciones se compone de dos procesos principales: normalización y fusión propiamente dicha.

El proceso de normalización transforma los resultados de las diferentes modalidades biométricas de manera que se encuentren en un rango equivalente o comparable. Si no se realizara esta transformación, una modalidad biométrica con un amplio rango de valores podría eliminar la contribución de otra modalidad con un rango menor.

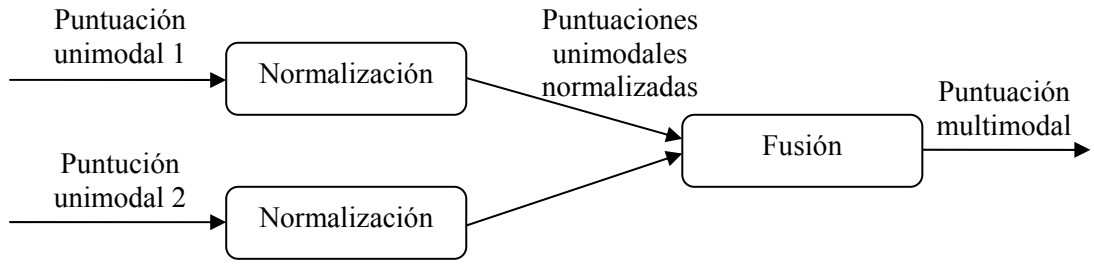


Figura 3-5: Esquema de fusión a nivel de puntuaciones de reconocimiento.

El proceso de fusión obtiene una puntuación multimodal y una decisión a partir de las puntuaciones normalizadas de las biometrías. Existen dos enfoques principales para la obtención de esta decisión: la combinación mediante cálculos de las puntuaciones obtenidas por cada sistema unimodal y el entrenamiento de modelos de clasificación más complejos a partir del vector de valores formado por estas puntuaciones.

La figura 3-5 muestra el esquema de la fusión a nivel de puntuaciones de reconocimiento donde se representan los procesos de normalización y de fusión. Una vez obtenida la puntuación multimodal ésta se puede tratar de la misma manera que la de un sistema unimodal, comparándola con un umbral o con otras puntuaciones multimodales para tomar una decisión de reconocimiento.

Son diversos los trabajos en que se comparan las diferentes técnicas de normalización y se comprueba su eficacia al combinarlas con diversas técnicas de fusión. Por ejemplo, en (Snelick et al., 2003), se utiliza información de caras y huellas dactilares para realizar la comparación y también (Snelick et al., 2005) hace una comparativa del estado del arte en la fusión multimodal. En (Jain et al., 2005) se comparan las técnicas convencionales de normalización y se demuestra la sensibilidad de las técnicas *min-max* y *z-score* a la existencia de *outliers*, valores que quedan fuera de los resultados habituales para una biometría, y la resistencia a este problema de los estimadores de *tanh*.

Bolle et al. proponen en (Bolle et al., 2000) utilizar intervalos de confianza para dar validez a los resultados de reconocimiento obtenidos por cada biometría unimodal. De esta manera, además de normalizar los resultados de reconocimiento se tiene en cuenta la validez de cada uno de ellos.

3.5.1 Normalización de puntuaciones

Los sistemas unimodales utilizan diversas técnicas para la normalización de puntuaciones. Por ejemplo, en el caso de las puntuaciones de voz existe toda una familia de técnicas donde las puntuaciones se normalizan restando la media y dividiendo por la desviación estándar, estimadas ambas a partir de la distribución de puntuaciones de impostores. Es el caso de la normalización cero (ZNorm: Zero Normalization), ampliamente utilizada en verificación de locutor, donde la media y la desviación estándar se estiman a partir de distribuciones dependientes de usuario (Auckenthaler et al., 2000). En este caso, la estimación de los parámetros de la normalización se puede hacer de manera previa al reconocimiento. Otras de estas técnicas son HNorm, donde se procesa de forma independiente la información de cada tipo de terminal telefónico, o TNorm, en cuyo caso la media y la varianza es dependiente de cada señal de test. También existe la normalización HTNorm, combinación de las dos anteriores o la CNorm, específica para comunicaciones telefónicas móviles. Todas estas normalizaciones modifican el rendimiento de los sistemas unimodales (Auckenthaler et al., 2000; Bimbot et al., 2004).

En lo que refiere a esta tesis, suponemos que las técnicas unimodales utilizadas están optimizadas en su rendimiento y que la mejora de los sistemas unimodales de extracción de parámetros o puntuaciones por medio de la normalización queda fuera del alcance de esta tesis. Por este motivo, se han utilizado y desarrollado normalizaciones monótonamente crecientes, de manera que los resultados de reconocimiento obtenidos de forma individual por cada parámetro o puntuación se mantienen invariables.

Algunas de las técnicas de normalización de puntuaciones más comúnmente utilizados que cumplen la condición anterior son, *min-max*, *z-score*, que sería el equivalente a las normalización de media y desviación estándar utilizadas en la verificación de locutor, los métodos adaptativos basados en formas cuadráticas, *QQ* y *QLQ* y normalizaciones basadas en la función tangente hiperbólica (Snelick et al., 2005).

La normalización *min-max* (*MM*) transforma de forma lineal las puntuaciones de una biometría de forma que su menor valor sea cero y el mayor sea uno. La ecuación (3.1) muestra la manera de realizar la normalización *MM*, donde x_{MM} es la biometría normalizada, A es la modalidad biométrica original y *min(A)* y *max(A)* son los puntos extremos del rango de puntuaciones.

$$x_{MM} = \frac{a - \min(A)}{\max(A) - \min(A)} \tag{3.1}$$

La normalización *z-score* (*ZS*) modifica la media y la varianza globales de las puntuaciones de una biometría multimodal para que valgan cero y uno respectivamente. La ecuación (3.2) muestra la transformación a aplicar para realizar la normalización *z-score*, donde x_{ZS} es la biometría normalizada, A es la biometría unimodal original, *mean(A)* es la media estadística de las puntuaciones y *std(A)* es su desviación estándar. La media y la varianza se pueden calcular sobre el total de las puntuaciones o únicamente sobre las puntuaciones de impostores.

$$x_{ZS} = \frac{a - mean(A)}{std(A)} \tag{3.2}$$

Indovina et al. presentaron en (Indovina et al., 2003) diversos métodos de normalización adaptativos con la finalidad de reducir el solapamiento de los lóbulos de clientes e impostores y, en consecuencia, incrementar el rendimiento del sistema de reconocimiento. Estas normalizaciones dependen del centro (c) y la amplitud (w) de la región de solapamiento.

La figura 3-6 muestra el histograma de las puntuaciones de clientes e impostores y la región de solapamiento para una modalidad biométrica.

Dos de estos métodos adaptativos, los que obtuvieron mejores resultados, son dos-cuádricas (*QQ: two-quadrics*) y cuádrica-línea-cuádrica (*QLQ: quadric-line-quadric*) y se deben aplicar sobre las puntuaciones normalizadas con *min-max*.

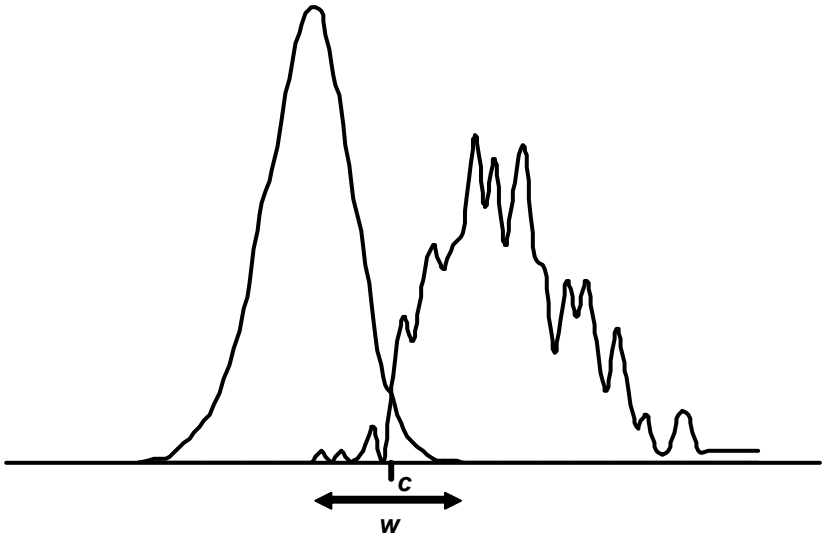


Figura 3-6: Solapamiento de los lóbulos de clientes e impostores en las puntuaciones de una modalidad biométrica.

QQ se compone de dos cúadricas que cambian su concavidad en el centro de la región de solapamiento (c). La relación entre las puntuaciones normalizadas x_{QQ} y las puntuaciones normalizadas mediante *min-max* x_{MM} es la representada en la ecuación (3.3).

$$x_{QQ} = \begin{cases} \frac{1}{c} x_{MM}^2, & x_{MM} \le c \\ c + \sqrt{(1-c)(x_{MM} - c)} & otherwise \end{cases} \tag{3.3}$$

En la normalización *QLQ* la zona de solapamiento se mantiene intacta mientras las otras regiones se transforman mediante dos funciones cuadráticas. La relación entre las puntuaciones biométricas normalizadas x_{QLQ} y las puntuaciones normalizadas mediante *min-max* x_{MM} se muestra en la ecuación (3.4).

$$x_{QLQ} = \begin{cases} \frac{1}{\left(c - \frac{w}{2}\right)} x_{MM}^2, & x_{MM} \le \left(c - \frac{w}{2}\right) \\ x_{MM}, & \left(c - \frac{w}{2}\right) < x_{MM} \le \left(c + \frac{w}{2}\right) \\ \left(c + \frac{w}{2}\right) + \sqrt{\left(1 - c - \frac{w}{2}\right)\left(x_{MM} - c - \frac{w}{2}\right)}, & otherwise \end{cases} \tag{3.4}$$

La normalización *tanh* (Jain et al., 2005) proyecta las puntuaciones en el rango [-1, 1] mediante una transformación no lineal. Las puntuaciones cercanas a la media se transforman de forma cuasi-lineal mientras que para los valores altos y bajos de las puntuaciones se realiza una compresión de los datos. La normalización x_{TANH} se calcula como

$$x_{TANH} = \frac{1}{2}\left\{\tanh\left(k\frac{a-\mu_{GH}}{\sigma_{GH}}\right)+1\right\} \tag{3.5}$$

donde μ_{GH} y σ_{GH} son, respectivamente, la media y la desviación estándar de las puntuaciones de clientes según la estimación de Hample (Hampel et al., 1986), y k es una constante ajustable. La principal ventaja de esta normalización es la reducción del efecto de los valores fuera de rango o *outliers*, que son absorbidos por la compresión de los valores extremos.

Otras técnicas de normalización comunmente utilizadas son el escalado decimal, la desviación absoluta de media (*MAD: Median Absolute Deviation*) o la función de doble sigmoide. Sin embargo, se ha demostrado que estas técnicas son, por lo general, menos eficientes que las normalizaciones basadas en *tanh* (Jain et al., 2005).

Más recientemente, se ha propuesto la estimación de la función densidad de probabilidad de las puntuaciones de las modalidades biométricas como un primer paso en la normalización de dichas puntuaciones. Una de estas propuestas (Jain et al., 2005) utiliza el método de estimación de la ventana de Parzen (Duda et al., 2001) para transformar las puntuaciones en probabilidades.

3.5.2 Fusión de puntuaciones

En la fusión a nivel de puntuaciones pueden utilizarse tanto combinaciones aritméticas o lógicas de las puntuaciones como otras técnicas de fusión entrenada de mayor complejidad (Duin, 2002; Rolli et al., 2002b). El primer grupo de técnicas son más fáciles de implementar mientras que el segundo puede proporcionar mayor adaptación, sobretodo, en sistemas complejos.

3.5.2.1 *Combinación aritmética o lógica*

Las técnicas de fusión mediante combinación de puntuaciones de reconocimiento realizan operaciones aritméticas o estadísticas sobre dichos resultados para obtener una medida de confianza que permita tomar una decisión global de reconocimiento.

Para ello, se suelen realizar sumas o productos ponderados de los resultados de cada biometría, como en la suma simple (*SS: simple sum*) o en la ponderación en función del resultado de reconocimiento de cada biometría unimodal (*MW: matcher weighting*), o se realizan operaciones estadísticas como en el caso del cálculo del mínimo o el máximo valor de reconocimiento (*MIN: minimum y MAX: maximum*) (Indovina et al., 2003).

La clave para el buen funcionamiento de estas técnicas es la elección de los parámetros de ponderación de cada biometría. Por ello, se han presentado diversos trabajos en que se proponen técnicas para obtener estos valores. Este es el caso en (Maison et al., 2001) donde se exploran diversas técnicas para determinar de manera óptima los pesos a partir de la información audio-

visual obtenida de diversas grabaciones de noticiarios, o en (Fox et al., 2005) donde se utilizan resultados de reconocimiento provenientes de caras, boca e información acústica buscando valores óptimos para la fusión lineal. En (Jain et al., 2002) se evalúa la importancia de la elección de los umbrales y los pesos que se deben utilizar para cada biometría.

Sin embargo, la utilización de cada una de estas técnicas no es excluyente dado que es posible el diseño de sistemas híbridos, en que se utilizan diferentes técnicas en diferentes rangos de valores de reconocimiento con el objetivo de aprovechar aquellas que obtengan un mejor rendimiento en cada rango. En (Lucey et al., 2003) y en (Fox et al, 2003b) se utilizan fusiones mediante suma, producto y combinación híbrida de las dos, para el reconocimiento audio-visual de locutor.

Todas estás técnicas son también aplicables de forma dependiente de la persona, de manera que los parámetros de fusión se entrenan para cada usuario de manera independiente. Así, por ejemplo, en (Indovina et al., 2003) se obtienen mejoras en el reconocimiento mediante modelos dependientes de usuarios. En este trabajo, el cálculo de la ponderación dependiente de usuario se basa en el concepto de lobos y corderos presentado por Doddington et al. (Doddington et al., 1998) para biometrías unimodales basadas en voz.

3.5.2.2 *Otras técnicas de clasificación entrenada*

La otra gran categoría de técnicas de fusión es la que utiliza clasificadores entrenados más complejos que una simple combinación lineal de puntuaciones. Las técnicas utilizadas para este fin son muy diversas tanto para la clasificación entre los diferentes usuarios en el caso de la identificación como para decidir entre cliente o impostor en el caso de la verificación.

En (Ben-Yacoub et al., 1999), (Verlinde et al., 2000) y (Roli et al., 2002a) se presentan una amplia representación de ellas en el entorno audio-visual. En concreto, en (Ben-Yacoub et al., 1999) destaca la utilización de *SVM, MLP*, árbol de decisión *C4.5*, discriminante lineal de Fisher, y clasificador Bayesiano.

Por otro lado, en (Verlinde et al., 2000) se evalúa el rendimiento de diversos clasificadores paramétricos y no paramétricos a partir de las biometrías vocal y visual. En sentido decreciente por lo que hace a su eficiencia, en este trabajo se evalúan *Logistic Regression, Maximum a Posteriori, k-Nearest Neighbors classifiers* (Higgins et al., 1993), *Multilayer Perceptrons, Binary Decision Trees, Maximum Likelihood, Quadratic classifiers y Linear classifiers*.

En (Roli et al., 2002b) se comparan experimentalmente se comparan varias reglas de fusión entrenadas y no entrenadas, en una fusión no balanceada.

Como ejemplos de la utilización de la teoría de Bayes en la fusión a nivel de puntuaciones de reconocimiento está el trabajo de Kittler et al. que en (Kittler et al., 1997a; Kittler et al., 1997b) analizan el problema de fusión en el marco de la teoría bayesiana para la verificación de imágenes

faciales, y en (Kittler et al., 1998) realizan una comparación experimental de diversos esquemas de clasificación. En (Duc et al., 1997) se utiliza la teoría de Bayes para estimar las desviaciones de las opiniones de diversos expertos y calibrar y conciliar sus opiniones para tomar una decisión conjunta.

También Bigun et al. comparan en (Bigun et al., 1997a; Bigun et al., 1997b) el promediado de los resultados de reconocimiento con un supervisor entrenado basado en estadística Bayesiana, que es el método que obtiene los mejores resultados. Los sistemas unimodales involucrados en la fusión consisten en algoritmos de autentificación entrenados a partir de parámetros de Gabor para caras y parámetros LPC para voz.

En (Ross et al., 2001) se utilizan árboles de decisión y análisis mediante discriminante lineal para transformar los vectores tridimensionales provenientes de expertos en caras, huellas dactilares y geometría de la mano en un nuevo subespacio que maximice la separación entre clases.

En lo que respecta a la utilización de redes neuronales en la fusión de puntuaciones de reconocimiento, en (Czyz et al., 2003) se compara la fusión de información procedente de caras y voz mediante un *MLP* y mediante la combinación lineal de los resultados de reconocimiento. En (Baker et al., 2002) se presentan las redes neuronales artificiales (*ANN*) como un método de fusión de biometrías y se aplican a la fusión de información de caras, voz y huellas dactilares.

Otro ejemplo es el *ICSI's 2005 Speaker Recognition System* (Mirghafori et al., 2005), en que los resultados de cuatro expertos se combinan mediante una red neuronal. Los cuatro subsistemas se componen de: un sistema de palabra clave condicional basado en *HMM*, un sistema de n-gramas basado en *SVM*, un sistema secuencial no paramétrico y un sistema tradicional basado en *GMM* con parámetros cepstrales. Los tres primeros sistemas han sido diseñados para aprovechar información de alto nivel y alta periodicidad mientras que el cuarto aporta información de bajo nivel.

Por último, las *SVM*s han demostrado obtener buenos resultados en un gran número de aplicaciones en el entorno de la fusión de biometrías a nivel de puntuaciones y el número de trabajos presentados en esta área son considerables. En (Ben-Yacoub, 1999) se muestra que la decisión final en la verificación de un sistema biométrico es una clasificación binaria y se propone resolverlo con una *SVM*. Esta aproximación obtiene mejoras en comparación con otros métodos propuestos.

En (Campbell et al., 2003a; Campbell et al., 2003b; Campbell et al., 2004) se utilizan *Support Vector Machines* para la tarea de verificación de locutor a partir de parámetros acústicos de alto nivel comparándolo con las técnicas tradicionales de frecuencia y concurrencia de símbolos.

En (Fierrez et al., 2003a; Fierrez et al., 2003b) se utiliza una *SVM* para el reconocimiento conjunto mediante huellas dactilares y firmas. En (Fierrez et al., 2004) y (Garcia et al., 2004) se ajusta el valor de control de la influencia de los vectores de parámetros mal clasificados en función de la calidad de las señales, de manera que se les asigna una menor relevancia a las señales de menor

calidad. Esta adaptación permite mejorar los resultados de reconocimiento respecto a una *SVM* clásica.

Por otro lado, en (Gurban et al., 2005) se apuesta por un sistema híbrido de *SVM* y *HMM* en la tarea de reconocimiento de locutor basado en texto independiente. En este trabajo, se estudia la fusión tanto a nivel de parámetros como a nivel de puntuaciones de reconocimiento para sistemas de voz y de caras. Los resultados prueban que la combinación es beneficiosa tanto en términos de resultados como práctica en términos computacionales.

Las técnicas de fusión presentadas hasta ahora se pueden utilizar también para fusionar resultados de reconocimiento obtenidos por sistemas independientes para una misma biometría. Así, en (Cheung et al., 2005), en un sistema audio-visual, los valores obtenidos por diferentes sistemas para la misma modalidad se combinan linealmente de manera dependiente del usuario y se realiza una segunda fusión multimodal mediante la utilización de *Support Vector Machines (SVM)*.

También Czyz et al. utilizan fusión en el diseño de sus sistemas faciales de reconocimiento. En (Czyz et al., 2002) se combinan expertos reconocedores de caras basados en LDA y técnicas probabilísticas, que proporcionan información para realizar la fusión. En (Czyz et al., 2004a) y (Czyz et al., 2004b), se realiza una doble fusión en el reconocimiento de caras. En primer lugar, se fusionan los resultados obtenidos en tramas de vídeo sucesivas y, en segundo lugar, los resultados obtenidos por diversos algoritmos de autentificación de caras se combinan para obtener una decisión final.

3.6 Fusión a nivel de decisión

En la fusión a nivel de decisión cada sistema obtiene los parámetros necesarios para compararlos con el modelo de cada persona y toma su propia decisión de reconocimiento. El sistema de fusión debe tomar una decisión final a partir de las decisiones de cada uno de los sistemas.

Las técnicas más directas en el caso de un sistema de verificación son la aplicación de operaciones lógicas como *OR* y *AND* sobre las decisiones individuales (Bolle et al., 2004).

Por otro lado, una de las técnicas más utilizadas para realizar la fusión a este nivel es el votador por mayoría (Zuev et al., 1996) en que se escoge la decisión respaldada por el mayor número de sistemas unimodales.

Para aplicar esta técnica en la verificación de personas es necesario que en la decisión estén involucrados al menos tres sistemas unimodales y que el número de sistemas sea impar. Si no fuera así se podrían producir empates en la decisión. Si el sistema es de identificación, un votador de

mayoría puede llegar a precisar una gran cantidad de expertos para que existan coincidencias en las identificaciones.

Basándose en el votador por mayoría, se han desarrollado sistemas en que las decisiones son ponderadas en función de la fiabilidad de cada biometría. En este principio se basa la teoría de la evidencia de Dempster-Shafer que amplía el concepto bayesiano de la probabilidad introduciendo el concepto de credibilidad o fiabilidad.

En (Wang et al., 2004) se comparan 13 combinaciones de métodos para el reconocimiento conjunto de voz y huella dactilar tanto en verificación como en identificación. Los resultados experimentales demuestran que las *SVM*s y el método Dempster-Shafer son superiores a otros esquemas de reconocimiento. En (Tao et al., 2009) se propone un esquema óptimo de fusión a nivel de decisión mediante las reglas AND o OR, basado en la optimización de los umbrales de decisión.

En (Chatzis et al., 1999) se utilizan algoritmos de clustering para la fusión a nivel de decisión utilizando la información de diferentes modalidades biométricas.

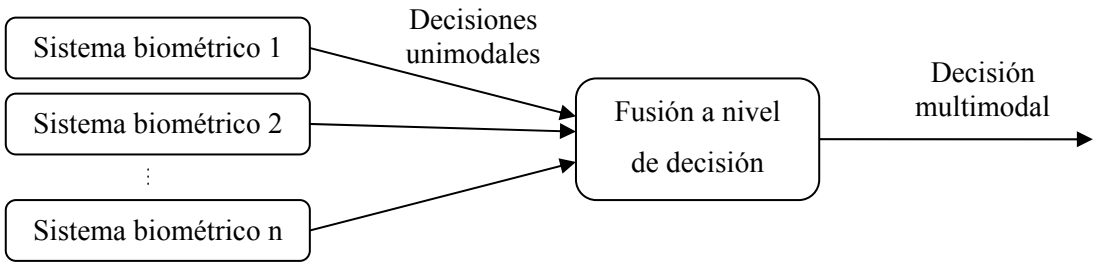


Figura 3-7: Esquema de fusión a nivel de decisión.

3.7 Evaluación de las técnicas multimodales

Para evaluar los sistemas multimodales y comparar así las técnicas de normalización y fusión utilizadas se utilizan las mismas medidas que en el caso del reconocimiento unimodal.

Como ya se ha comentado en el apartado 2.4.4, algunas de las medidas más utilizadas en los sistemas de verificación son la tasa de error equivalente (*EER: Equal Error Rate*), y la mínima tasa de error total (*HTER: Half Total Error Rate*). En ocasiones, también se evalúan los sistemas fijando la tasa de falsas aceptación (*FAR: False acceptance rate*) o tasa de falso rechazo (FRR: False rejection *rate*) y comparando los valores de la otra tasa de error para el umbral establecido.

De la misma forma, para evaluar la relación entre las dos tasas de error se utilizan el diagrama FAR-FRR, la curva ROC y la curva DET. En la figura 3-8 se muestra una curva DET comparando

los sistemas unimodales de reconocimiento de locutor y de reconocimiento de caras y un sistema multimodal basado en normalización *min-max* con fusión *Simple Sum*, a partir de las puntuaciones de caras y espectro de voz del apartado 6-2.

En los sistemas de identificación, se utiliza como medida para la verificación del sistema la tasa de error de identificación (*FIR: False Identification Rate*).

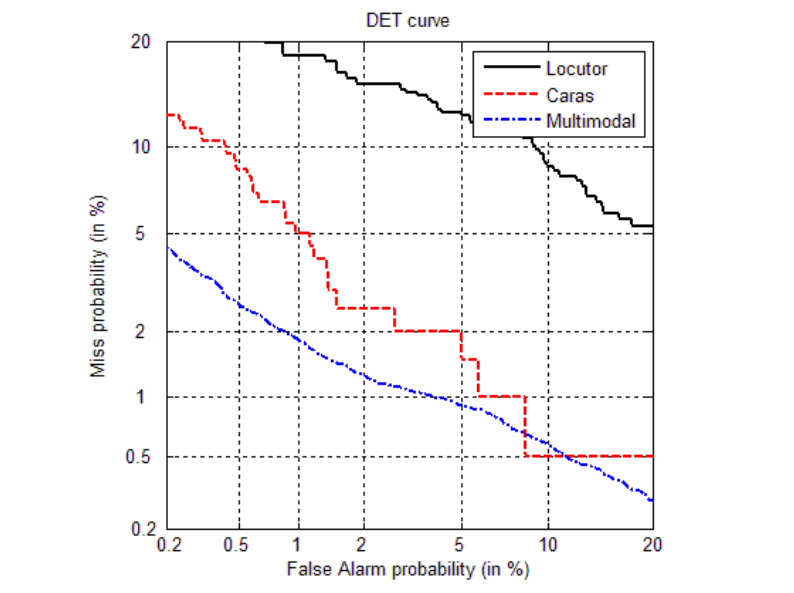


Figura 3-8: Curva DET para sistemas unimodales y multimodal.

3.7.1 Bases de datos multimodales y quiméricas

Para obtener estos indicadores y gráficas es necesario aplicar las técnicas de fusión sobre un conjunto de datos común. Estos datos se obtienen a partir de bases de datos, que pueden ser bases de datos multimodales, en que varias modalidades biométricas de diversos individuos se obtienen y se registran en una o varias sesiones. Diversas universidades o consorcios han grabado bases de datos multimodales. Aquí se presentan las más relevantes:

BANCA: Base de datos multilenguaje que contiene caras y voz en inglés, francés, italiano y español. La base de datos consta de 208 sujetos, 52 por cada idioma. Creador: Varios. Distribuidor: *University of Surrey* (Bailly-Baillière et al., 2003).

XM2VTS: Base de datos multimodal europea para voz y caras. Consta de grabaciones de vídeo conteniendo la pronunciación de frases en inglés y de imágenes fijas de caras de 295 sujetos. Creador: *University of Surrey*. Distribuidor: *University of Surrey* (Lüttin et al., 1998).

BIOMET: Contiene voz y caras grabadas con videocámara, caras grabadas con infrarojos, hullas de manos, huellas dactilares y firmas. Grabada en Francia. Hay 236 usuarios comunes a estas biometrías aunque algunas de ellas llegan a los 327. Distribuidor: ELDA (*European Language Resources Distribution Agency*) (Carcia-Salicetti et al., 2003).

MCYT: Incluye huellas dactilares y firmas de 330 usuarios. Grabada es España. Creador: ATVS-UPM. Distribuidor: ATVS (*Biometric Research Lab*) (Ortega-Garcia et al., 2003).

BIOSEC: Consiste en imágenes de huellas dactilares adquiridas con tres sensores diferentes, imágenes faciales frontales obtenidas con una webcam, imágenes de iris obtenidas con un sensor de iris y segmentos de voz obtenidos con un micrófono cercano y una webcam. Incluye datos multimodales de 2 sesiones para 200 individuos (Fierrez et al., 2007).

Otra forma de obtener una base de datos multimodal es combinar diversas bases de datos unimodales para crear lo que se denomina una base de datos quimérica. Para ello se combinan las pruebas realizadas a los usuarios de las diferentes bases de datos creando de esta manera nuevos usuarios multimodales (Poh et al., 2005a).

En esta tesis se han utilizados las dos opciones. En los experimentos del apartado 6 hemos creado bases de datos quiméricas a partir de los resultados de reconocimiento de caras de la base de datos XM2VTS y de voz de las bases de datos POLYCOST (Melin et al., 1996; Petrovska, 1998) y Switchboard-I (Campbell et al., 1999; Godfrey et al. 1990), mientras que para los resultados del proyecto Agatha, en el apartato 7, se ha utilizado la base de datos XM2VTS como base de datos multimodal.

Para realizar todo el proceso de reconocimiento, el conjunto de datos seleccionado se suele dividir en tres subconjuntos que se denominan de entrenamiento, de desarrollo y de evaluación o test.

En los sistemas de fusión a nivel de parámetros, estos subconjuntos se utilizan de forma similar al caso de los sistemas unimodales, es decir, el primer subconjunto se utiliza para entrenar los modelos de los usuarios, en el caso de la fusión, modelos multimodales, los datos de desarrollo se utilizan para establecer los umbrales de los sistemas de verificación o identificación y, finalmente, los datos de test se utilizan para obtener las medidas de reconocimiento.

En el caso de los sistemas de fusión a nivel de puntuaciones o decisión, el primer subconjunto se utiliza para el entrenamiento de los modelos unimodales de las diversas modalidades biométricas involucradas en el sistema multimodal, los datos de desarrollo se utilizan para el entrenamiento de los sistemas de normalización y fusión multimodales y, del mismo modo que en el caso anterior, el subconjunto de test se utiliza para obtener los resultados definitivos de reconocimiento.

3.7.2 Comparativa de sistemas de reconocimiento

Decidir que un sistema de reconocimiento es más fiable que otro u otros a partir de uno de los indicadores presentados, cuando se aplican sobre un conjunto de datos particular, puede resultar poco fiable estadísticamente.

Una de las técnicas más utilizadas para comparar estadísticamente los resultados es el cálculo de los intervalos de confianza. Para ello, se dividen los datos de test en *N* subgrupos y se calcula la varianza de los resultados obtenidos con dichos subgrupos. Aproximando la distribución de los resultados a una distribución gaussiana se pueden calcular los intervalos en que se encontrarán los resultados con una cierta probabilidad.

Otra forma de comparar dos sistemas de reconocimiento es considerar la diferencia de las dos tasas de reconocimiento como una variable aleatoria. Si se aproximan las dos distribuciones binomiales como distribuciones normales y se consideran las tasas de error como independientes, entonces esta diferencia también tendrá aproximadamente una distribución normal.

Sin embargo, esta comparativa presupone que los experimentos de test se realizan de forma independiente, cosa que en nuestro caso no es cierta, dado que se utilizan los mismos datos para probar las dos técnicas a comparar (Kuncheva, 2004). Dietterich demuestra en (Dietterich, 1998) que corrigiendo esta dependencia de los datos, se obtiene una estadística que es la raíz cuadrada de

$$x^2 = \frac{\left(|N_{01} - N_{10}| - 1\right)^2}{N_{01} + N_{10}} \tag{3.6}$$

donde N_{01} es el número de test en que la primera técnica a comparar falla y la otra acierta y N_{10} es el número de test en que la primera acierta y la otra falla.

Este cálculo se corresponde con el test de McNemar y se considera que sigue aproximadamente una distribución χ^2 con un grado de libertad. La forma más simple de llevar a cabo este test es calcular x^2 y compararlo con las tablas de la distribución χ^2. Para una fiabilidad del 95% x^2 debe ser mayor de 3.841. Dietterich recomienda la utilización de dicho test frente a la diferencia de las dos tasas de error.

4 Normalización de media y varianza basada en las estadísticas separadas de clientes e impostores

En este apartado, se van a presentar diversas técnicas para la normalización de las puntuaciones de clientes e impostores en sistemas de verificación de personas. En concreto, las técnicas presentadas normalizan la media y la varianza de las puntuaciones teniendo en cuenta las estadísticas separadas de clientes e impostores.

4.1 Normalización conjunta de medias

Tal y como se ha comentado en el apartado anterior, cualquier transformación monótonamente creciente aplicada sobre las puntuaciones obtenidas a partir de un sistema de reconocimiento de personas no afecta al rendimiento del sistema. Por lo tanto, se puede añadir un número constante real a un conjunto de puntuaciones biométricas o se pueden multiplicar por una constante real positiva sin modificar los resultados de reconocimiento.

Se asume que en la fase de entrenamiento se conocen las identidades supuestas y reales. Se definen a_C y a_I respectivamente como las puntuaciones de clientes e impostores de una modalidad biométrica unimodal y x_C y x_I como las puntuaciones normalizadas calculadas como

$$x_C = k_1 \cdot a_C + k_2 \tag{4.1}$$

$$x_I = k_1 \cdot a_I + k_2 \tag{4.2}$$

donde k_1 es una constante real positiva y k_2 es una constante real. Las puntuaciones x_C y x_I proporcionarán los mismos resultados de verificación que las puntuaciones originales a_C y a_I.

En el proceso de normalización, pocos autores han tenido en cuenta las estadísticas separadas de las puntuaciones de clientes e impostores (Poh et al., 2005b; Poh et al., 2008). En esta técnica, se propone la utilización de esta información para la normalización de las medias de los dos conjuntos de puntuaciones. De esta manera, el valor absoluto de las medias de x_C y x_I se fija a un valor predefinido $\mu_x=\mu_{xC}=-\mu_{xI}$ ajustando el valor de las constantes k_1 y k_2. En consecuencia, la suma de las medias de las puntuaciones de clientes e impostores será cero. Si los valores de μ_{xC} y μ_{xI} se fijan a uno y menos uno respectivamente, los valores de k_1 y k_2 obtenidos son:

$$k_1 = \frac{2}{\mu_{aC} - \mu_{aI}} \tag{4.3}$$

$$k_2 = -\frac{\mu_{aC} + \mu_{aI}}{\mu_{aC} - \mu_{aI}} \tag{4.4}$$

donde μ_{aC} y μ_{aI} son las medias de las puntaciones de clientes e impostores, a_C y a_I respectivamente. Definimos entonces las puntuaciones x_C y x_I como puntuaciones con las medias normalizadas de forma conjunta (*JMN: Joint Mean Normalization*).

Si esta normalización se aplica a dos conjuntos de puntuaciones biométricas a y b, se obtienen dos conjuntos de puntuaciones con las medias normalizadas de forma conjunta, x e y, es decir,

$$x = k_{1a}a + k_{2a} \tag{4.5}$$

$$y = k_{1b}b + k_{2b} \tag{4.6}$$

donde k_{1a}, k_{2a}, k_{1b} y k_{2b} se calculan por medio de las ecuaciones (4.3) y (4.4) para cada una de las modalidades biométricas unimodales.

Una vez que las puntuaciones de las dos modalidades biométricas han sido normalizadas, se deben combinar para obtener una única puntuación que permita tomar una decisión de reconocimiento. Uno de los métodos de fusión más directos es la suma simple (*SS: Simple Sum*), que consiste en la suma de las puntuaciones unimodales. Sin embargo, para mantener inalteradas las medias de puntuaciones de clientes e impostores respecto a los sistemas unimodales normalizados, vamos a aplicar una semisuma, es decir, se define la puntuación multimodal u como

$$u = \frac{1}{2}(x + y) \tag{4.7}$$

En este caso, las medias de las puntuaciones de clientes e impostores tanto de los sistemas unimodales como del sistema multimodal son los mismos

$$\mu_{xC} = -\mu_{xI} = \mu_{yC} = -\mu_{yI} = \mu_{uC} = -\mu_{uI} = \mu \tag{4.8}$$

donde μ_{xC}, μ_{yC}, μ_{uC} y μ_{xI}, μ_{yI}, μ_{uI} son las medias de las puntaciones de clientes e impostores x_C, y_C, u_C y x_I, y_I, u_I respectivamente. Si en la ecuación (4.7) se sustituyen los valores de x e y por las

expresiones en (4.5) y (4.6) se obtienen las puntuaciones JMN-HS (*Joint Mean Normalization – Half Sum*):

$$u_{jmn-hs} = \frac{1}{2}\left(\frac{2}{\mu_{aC} - \mu_{aI}} a - \frac{\mu_{aC} + \mu_{aI}}{\mu_{aC} - \mu_{aI}} + \frac{2}{\mu_{bC} - \mu_{bI}} b - \frac{\mu_{bC} + \mu_{bI}}{\mu_{bC} - \mu_{bI}} \right) \tag{4.9}$$

donde μ_{bC} y μ_{bI} son respectivamente las medias de las puntuaciones de clientes e impostores de la biometría *b*, de manera análoga a como se han definido anteriormente para la otra modalidad biométrica.

Tal y como se ha explicado anteriormente, la suma de una constante real o la multiplicación por una constante positiva real aplicadas sobre las puntuaciones biométricas no afectan el rendimiento del sistema. Para simplificar el cálculo de las puntuaciones del sistema de fusión, esta propiedad se puede aplicar a las puntuaciones JMN-HS para obtener un nuevo conjunto de puntuaciones multimodales que se pueden expresar como:

$$v_{jmn-hs} = \frac{a}{\mu_{aC} - \mu_{aI}} + \frac{b}{\mu_{bC} - \mu_{bI}} \tag{4.10}$$

Este resultado permite calcular las puntuaciones biométricas multimodales a partir de las puntuaciones de clientes e impostores y las medias estadísticas de las puntuaciones *a* y *b*.

A partir de este punto, se describe el efecto de la combinación de la normalización JMN y la fusión HS sobre la estadística de la biometría multimodal. Las distribuciones de las puntuaciones de clientes e impostores se solapan y este solapamiento produce los errores de verificación. Se puede suponer que, si las varianzas de las puntuaciones de clientes e impostores se reducen, habrá un menor solapamiento y, en consecuencia, el sistema mejorará su rendimiento. Por este motivo, resulta de interés estudiar los casos en que las varianzas de las puntuaciones multimodales se reducen respecto a las respectivas varianzas unimodales. Este estudio sólo tiene sentido cuando las medias de las puntuaciones de clientes e impostores son fijas, como en el caso de JMN.

La varianza de las puntuaciones multimodales de clientes tras aplicar JMN y la semisuma se puede calcular como (Leon-Garcia, 1993):

$$\sigma_{uC}^2 = \mathrm{E}\left[(u_C - \mu_{uC})^2 \right] = \mathrm{E}\left[\left(\frac{1}{2}\left(x_C - \mu_{xC} + y_C - \mu_{yC} \right) \right)^2 \right] \tag{4.11}$$

$$\sigma_{uC}^2 = \frac{1}{4}\mathrm{E}\left[(x_C - \mu_{xC})^2 + (y_C - \mu_{yC})^2 + 2(x_C - \mu_{xC})(y_C - \mu_{yC}) \right] \tag{4.12}$$

La varianza de las puntuaciones de impostores se calcularía de forma análoga.

Finalmente, si se considera que x_G y y_G, y x_I y y_I están incorrelados, las varianzas de las puntuaciones de clientes e impostores son

$$\sigma_{uC}^2 = \frac{1}{4}\left(\sigma_{xC}^2 + \sigma_{yC}^2\right) \tag{4.13}$$

$$\sigma_{uI}^2 = \frac{1}{4}\left(\sigma_{xI}^2 + \sigma_{yI}^2\right) \tag{4.14}$$

donde σ_{xC}, σ_{yC}, σ_{xI} y σ_{yI} son las desviaciones estándar de las puntuaciones de clientes e impostores de los sistemas unimodales con las medias normalizadas de forma conjunta.

Dado que el objetivo de la fusión es una reducción de las varianzas multimodales, vamos a buscar la relación entre las varianzas unimodales que produce una reducción de las varianzas multimodales respecto a las unimodales. Para que esto suceda, para las varianzas de clientes, se deben cumplir de forma simultánea las siguientes ecuaciones

$$\sigma_{uC}^2 = \frac{1}{4}\left(\sigma_{xC}^2 + \sigma_{yC}^2\right) < \sigma_{xC}^2 \tag{4.15}$$

$$\sigma_{uC}^2 = \frac{1}{4}\left(\sigma_{xC}^2 + \sigma_{yC}^2\right) < \sigma_{yC}^2 \tag{4.16}$$

La varianza de las puntuaciones de clientes de *y* se puede expresar en función de la varianza de las puntuaciones de clientes de *x* de la siguiente manera:

$$\sigma_{yC}^2 = k \cdot \sigma_{xC}^2 \tag{4.17}$$

y las inecuaciones anteriores se pueden resolver aplicando (4.17) en (4.15) y (4.16), es decir,

$$\frac{1}{4}\left(\sigma_{xC}^2 + k\sigma_{xC}^2\right) < \sigma_{xC}^2 \tag{4.18}$$

$$\frac{1}{4}\left(\sigma_{xC}^2 + k\sigma_{xC}^2\right) < k\sigma_{xC}^2 \tag{4.19}$$

Finalmente, se obtienen las limitaciones $\sigma_{yC}^2 < 3\sigma_{xC}^2$ y $\sigma_{xC}^2 < 3\sigma_{yC}^2$. En consecuencia, la varianza de las puntuaciones multimodales de clientes se reduce respecto a las varianzas las puntuaciones de clientes unimodales cuando la varianza de clientes de cada modalidad es menor que tres veces la varianza de clientes de la otra modalidad. Si se aplica el mismo resultado para las varianzas de las puntuaciones de impostores, las varianzas separadas se reducen cuando

$$\frac{1}{3}\sigma_{yC}^2 < \sigma_{xC}^2 < 3\sigma_{yC}^2 \tag{4.20}$$

$$\frac{1}{3}\sigma_{yI}^2 < \sigma_{xI}^2 < 3\sigma_{yI}^2 \tag{4.21}$$

Obviamente, la reducción de las varianzas unimodales implica una reducción de las varianzas multimodales. Sin embargo, para unos valores determinados de las varianzas unimodales, los

menores valor de σ_{uG} y σ_{uI} se obtienen respectivamente cuando $\sigma_{xG} = \sigma_{yG}$ y $\sigma_{xI} = \sigma_{yI}$. En este caso, las desviaciones estándar se reducen con un factor de $\sqrt{2}$.

En resumen, si se aplican la normalización JMN y la fusión HS sobre puntuaciones biométricas incorreladas con varianzas similares, las varianzas de las puntuaciones de clientes e impostores se verán reducidas y los resultados de verificación serán muy probablemente mejorados respecto a los obtenidos por los sistemas biométricos unimodales.

Para ilustrar la necesidad del proceso de normalización y, en concreto, la normalización conjunta de medias, se va a utilizar como ejemplo la fusión de las puntuaciones obtenidas por dos sistemas unimodales, un sistema de verificación de locutor y un sistema de verificación facial. La figura 4-1 muestra los histogramas de las puntuaciones de clientes e impostores para el sistema de reconocimiento de locutor (a) y para el sistema de reconocimiento facial (b).

El rango de puntuaciones en el sistema de locutor va de -1583700.844 a 387998.0625 mientras que para el sistema de reconocimiento facial, el rango de puntuaciones va de 0.229006 a 0.873828. En la figura (c) se puede observar el efecto de esta gran diferencia en los rangos de puntuaciones de las dos modalidades en un histograma conjunto.

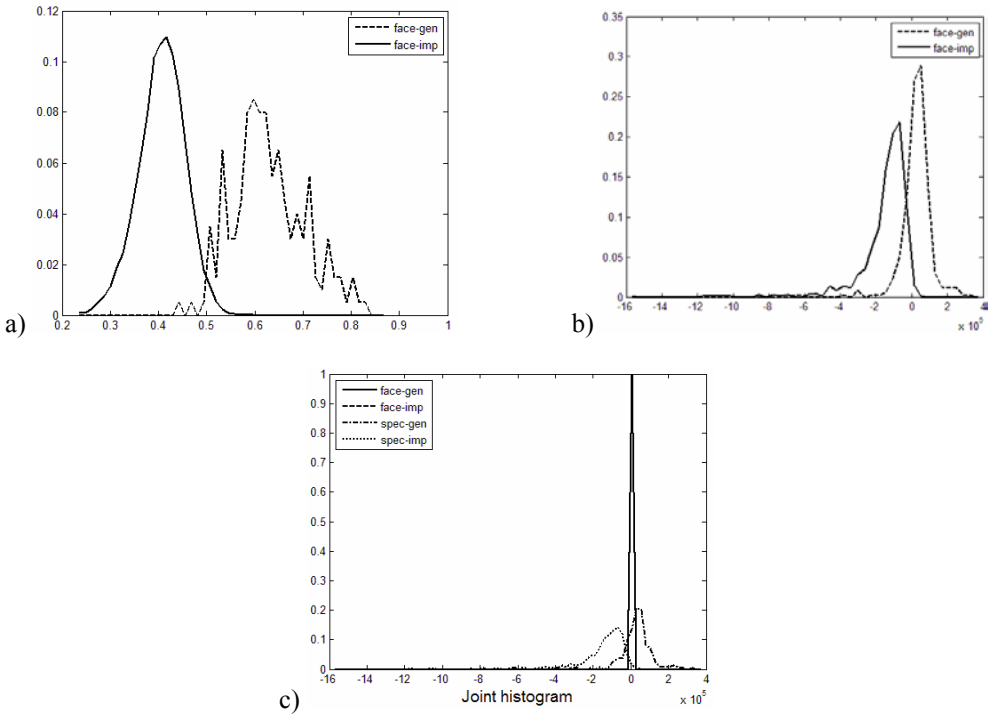


Figura 4-1: Histogramas de las puntuaciones para los sistemas de reconocimiento de locutor (a) y de reconocimiento facial (b) e histograma conjunto (c).

Si, por ejemplo, se utiliza la suma simple como método de fusión sin homogeneizar estas puntuaciones, las puntuaciones del sistema facial quedarán absorbidas por las puntuaciones del sistema de reconocimiento de locutor por lo que el sistema multimodal sería equivalente al sistema de reconocimiento de locutor.

Sin embargo, si para homogeneizar las puntuaciones se utiliza la normalización conjunta de medias, el histograma conjunto de las puntuaciones de las dos modalidades es el mostrado en la figura 4-2.

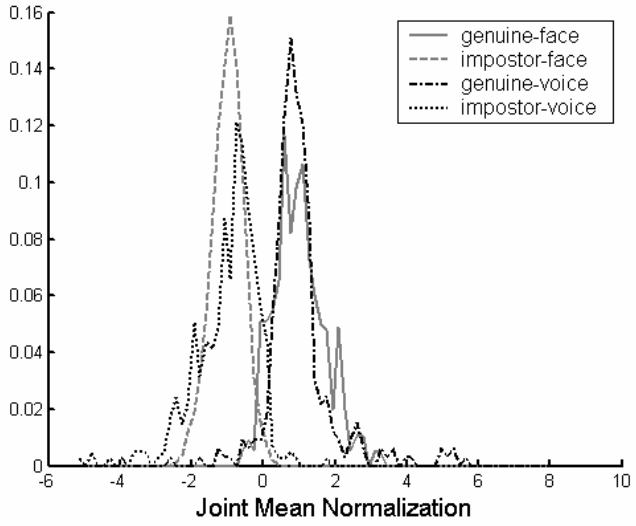


Figura 4-2: Histograma conjunto de las puntuaciones normalizadas.

Al aplicar, por ejemplo, la ponderación en función del resultado de reconocimiento de cada biometría unimodal (*MW: matcher weighting*) sobre las puntuaciones normalizadas, se consigue reducir el EER de los sistemas de voz y caras, de un 9.52% y un 2.5% respectivamente, al 1.5% del sistema multimodal.

En la figura 4-3 se puede observar el efecto de diversas normalizaciones del estado del arte y de la normalización conjunta de medias sobre el histograma de las puntaciones de las modalidades biométricas basadas en la cara y en el espectro de voz del apartado 6.2.

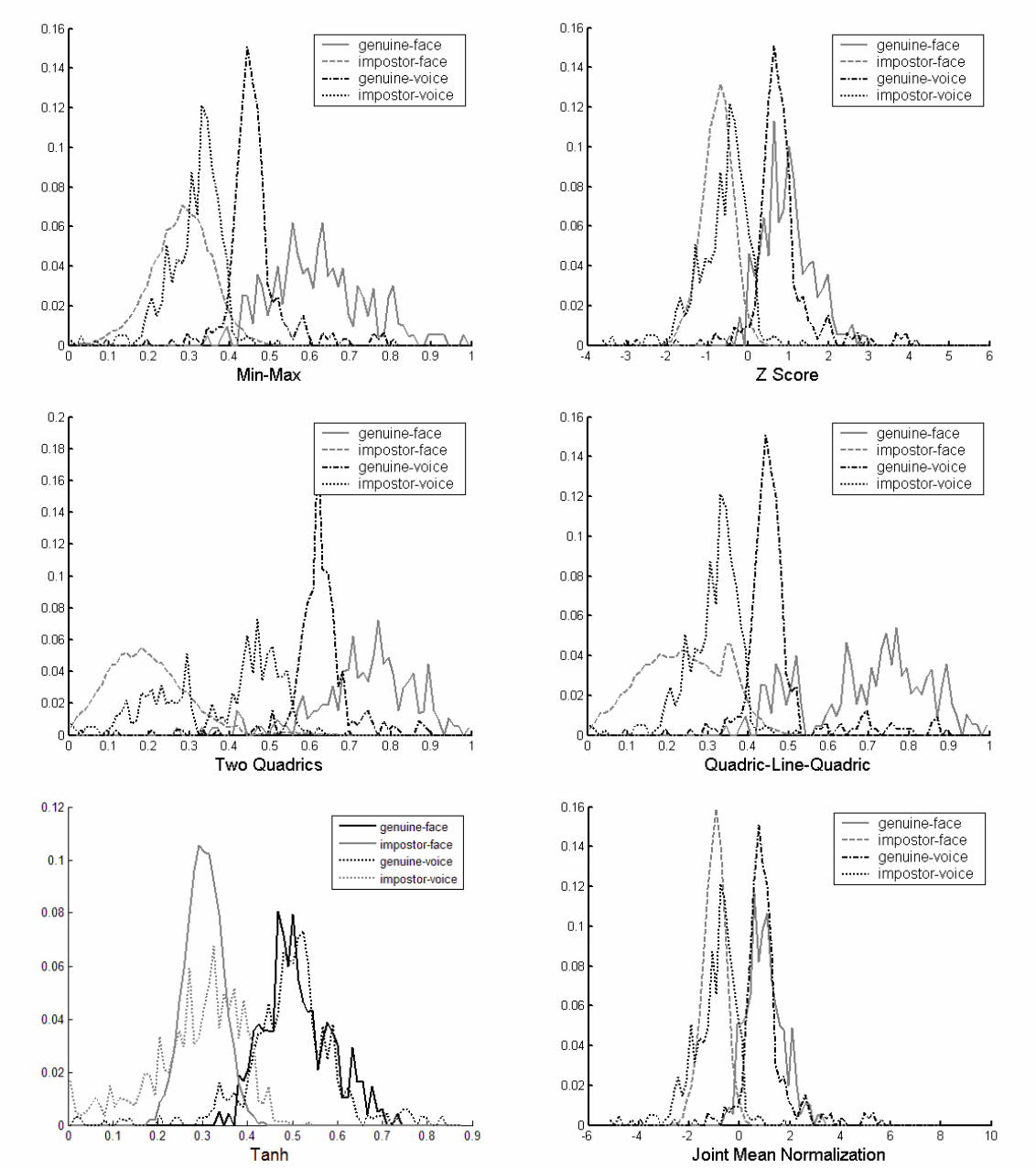


Figura 4-3: Distribución de las puntaciones de modalidades biométricas basadas en la cara y en la voz, normalizadas mediante diversas técnicas de normalización.

4.2 Técnicas de minimización de las varianzas

En la normalización conjunta de medias, se utiliza la suma simple como método de fusión para obtener la puntuación multimodal a partir de las puntuaciones unimodales. En este apartado, se van a presentar nuevas técnicas de normalización que unidas a la suma simple minimicen la varianza de las puntuaciones de clientes e impostores. La propuesta consiste en una combinación lineal de las

puntuaciones multimodales cuyos pesos se escogen de manera que se minimicen la varianzas separadas del valor resultante, es decir,

$$u = \alpha \cdot x + (1-\alpha) \cdot y \tag{4.22}$$

donde α es un número real positivo. Los pesos α y $(1-\alpha)$ garantizan que, si x e y tienen la misma media para las puntuaciones de clientes e impostores, como en el caso de las puntuaciones con normalización conjunta de medias, u tendrá las mismas medias que x e y.

$$u_C = \alpha \cdot x_C + (1-\alpha) \cdot y_C \qquad \mu_{uC} = \alpha \cdot \mu_{xC} + (1-\alpha) \cdot \mu_{yC} = \mu_C \tag{4.23) (4.24}$$

$$u_I = \alpha \cdot x_I + (1-\alpha) \cdot y_I \qquad \mu_{uI} = \alpha \cdot \mu_{xI} + (1-\alpha) \cdot \mu_{yI} = \mu_I \tag{4.25) (4.26}$$

Suponiendo que x_C e y_C, y x_I e y_I están incorrelados, las varianzas de u pueden ser calculadas como:

$$\sigma_{uC}^2 = \alpha^2 \sigma_{xC}^2 + (1-\alpha)^2 \sigma_{yC}^2 \qquad \sigma_{uI}^2 = \alpha^2 \sigma_{xI}^2 + (1-\alpha)^2 \sigma_{yI}^2 \tag{4.27) (4.28}$$

Desafortunadamente, en la mayor parte de los casos no existirá ningún valor de α que minimice las dos ecuaciones simultáneamente. Por lo tanto, para encontrar un único valor para α, se debe definir un nuevo criterio de minimización. En esta tesis, se propone minimizar la suma de las desviaciones estándar de las puntuaciones de clientes e impostores o la suma de las correspondientes varianzas, en lugar de minimizar las varianzas de forma separada. En los siguientes apartados se presentan estas dos opciones.

4.2.1 Minimización de la suma de las desviaciones estándar

El objetivo de la normalización para la minimización de la suma de las desviaciones estándar (*MSDSW : Minimum standard deviation sum weighting*) es encontrar un valor de α en (4.22) que minimice la suma de las desviaciones estándar de clientes e impostores $\sigma_{uC} + \sigma_{uI}$. Si se deriva esta suma con respecto a α y se iguala el resultado a cero, se llega a la siguiente expresión para obtener la variable α:

$$\frac{\alpha^2 \sigma_{xC}^2 + (1-\alpha)\sigma_{yC}^2}{\alpha^2 \sigma_{xI}^2 + (1-\alpha)\sigma_{yI}^2} = -\sqrt{\frac{\alpha^2 \sigma_{xC}^2 + (1-\alpha)^2 \sigma_{yC}^2}{\alpha^2 \sigma_{xI}^2 + (1-\alpha)^2 \sigma_{yI}^2}} \tag{4.29}$$

El valor de α que minimiza la suma de las desviaciones estándar cumple esta ecuación y, dado que no se puede obtener una expresión cerrada para encontrar este valor, la ecuación se debe resolver mediante la utilización de métodos iterativos.

4.2.2 Minimización de la suma de las varianzas

La suma de las varianzas $\sigma^2_{uC} + \sigma^2_{uI}$ se minimiza cuando su derivada respecto a α es igual a cero, es decir,

$$\alpha\sigma^2_{xC} - (1-\alpha)\sigma^2_{yC} + \alpha\sigma^2_{xI} - (1-\alpha)\sigma^2_{yI} = 0 \qquad (4.30)$$

Los valores de α y $1\text{-}\alpha$ que minimizan la suma de las varianzas son, en consecuencia,

$$\alpha = \frac{\sigma^2_{yC} + \sigma^2_{yI}}{\sigma^2_{xC} + \sigma^2_{xI} + \sigma^2_{yC} + \sigma^2_{yI}} \qquad 1-\alpha = \frac{\sigma^2_{xC} + \sigma^2_{xI}}{\sigma^2_{xC} + \sigma^2_{xI} + \sigma^2_{yC} + \sigma^2_{yI}} \qquad (4.31)\ (4.32)$$

Mediante la aplicación de la minimización de la suma de las varianzas (*MVSW: Minimum variance sum weighting*), las puntuaciones de cada modalidad biométrica previamente normalizadas con JMN se multiplican por un factor proporcional a la suma de las varianzas de la otra biometría unimodal. Se puede esperar que la modalidad más precisa tenga una suma de sus varianzas sea menor y, en consecuencia, el mayor peso multiplicativo. Sin embargo, una gran diferencia entre la suma de las varianzas de las modalidades biométricas unimodales producen pesos con valor cercanos a cero y uno y, en consecuencia, el uso de una única biometría.

En este caso, la suma de las varianzas de la biometría multimodal es igual a la mitad de la media armónica de la suma de las varianzas de cada modalidad biométrica

$$\sigma^2_{uC} + \sigma^2_{uI} = \frac{1}{\frac{1}{\sigma^2_{xC} + \sigma^2_{xI}} + \frac{1}{\sigma^2_{yC} + \sigma^2_{yI}}} \qquad (4.33)$$

Si $\sigma^2_{yC} + \sigma^2_{yI}$ es cero entonces $\sigma^2_{uC} + \sigma^2_{uI} = \sigma^2_{xC} + \sigma^2_{xI}$. Cualquier otro valor de $\sigma^2_{yC} + \sigma^2_{yI}$ incrementará el valor del denominador de esta expresión y hará disminuir el valor de la suma de las varianzas multimodales, que será por lo tanto, en todos los casos, menor o igual que la suma de las varianzas de las puntuaciones de clientes e impostores de *x*. El mismo argumento es válido para las varianzas de la biometría *y*. Por lo tanto, dado que ninguna varianza es cero en los sistemas biométricos, esta relación garantiza que la suma de las varianzas de la biometría multimodal será menor que la suma de las varianzas de cualquiera de los sistemas unimodales.

$$\sigma^2_{uC} + \sigma^2_{uI} < \sigma^2_{xC} + \sigma^2_{xI} \qquad \sigma^2_{uC} + \sigma^2_{uI} < \sigma^2_{yC} + \sigma^2_{yI} \qquad (4.34)\ (4.35)$$

Dado que existe una reducción de la suma de las varianzas, se puede esperar que exista también una mejora respecto a los resultados obtenidos por los sistemas unimodales.

Si, de la misma forma que hemos hecho en el caso de la normalización conjunta de medias con semisuma, se simplifica la expresión correspondiente al sistema formado por MVSW y SS, se puede definir una biometría multimodal v_{mvsw} que proporciona los mismos resultados de

reconocimiento que la expresión en (4.22) sustituyendo (4.31) y (4.32). Las puntuaciones de esta biometría se puden expresar como

$$v_{mvsw} = \frac{\mu_{aC} - \mu_{aI}}{\sigma_{aC}^2 + \sigma_{aI}^2} a + \frac{\mu_{bC} - \mu_{bI}}{\sigma_{bC}^2 + \sigma_{bI}^2} b \tag{4.36}$$

Esta expresión únicamente depende de las puntuaciones y las estadísticas de primer y segundo orden de las modalidades biométricas unimodales.

Los pesos en la expresión (4.36) tienen ciertas semejanzas con el método *d-distance* utilizado en (Indovina et al., 2003) para la ponderación dependiente de usuario. Sin embargo, *d-distance* se ha probado como método de fusión en algunos de los conjuntos experimentales de esta tesis sin obtener resultados satisfactorios.

Estos resultados se pueden extrapolar a la fusión de más de dos biometrías. En ese caso, la puntuación multimodal se calcularía a partir de N puntuaciones unimodales como

$$u = \sum_{i=1}^{N} \alpha_i \cdot x_i \tag{4.37}$$

donde u es la biometría multimodal, x_i representa las N biometrías unimodales tras la normalización conjunta de medias y α_i son los pesos aplicados a cada biometría unimodal. En este caso, la suma de las varianzas de la biometría multimodal es igual a la media armónica de la suma de las varianzas de cada modalidad biométrica dividido por N, por lo que queda garantizado que la suma de las varianzas multimodales será menor que todas y cada una de las sumas de las varianzas unimodales

$$\sigma_{uC}^2 + \sigma_{uI}^2 = \frac{1}{\sum_{i=1}^{N} \frac{1}{\sigma_{x_iC}^2 + \sigma_{x_iI}^2}} \tag{4.38}$$

Esta minimización de la suma de las varianzas se consigue con

$$\alpha_i = \frac{1}{\sum_{j=1}^{N} \frac{\sigma_{x_iC}^2 + \sigma_{x_iI}^2}{\sigma_{x_jC}^2 + \sigma_{x_jI}^2}} = \frac{1}{\sigma_{x_iC}^2 + \sigma_{x_iI}^2} \cdot \frac{1}{\sum_{j=1}^{N} \frac{1}{\sigma_{x_jC}^2 + \sigma_{x_jI}^2}} \tag{4.39}$$

En esta ecuación el segundo multiplicando es constante para el cálculo de todos los pesos α_i, por lo que el factor multiplicador relevante para cada modalidad biométrica es la inversa de la suma de las propias varianzas de las puntuaciones de clientes e impostores. Por lo tanto, igual que en el caso bimodal, la biometría con menor suma de varianzas es la que tiene un mayor peso multiplicativo.

La expresión simplificada del sistema formado por MVSW y SS es en este caso

$$v_{mvsw} = \sum_{i=1}^{N} \frac{\mu_{a_i C} - \mu_{a_i I}}{\sigma^2_{a_i C} + \sigma^2_{a_i I}} a_i \tag{4.40}$$

donde a_i son las N biometrías unimodales y la definición de medias y varianzas es equivalente a la del resto del apartado. De nuevo, en esta expresión, el peso de cada modalidad biométrica multimodal depende únicamente de las propias puntuaciones y estadísticas de primer y segundo orden.

5 Normalización mediante ecualización de histograma

La ecualización de histograma (*HEQ: Histogram Equalization*) iguala la función de distribución acumulada de un cierto conjunto de datos a una distribución de referencia. Esta técnica se puede ver como una extensión de la normalización estadística realizada por la normalización *z-score* a toda la estadística de una modalidad biométrica y no sólo a su media y su varianza. Cuando las puntuaciones de todas las biometrías se ecualizan a la misma distribución de referencia, las características estadísticas de las puntuaciones ecualizadas se homogenizan y, en consecuencia el conjunto de puntuaciones se normaliza en sus propiedades estadísticas.

La ecualización de histogramas es un método ampliamente utilizado en el procesado y mejora de imágenes (Jain, 1986). HEQ realiza un mapeo monótono y no lineal que asigna los valores de la intensidad de los píxeles de la señal de entrada de manera que se controle la distribución del histograma de intensidades de la señal de salida con el fin de conseguir una distribución uniforme de intensidades o de realzar algunos niveles de intensidad.

Este método también se ha utilizado en métodos adaptativos de reconocimiento del habla o para la corrección de efectos no lineales típicamente introducidos por micrófonos, amplificadores, circuitos de control automático de la ganancia, etc. (Hilger et al., 2001; Balchandran et al., 1998). Además, también se ha utilizado en la verificación robusta de locutor para la adaptación (*warping*) de las tramas de coeficientes cepstrales sobre un intervalo específico (Pelenacos et al., 2001).

Finalmente, en la normalización de rango (*rank normalization*) de las puntuaciones propuesta en (Stolcke et al., 2005) para fusión biométrica multimodal, cada valor de las características es reemplazado por su rango en la distribución de refefencia, lo que es equivalente a la ecualización de histograma a una distribución uniforme.

El objetivo de la ecualización de histogramas es encontrar una transformación no lineal que reduzca las diferencias entre la distribución estadística de una señal y la distribución estadística de

una señal de referencia. La implementación de esta técnica que se ha llevado a cabo en esta tesis se basa en la utilización del histograma acumulado de las señales *F(.)*. Tanto en el histograma acumulado de la señal origen como en el de la señal de referencia, se definen *N* intervalos con la misma probabilidad de ocurrencia, es decir, con el mismo número de características (parámetros o puntuaciones) de entrenamiento en su interior. Cada intervalo en el histograma de la señal de referencia es representado por un valor de referencia (x_i) y por el valor máximo de la distribución acumulada ($F_m(x_i)$),

$$F_m(x_i) = \frac{\sum_{j=1}^{i} k_j}{M} = \frac{i}{N} \tag{5.1}$$

donde k_i es el número de características en el intervalo número *i* y *M* es el número total de puntuaciones.

En el proceso de ecualización, todas las características en cada intervalo del histograma de la señal de origen se asignan al correspondiente intervalo del histograma de referencia. $F(x_i)$ establece los límites de los intervalos en la distribución que se va a ecualizar. Todas las características en un conjunto a ecualizar con valores en su distribución acumulada entre $F(x_{i-1})$ y $F(x_i)$ se considerarán incluidas en el intervalo *i* del conjunto de características y, en consecuencia, se transformarán a su correspondiente valor x_i.

En esta tesis se ha tomado como valor de referencia de cada intervalo aquella característica con probabilidad acumulada igual al promedio de las probabilidades acumuladas de los puntos extremos del intervalo, es decir,

$$F(x_i) = \frac{F_m(x_i) + F_m(x_{i-1}) + \frac{1}{M}}{2} \tag{5.2}$$

5.1 Ecualización de histogramas para la normalización de parámetros y puntuaciones

En esta tesis, se utiliza la ecualización de histogramas desde dos aproximaciones diferentes. En la primera de ellas el histograma de una de las modalidades biométricas involucradas en el proceso de fusión se toma como el histograma de referencia en el proceso de ecualización. En la segunda aproximación, se utiliza una distribución definida a priori como referencia.

En el primer caso, una vez que se han ecualizado los parámetros o puntuaciones, las características de cada una de las modalidades biométricas tienen las mismas propiedades estadísticas que la escogida como la referencia para la ecualización.

La primera fase del proceso de ecualización es la división del histograma acumulado de las características biométricas a ecualizar en *M* intervalos con la misma probabilidad de ocurrencia (1/*M*). Una vez que los intervalos de origen se han definido, la distribución acumulada de referencia (*F(.)*) también se divide en el mismo número de intervalos, y un valor del intervalo se escoge como el punto representativo de cada intervalo. En el proceso de ecualización, todas las puntuaciones de la modalidad biométrica de origen incluidas en el mismo intervalo se relacionan con el punto representativo en el intervalo correspondiente de la distribución de referencia. La figura 5-1 ilustra el proceso de ecualización para puntuaciones biométricas. La puntuación *x* y todas las puntuaciones en el mismo intervalo se transforman a la puntuación ecualizada *y* de la distribución de referencia.

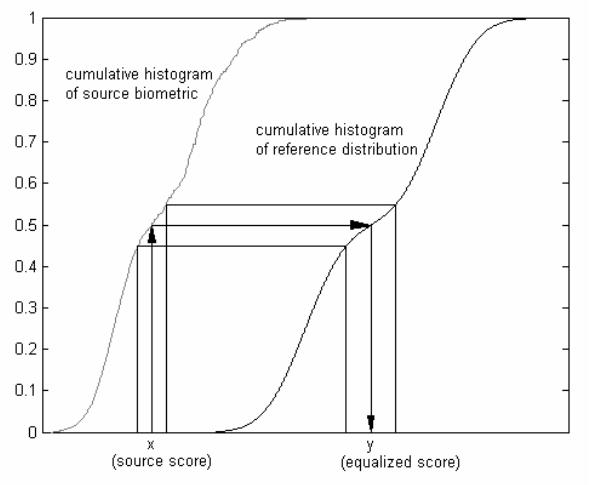


Figura 5-1: Transformación de la distribución acumulada realizada por HEQ.

Cuando se realiza el proceso de ecualización sobre un histograma de referencia calculado a partir de un conjunto de puntuaciones, el punto representativo de cada intervalo se escoge como aquel cuya probabilidad acumulada es el promedio de las probabilidades acumuladas de los puntos extremos del intervalo. La estimación de la probabilidad acumulada de una puntuación *y*, $\widetilde{F}(y)$, perteneciente al intervalo número *r* se puede calcular como el rango de *y* en una versión de las puntuaciones en que estas se ordenan de forma ascendente *rank(y)* dividido por el número total de puntuaciones *T*. La función *rank* para una puntuación en el intervalo *r* se puede formular en función del número de datos N_i en cada intervalo *i*, como la suma del número de puntuaciones en los anteriores *r-1* intervalos más el ranking de la puntuación en su correspondiente intervalo,

rank$_r$(y). Además, como el número de datos es el mismo para todos los intervalos, el número de datos en *r-1* intervalos dividido por el número total de datos *T* es equivalente a la división entre el número de intervalos *r-1* y el número de intervalos *M*, es decir,

$$\widetilde{F}(y) = \frac{rank(y)}{T} = \frac{\sum_{i=1}^{r-1} N_i + rank_r(y)}{T} = \frac{r-1}{M} + \frac{rank_r(y)}{T} \tag{5.3}$$

Tal y como se ha dicho anteriormente, el punto representativo para cada intervalo se escoge como aquel cuya probabilidad acumulada es el promedio de las probabilidades acumuladas de los puntos extremos del intervalo. Dado que en la expresión final de la ecuación (5.3) tanto el primer término como *T* son constantes para cada intervalo, el punto representativo para el intervalo *r* será aquel cuyo *rank$_r$(.)* sea el promedio del valor de *rank$_r$(.)* de los valores extremos, que son 1 y N_r. En consecuencia, el punto representativo es aquel que cumple *rank$_r$(y) = (N_r+1)/2*, lo que corresponde con la definición de la mediana de un conjunto de números. En conclusión, el punto representativo para cada intervalo es la mediana de las puntuaciones del intervalo.

En la figura 5-2, se muestran los histogramas de las puntuaciones para la normalización de rango y para la ecualización de histogramas sobre las puntuaciones de caras y de espectro de voz del apartado 6.2. Se puede observar que la normalización de rango modifica considerablemente la forma de los histogramas, mientras que HEQ adapta el histograma de las puntuaciones a la distribución de las puntuaciones de caras.

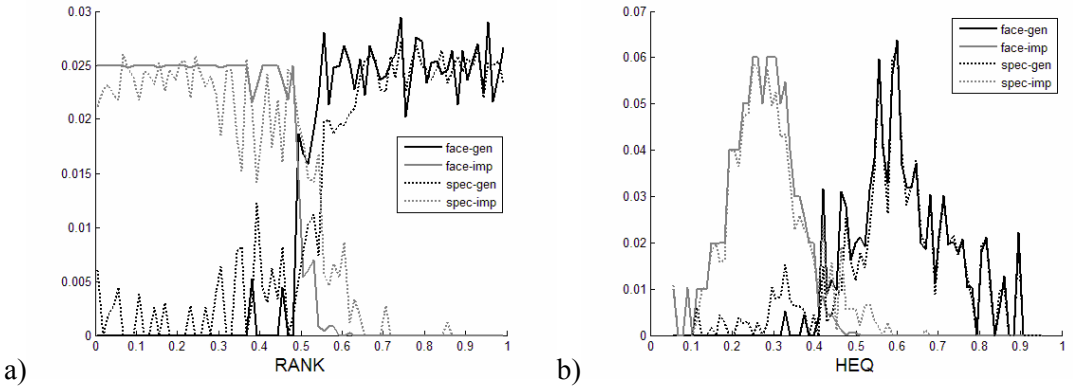


Figura 5-2: Histograma de las puntuaciones para la normalización de rango (a) y la ecualización de histogramas (b) tomando como referencia el histograma de las puntuaciones de caras para modalidades biométricas caras y voz.

Por otro lado, como se ha mencionado anteriormente, proponemos la ecualización sobre una distribución definida a priori.

5.2 Ecualización a gaussiana

En el caso de la ecualización sobre una distribución definida a priori, el punto representativo para cada intervalo se puede calcular a partir de la distribución de referencia de la siguiente manera

$$\frac{r-1}{M}+\frac{N_x}{2T_x}=\int_{-\infty}^{y_r} f_{ref}(\beta)d\beta \tag{5.4}$$

dondo N_x es la cantidad de datos de la modalidad biométrica de origen x en cada intervalo, T_x es el número total de datos, $f_{ref}(\beta)$ es la distribución de referencia, y y_r es el punto representativo en la distribución objetivo para el intervalo r. Por medio de la ecuación (5.4) se puede construir una tabla de búsqueda que permita encontrar el valor de referencia para cada intervalo.

Para el caso de verificación de hablantes, en (Pelenacos et al., 2001) se utiliza la ecualización a una distribución normal para la adaptación (*warping*) de tramas de parámetros de voz obtenidas sobre un intervalo de tiempo específico. Con ello se reduce el efecto del ruido aditivo en canales lineales.

En esta tesis, se han ecualizado parámetros y puntuaciones obtenidos por sistemas expertos unimodales a una distribución normal con media cero y la misma varianza que la distribución original. Esta técnica se ha denominado ecualización a gaussiana (*GEQ: Gaussian Equalization*). En este caso la función de referencia se define como

$$f_{ref}(\beta)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{\beta^2}{2\sigma^2}} \tag{5.5}$$

Esta técnica resulta útil para la normalización de parámetros. Sin embargo, los histogramas de puntuaciones de modalidades biométricas habitualmente se componen de dos lóbulos, uno para las puntuaciones de clientes y otro correspondiente a las puntuaciones de impostores, mientras que la distribución gaussiana se compone de un solo lóbulo.

En la figura 5-3, donde los histogramas de las puntuaciones de caras y espectro de voz se muestran para este tipo de ecualización, se puede observar que las puntuaciones de clientes e impostores se concentran alrededor del valor cero. Por este motivo, se puede esperar que esta técnica de normalización no obtenga buenos resultados para la normalización de puntuaciones. De hecho, los resultados obtenidos con esta técnica se han visto mejoradas por otras técnicas de normalización probadas en esta tesis para la fusión de puntuaciones. Estos resultados evidencian la importancia de la elección de la distribución de referencia en el proceso de ecualización.

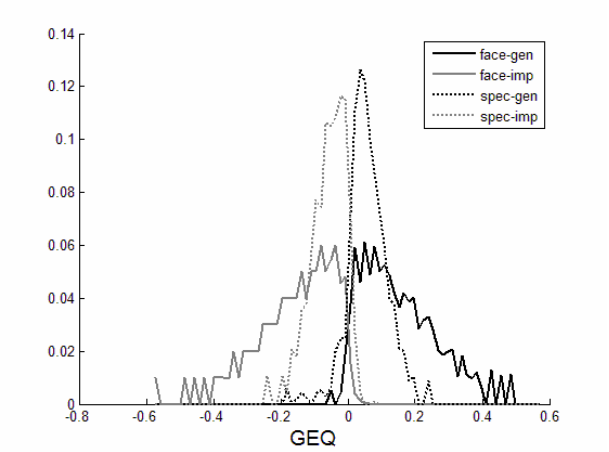


Figura 5-3: Histograma de las puntuaciones para la ecualización a gaussiana para caras y espectro de voz.

5.3 Ecualización de doble gaussiana

Según lo comentado en el apartado anterior, parece razonable que la selección de la distribución de referencia en el proceso de ecualización de puntuaciones tenga en cuenta las características estadísticas de las distribuciones biométricas. Por otro lado, en el apartado 4 se muestra que la información relativa a la estadística separada de clientes e impostores se puede utilizar en técnicas de normalización para mejorar los resultados obtenidos en sistemas de fusión multimodales.

En este apartado se propone una técnica par a la ecualización de puntuaciones, donde la distribución de referencia modela de forma separada la estadística de puntuaciones clientes e impostores mediante la suma de dos funciones gaussianas independientes. En esta propuesta, las medias de las distribuciones de clientes e impostores se han fijado a valores concretos. Como en el caso de JMN, se ha establecido la media de la distribución de clientes a 1 y la media de la distribución de impostores a -1.

Sin embargo, en lugar de utilizar la desviación estándar global como en GEQ, las desviaciones estándar de clientes e impostores se han estimado para modelar el solapamiento entre los lóbulos de clientes e impostores de las distribuciones separadas de las puntuaciones originales.

Esta técnica se ha denominado ecualización de doble gaussiana (*BGEQ: Bi-Gaussian Equalization*). En una primera versión de esta técnica y buscando la simplicidad, se dará el mismo valor a las dos desviaciones estándar. La distribución de referencia es, en este caso,

$$f_{ref}(\beta) = \frac{1}{2\sigma\sqrt{2\pi}}\left[e^{-\frac{(\beta+1)^2}{2\sigma^2}} + e^{-\frac{(\beta-1)^2}{2\sigma^2}} \right] \tag{5.6}$$

donde σ es la desviación estándar de ambas gaussianas.

Para calcular la desviación estándar σ, se debe considerar que una opción para modelar el solapamiento entre los lóbulos de clientes e impostores es mantener el error de reconocimiento de la modalidad biométrica original en la distribución de referencia dado que, de esta manera, el número de errores relativos, es decir, el número de datos en las regiones de solapamiento, se mantiene. Para esta técnica, tanto la tasa de error equivalente (EER) como la mínima tasa de error total (HTER) de la modalidad biométrica original se han utilizado para calcular la desviación estándar para BGEQ.

5.3.1 EER de la doble gaussiana igual al de la modalidad biométrica

La figura 5-3 muestra la transformación del histograma de la biometría original al histograma de las puntuaciones ecualizadas a doble gaussiana para puntuaciones de caras y espectro de voz de los datos de entrenamiento del apartado 6.2, utilizando EER como medida de error. Los histogramas de la parte superior de la figura, 5(a) y 5(b), muestran los histogramas originales para ambas modalidades.

Los EER son 2.50% y 9.52% respectivamente para los sistemas de caras y espectro de voz. En la parte central de la figura, 5(c) y 5(d), se muestran para ambas modalidades las distribuciones de referencia para la ecualización a doble gaussiana.

Dado que las gaussianas de clientes e impostores se han diseñado con la misma desviación estándar, el EER de la distribución de referencia se obtendrá definiendo el umbral como el promedio de las medias de las dos gaussianas. Por lo tanto, en este caso, dado que las medias de las distribuciones se han fijado a 1 y -1, el umbral será cero. Como el EER se puede calcular como la probabilidad de que el valor de una puntuación de impostor sea mayor que el umbral o que una puntuación de cliente sea menor que dicho umbral, la desviación estándar de las gaussianas para las caras y el espectro de voz deben cumplir la expresión en (5.7) para conseguir que el EER de la distribución de referencia sea igual al EER de la modalidad original.

$$\int_0^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\beta+1)^2}{2\sigma^2}} d\beta = EER \tag{5.7}$$

Según esta ecuación, la desviación estándar de la distribución de doble gaussiana se puede calcular para cada modalidad biométrica a partir del EER. En concreto, de acuerdo con los EER obtenidos en el apartado 6, 2.50% y 9.25% para las caras y el espectro de voz, la desviación estándar de las gaussianas son 0.5102 y 0.7637 respectivamente. La figura 5(e) muestra los histogramas de las puntuaciones ecualizadas con doble gaussiana.

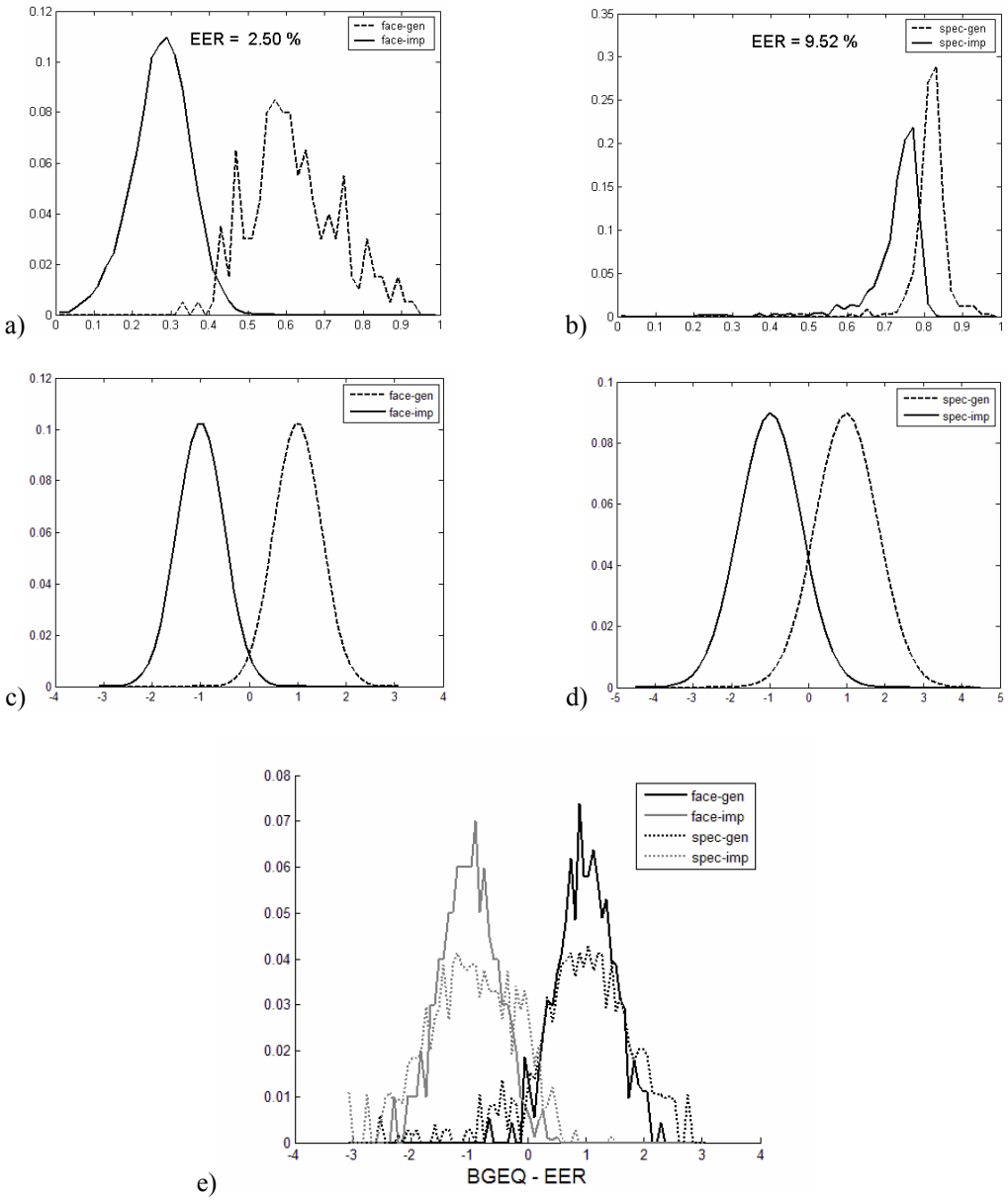


Figura 5-4: Ilustración del proceso de ecualización a doble gaussiana. Histogramas de las modalidades biométricas caras y espectro de voz ((a) y (b)), distribuciones de doble gaussiana de referencia ((c) y (d)) e histograma de las puntuaciones tras la ecualización a doble gaussiana (e).

5.3.2 HTER de la doble gaussiana igual al de la modalidad biométrica

En el caso en que se utiliza HTER como medida de error, las consideraciones anteriores son válidas cambiando EER por HTER, dado que de forma análoga al caso anterior, en una distribución suma de dos gaussianas con la misma varianza, el mínimo HTER se consigue con umbral cero.

Por lo tanto, para que el HTER de la doble gaussiana sea igual al de la modalidad biométrica original, es suficiente con cambiar EER por HTER en las expresiones anteriores. En consecuencia, la varianza de las distribuciones gaussianas se calcula en este caso según la ecuación (5.8).

$$\int_0^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\beta+1)^2}{2\sigma^2}} d\beta = HTER \tag{5.8}$$

Como se puede observar, dicha varianza se puede calcular únicamente a partir de su HTER.

5.3.3 Mismas desviaciones estándar de las distribuciones de clientes e impostores

En otra versión de la técnica BGEQ, las desviaciones estándar de los conjuntos originales de puntuaciones de clientes e impostores se mantienen en la distribución de referencia. Teniendo en cuenta que las puntuaciones se deben transformar de forma afín para conseguir que las medias de clientes e impostores sean 1 y -1 respectivamente, las desviaciones estándar de clientes e impostores de las distribuciones gaussianas $\sigma_{c,i}$ se deben transformar según

$$\sigma_{c,i} = 2\frac{\sigma_{c,i}^o}{\mu_c^o - \mu_i^o} \tag{5.9}$$

donde $\sigma_{c,i}^o$ son las desviaciones estándar de clientes e impostores originales, y μ_c^o y μ_i^o son las medias de clientes e impostores originales. La correspondiente distribución de referencia es, en este caso,

$$f_{ref}(\beta) = \frac{1}{2\sqrt{2\pi}} \left[\frac{e^{-\frac{(\beta+1)^2}{2\sigma_i^2}}}{\sigma_i} + \frac{e^{-\frac{(\beta-1)^2}{2\sigma_c^2}}}{\sigma_c} \right] \tag{5.10}$$

5.3.4 Algoritmo de cálculo de la ecualización de doble gaussiana

El algoritmo para la aplicación de la técnica de ecualización de doble gaussiana en la tarea de fusión multimodal es la siguiente para cada modalidad:

1. Fase de entrenamiento:
    - Calcular las desviaciones estándar de las gaussianas que componen la distribución de referencia de doble gaussiana a partir de los datos de entrenamiento de la modalidad biométrica.
        - A partir de la tasa de error equivalente (EER) o la mínima tasa media de error total (HTER) en el caso de utilizar una sola desviación estándar para las dos gaussianas, utilizando las ecuaciones (5.7) y (5.8) respectivamente y según tablas de la distribución normal estándar o software que implemente esta función.

- A partir de las desviaciones estándar originales de clientes e impostores si se mantienen las desviaciones estándar de las distribuciones originales.

- Dividir los datos de entrenamiento en *N* intervalos con la misma cantidad de datos en cada uno de ellos.

- Construir una tabla de correspondencia de la siguiente manera:

  - Primer intervalo: El menor valor del rango de correspondencia se establece como -∞ (que puede representarse por el menor valor de la puntuación de la modalidad biométrica proporcionado por el proveedor del sistema biométrico unimodal). El valor mayor del rango de correspondencia se estable como el punto medio entre el mayor valor del primer intervalo y el menor valor del segundo intervalo (la mitad de la suma). El valor de referencia del primer intervalo se puede encontrar por medio de (5.4) con $r = 1$.

$$\frac{N_{x,1}}{2T_x} = \int_{-\infty}^{y_1} f_{ref}(\beta) d\beta \tag{5.11}$$

  - Resto de intervalos: El valor menor del rango de correspondencia se estable como el valor mayor del rango anterior. El valor mayor se establece como el promedio del valor mayor en el intervalo actual y el valor menor del siguiente intervalo. El valor de referencia del intervalo se puede encontrar por medio de (5.4) para el correspondiente valor de *r*.

  - Último intervalo: Para el último intervalo, el valor mayor del rango de correspondencia se establece como ∞ (o mayor valor proporcionado por el proveedor). El resto de los valores se pueden calcular como en el caso anterior.

- Para todas las puntuaciones de entrenamiento de la modalidad biométrica, se debe encontrar el valor de referencia en la tabla de correspondencia para construir las puntuaciones ecualizadas de entrenamiento.

- Una vez que la ecualización de doble gaussiana se ha llevado a cabo para todas las modalidades biométricas, se puede entrenar el método de fusión escogido con las puntuaciones ecualizadas.

2. Fase de pruebas o test:

   - Para todas las puntuaciones de test de cada modalidad biométrica, se debe encontrar el valor de referencia en su correspondiente tabla de búsqueda. Con ello se obtienen las puntuaciones ecualizadas.

   - Aplicar el método de fusión seleccionado sobre las puntuaciones ecualizadas.

6 Experimentos de fusión de puntuaciones de locutor y caras

En este apartado se van a presentar los resultados de reconocimiento obtenidos por diferentes sistemas de fusión de puntuaciones a partir de sistemas expertos de reconocimiento de locutor y caras. Los sistemas de locutor se basan en parámetros de espectro de voz y prosodia y el sistema de reconocimiento de caras se basa en el algoritmo NMFFaces (Zafeiriou et al., 2005a).

En concreto, en el subapartado 6.1 se van a presentar los resultados de un sitema bimodal que utiliza puntuaciones de espectro de voz y caras. Para ello se van a comparar algunas de las técnicas de normalización más convencionales con las técnicas de normalización de media y varianza basadas en las estadísticas separadas de clientes e impostores (JMN, MSDSW y MVSW). Posteriormente estas técnicas se combinan con técnicas no afines de normalización (QLQ y HEQ).

En el subapartado 6.2 se introduce la prosodia como una tercera fuente de información para la fusión multimodal y se exploran diferentes estrategias de fusión. En este apartado se compara la normalización HEQ con las convencionales ZS y TANH. Además se explora el efecto de realizar ecualización de histograma tomando como referencia los histograms de las diferentes modalidades biométricas.

Finalmente, en el apartado 6.3 se realiza una amplia comparativa de diferentes técnicas de normalización, incluyendo las técnicas basadas en ecualización presentadas en esta tesis, siguiendo la estrategia de fusión que mejores resultados ha conseguido en el apartado anterior. Se incluyen en la comparativa diversas normalizaciones del estado del arte y las normalizaciones HEQ, GEQ y BGEQ en sus diversas variantes.

6.1 Fusión bimodal de espectro de voz y caras mediante normalizaciones estadísticas.

En este apartado, se van a presentar los sistemas de reconocimiento de voz y caras involucrados en los experimentos de fusión y los resultados experimentales obtenidos para un sistema de fusión bimodal de puntuaciones de voz y caras. Se van a comparar técnicas de normalización del estado del arte con técnicas de normalización de media y varianza basadas en las estadísticas separadas de clientes e impostores (JMN, MSDSW y MVSW). Posteriormente estas técnicas se combinan con técnicas no afines de normalización (QLQ y HEQ).

6.1.1 Sistemas unimodales.

En este apartado se revisan los sistemas unimodales de reconocimiento de locutor y caras. En este caso los sistemas de voz se basan en la información espectral, en concreto parámetros mel-cepstrum con GMM. La información facial se basa en los parámetros NMFFaces.

6.1.1.1 *Información espectral de voz*

Uno de los principales requisitos de cualquier tarea de reconocimiento de voz/locutor es construir un conjunto de datos, a partir de la señal de voz de entrada, de tal forma que se facilite la tarea del algoritmo de clasificación. Aquí se utilizarán los Mel-Frequency Cepstrum Coefficients (MFCC) basados en la percepción auditiva humana.

Para calcular los parámetros MFCC, dada una señal de entrada, la dividimos en tramas y calculamos la Transformada de Fourier para cada una de la siguiente manera:

$$X_a[k] = \sum_{n=0}^{N-1} x[n] e^{-j 2\pi n k / N}, \qquad 0 \le k < N \tag{6.1}$$

Se define ahora el banco de M filtros triangulares, solapados, de la siguiente forma:

$$H_m[k] = \begin{cases} 0 & k < f[m-1] \\ \dfrac{2(k - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & f[m-1] \le k \le f[m] \\ \dfrac{2(f[m+1] - k}{(f[m+1] - f[m-1])(f[m+1] - f[m])} & f[m-1] \le k \le f[m] \\ 0 & k < f[m-1] \end{cases} \tag{6.2}$$

Este banco de filtros promedia el espectro alrededor de cada frecuencia central y tiene bandas frecuenciales con ancho de banda creciente. Tiene la siguiente forma:

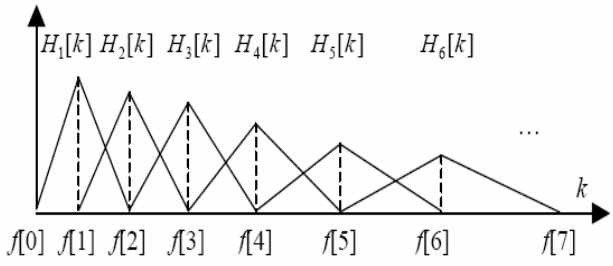


Figura 6-1: Banco de filtros para cálculo de MFCC.

El uso de estos filtros permite mapear las amplitudes del espectro sobre la escala Mel. Y el logaritmo de la energía a la salida de cada filtro se puede calcular como sigue:

$$S[m] = \ln\left[\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right], \qquad 0 \le m < M \tag{6.3}$$

Finalmente, el MFCC es la transformada coseno discreta de la salida de los M filtros:

$$c[n] = \sum_{m=0}^{M-1} S[m]\cos(\pi n(m+1/2)/M), \qquad 0 \le m < M \tag{6.4}$$

donde M oscila entre 24 y 40 dependiendo de la aplicación. En cuanto al número de coeficientes cepstrales, para aplicaciones de reconocimiento de voz/locutor, se suelen utilizar los 13 primeros.

La forma de hablar de cada locutor se puede modelar estadísticamente. Los Modelos Ocultos de Harkov (*HMM: Hidden Harkov Models*) son una herramienta potente que nos permite caracterizar la voz humana. Un HMM consiste en una serie de estados, interconectados entre sí, y a los que se puede llegar desde otro estado con una determinada probabilidad. Además, y de ahí el significado de Oculto, cada estado tiene asociada una determinada distribución de probabilidad gaussiana. Dicha función de probabilidad modela cómo se distribuyen los vectores de características asociados a un determinado estado.

Para aplicaciones de reconocimiento de locutor con texto independiente, la topología más común es un HMM con un único estado. Por eso también se le conoce como Gaussian Mixture Model (GMM).

El uso de distribuciones gaussianas se justifica con el Teorema Central del Límite. Según éste, cualquier distribución de probabilidad se puede conseguir mediante una combinación ponderada de múltiples distribuciones, cada una de ellas con una media y una varianza determinada, como se puede ver en la expresión siguiente:

$$g(x) = \sum_{i=1}^{N} \lambda_i N(x; \mu_i, \Sigma_i) \tag{6.5}$$

donde λ_i son los pesos que ponderan cada mezcla, *N(.)* es una distribución de probabilidad gaussiana, μ_i y Σ_i son la media y la matriz de covarianza de las gaussianas, y *N* es el número total de componentes. Estos parámetros se relacionan de la forma siguiente:

$$N(x; m_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} e^{-(1/2)(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)} \tag{6.6}$$

Habitualmente se utilizarán matrices de covarianza diagonales. Esto se hace por tres razones. En primer lugar, el modelado de un GMM con una matriz de covarianza completa se puede conseguir igualmente con una matriz de covarianza diagonal de orden superior. En segundo lugar, las matrices diagonales son computacionalmente más eficientes. Y en tercer lugar, se ha observado empíricamente que los GMMs de matrices de covarianza diagonales superan a los GMMs de matrices completas.

En este apartado, para obtener las puntuaciones de voz se ha usado un sistema de reconocimiento dependiente del texto basado en Modelos Ocultos de Markov. Las señales de voz se han segmentado en tramas de 20 ms con un desplazamiento de 10 ms previamente a extraer parámetros mel-cepstrum. Se han utilizado dos combinaciones diferentes de parámetros MFCC: 20 parámetros mel-cepstrum (MFCC20) y 60 parámetros mel-cepstrum (MFCC60), incluyendo los 20 originales, 20 deltas y 20 aceleraciones. El sistema de reconocimiento está basado en GMM con 32 gaussianas por modelo.

6.1.1.2 *Información facial*

Los sistemas de reconocimiento facial se basan comúnmente en la descomposición de la cara en partes distribuidas de forma dispersa: ojos, nariz, boca, etc. Zafeiriou et al. utilizan la factorización no negativa de matrices (*NMF: Non-negative matrix factorization*) para conseguir que la representación distribuida de parámetros localizados representen las partes constituyentes de la cara en las imágenes faciales (Zafeiriou et al., 2005a, Zafeiriou et al., 2005b).

La factorización no negativa de matrices es una técnica de reconocimiento facial basada en la apariencia que utiliza las técnicas convencionales de análisis de componentes. Dada una base de datos de imágenes de caras representadas por una matriz $\mathrm{X} \in \Re_+^{F\times L} = \left[x_{i,j}\right]$, donde $x_{i,j}$ es el elemento número *i* de la imagen número *j*, NMF permite encontrar dos matrices $\mathrm{Z} \in \Re_+^{F\times M} = \left[z_{i,k}\right]$ y $\mathrm{H} \in \Re_+^{M\times L} = \left[h_{k,j}\right]$ que cumplen

$$\mathrm{X} \approx \mathrm{ZH} \tag{6.7}$$

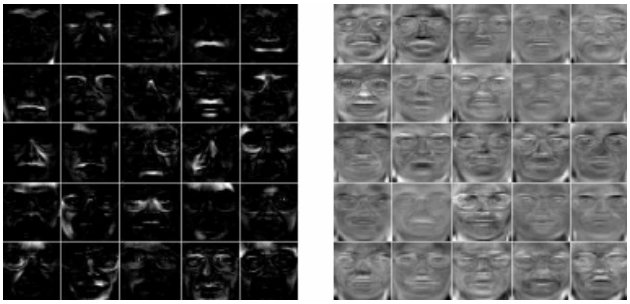
La imagen facial $\mathbf{u}_j$ tras la descomposición NMF se puede escribir como $\mathbf{u}_j \approx \mathrm{Z}\mathbf{h}_j$, donde $\mathbf{h}_j$ es la columna número *j* de H. Por lo tanto, las líneas de la matriz Z se pueden considerar como imágenes

base y el vector $\mathbf{h}_j$ como el correspondiente vector de ponderación. Los vectores $\mathbf{h}_j$ se pueden considerar también los vectores proyectados de un espacio de parámetros de menor dimensión.

NMF impone condiciones de no negatividad en los elementos de $z_{i,k}$ y de $h_{k,j}$. En consecuencia, solo se permiten combinaciones no sustractivas. Se considera que se corresponde mejor a la noción intuitiva de combinar diferentes partes de la cara para crear la imagen facial completa.

Dado que $\mathbf{x}_j \approx \mathbf{Z}\mathbf{h}_j$, una forma natural de computar la proyección de $\mathbf{x}_j$ a un espacio de menor dimensión utilizando NMF es $x'_j = \left(\mathrm{Z}^T\mathrm{Z}\right)^{-1}\mathrm{Z}^T x_j$.

El método NMF, no utiliza la información sobre como las diferentes imágenes faciales se separan en diferentes clases de caras. La forma más directa de explotar la información discriminativa en NMF es intentar descubrir las proyecciones discriminativas para los vectores de imágenes faciales tras la proyección a la matriz de imágenes base $\mathrm{Z}' = \left(\mathrm{Z}^T\mathrm{Z}\right)^{-1}\mathrm{Z}^T$. Supongamos que la base de datos de imágenes faciales contienen K clases diferentes (personas) con cada clase r conteniendo N_r imágenes. Y supongamos que la matriz X está organizada de la siguiente manera: la columna j de la base de datos X es la imagen número ρ de la clase r. Entonces, $j = \sum_{i=1}^{r-1} N_i + \rho$.



(a) (b)

Figura 6-2: Un conjunto de 25 imágenes base para NMF (a) y NMFFaces (b).

No hay ningún límite superior en el número de bases que se pueden construir usando la descomposición NMF y a menos que se cree un número de bases limitado la matriz S_ω es singular. Para resolver este problema se aplica la misma solución que en el caso de las Fisherfaces (Lee et al., 2001). El número total de parámetros discriminativos obtenidos según este proceso es *K-1*. Las puntuaciones de reconocimiento facial utilizado en los siguientes apartados se han calculado de esta manera según el método NMFFaces en el que las imágenes base finales son cercanas a las partes constitutivas de la cara.

La figura 6-2, obtenida de (Zafeiriou et al., 2005a), muestra un conjunto de 25 imágenes base para las técnicas NMF y NMFFaces.

6.1.2 Preparación de los experimentos

Los resultados presentados en este apartado se han obtenido mediante fusión de puntuaciones. Para conseguir este propósito, los resultados obtenidos por un sistema experto en reconocimiento de locutor y un sistema experto en reconocimiento de caras se han combinado para crear una base de datos quimérica con 10823 usuarios, mediante la combinación de 110 usuarios de la base de datos POLYCOST y 270 usuarios de la base de datos XM2VTS.

La base de datos utilizada en los experimentos de reconocimiento de locutor es la POLYCOST (Hernando et al., 1998), una base de datos de voz telefónica con una frecuencia de muestreo de 8 kHz. Esta base de datos contiene 134 hablantes. Se ha utilizado una frase en inglés pronunciada 10 veces por cada hablante.

La base de datos utilizada para los experimentos de reconocimiento facial es XM2VTS (Lüttin et al., 1998) de la Universidad de Surrey. Es una base de datos multimodal formada por imágenes de caras, secuencias de vídeo y grabaciones de voz de 295 personas. Para los experimentos de este apartado únicamente se han utilizado las imágenes faciales. Hay cuatro imágenes faciales por individuo.

Un conjunto de registros de cada base de datos se ha utilizado para entrenar los expertos de reconocimiento unimodales. Las técnicas de normalización y fusión se han entrenado a partir de las puntuaciones obtenidas por estos expertos para un segundo conjunto de registros. Finalmente, los resultados que se presentan se han obtenido mediante la aplicación de las técnicas de fusión implementadas sobre las puntuaciones unimodales obtenidas por un tercer conjunto de ocurrencias.

A partir de los sistemas expertos de reconocimiento de caras y de locutor, se disponía de 1488 experimentos de voz (para cada combinación de parámetros) y 33361 experimentos de caras. Las puntuaciones de todos los usuarios se han dividido en dos grupos, como se ha comentado anteriormente, para el entrenamiento y el testeo de los sistemas de fusión. Debido al gran número de experimentos necesarios para que las pruebas tengan validez estadística, según el número de errores obtenidos, ha sido necesaria la combinación de un usuario de una biometría con más de un usuario de la otra modalidad. Sin embargo, se ha intentado minimizar el número de combinaciones.

Mediante la combinación de las puntuaciones unimodales, se ha creado un total de 29480 experimentos unimodales para entrenar las técnicas de normalización y fusión y 30040 se han formado para testearlas. Todos los resultados mostrados en estos apartados se han obtenido con las puntuaciones de test.

De forma previa al entrenamiento de las técnicas de normalización y fusión, se ha eliminado el 1% de las puntuaciones de clientes e impostores de mayor o menor valor para evitar el efecto negativo de estas puntuaciones. Además, el número de intervalos en la ecualización de histogramas se ha establecido a 1000 y, para las SVM, se ha utilizado un kernel RFB, ***radial basis function***, y se ha puesto el valor 100 a la constante ***C*** que controla el compromiso entre permitir errores de entrenamiento y forzar márgenes rígidos. Estos parámetros se han ajustado para obtener los mejores resultados para estas técnicas.

Para las tablas en que se muestras sumas de varianzas de las puntuaciones de clientes e impostores, todas las varianzas se han calculado a partir de las puntuaciones normalizadas mediante normalización conjunta de medias, tanto para las puntuaciones unimodales como multimodales, con la media de las puntuaciones de clientes igual a uno y la media de las puntuaciones de impostores igual a menos uno, para que dichos indicadores estadísticos sean comparables.

6.1.3 Resultados

En la tabla I, se presentan los resultados correspondientes a la tasa de error equivalente (EER: Equal Error Rate) y la suma de las varianzas de las puntuaciones de clientes e impostores para los sistemas unimodales de reconocimiento de locutor y reconocimiento de caras.

	EER	Suma de Varianzas
Voz		
MFCC20	5.096 %	1.3284
MFCC60	2.670 %	0.7488
Caras	2.064 %	0.6970

Tabla 6-1: EER y suma de las varianzas de clientes e impostores para los sistemas unimodales basados en voz y caras.

Estos resultados muestras la relación directa que existe entre las tasas de reconocimiento y las varianzas de las puntuaciones unimodales. El sistema de voz que utiliza 20 parámetros mel-cepstrum obtiene los peores resultados y tiene la mayor suma de varianzas. Por otro lado, el sistema de reconocimiento facial obtiene la menor tasa de reconocimiento y, asimismo, la menor suma de varianzas.

En las siguientes tablas se muestra el EER obtenido por la combinación de las diferentes técnicas de normalización y fusión aplicadas sobre las puntuaciones de voz y caras. Para obtener estos resultados, se han comparado cinco métodos de normalización afines: *min-max* (MM), *z-score*

(ZS), normalización conjunta de medias (JMN), minimización de la suma de las desviaciones estándar (MSDSW) y minimización de la suma de las varianzas (MVSW), tres métodos de normalización no afines: dos-cuádricas (QQ), cuádrica-línea-cuádrica (QLQ) y ecualización de histogramas (HEQ) a una de las modalidades biométricas, con tres métodos de fusión: suma simple (SS), ponderación en función del resultado de reconocimiento de cada biometría unimodal o *matcher weighting* (MW) y una máquina de vector soporte (SVM).

Sin embargo, los resultados obtenidos mediante la normalización QQ han sido en todos los casos superados por los obtenidos mediante la normalización QLQ y, por este motivo, no se presentan. Además, no todas las combinaciones de normalización y fusión se han aplicado. Las técnicas de reducción de varianzas, MSDSW y MVSW han sido diseñados específicamente para la fusión con suma simple, y la SVM aplica implícitamente una normalización MM y, por este motivo, solo se presentan las combinaciones con esta técnica.

En primer lugar se comparan los resultados obtenidos con todas estas técnicas sobre el sistema de reconocimiento de locutor con 20 parámetros y el sistema de reconocimiento facial. En la tabla 6-2, se presentan los resultados obtenidos por las normalizaciones afines y todos los métodos de fusión.

	SS	MW	SVM
MM	0.776 %	1.099 %	**0.672 %**
ZS	0.752 %	0.832 %	-
JMN	0.849 %	0.789 %	-
MSDSW	0.789 %	-	-
MVSW	0.782 %	-	-

Tabla 6-2: Resultados de fusión para MFCC20 y caras para las normalizaciones afines.

El mejor resultado lo obtiene la fusión SVM. Para los métodos de fusión combinatorios, se obtienen diferencias menores al 5% en el número de errores para ZS-SS, MM-SS y para JMN-SS, JMN-MW, MSDSW-SS y MVSW-SS.

Las técnicas de normalización no afines se han aplicado sobre las puntuaciones previamente normalizadas con MM debido a la definición de las técnicas QQ y QLQ y por simplicidad en el caso de la ecualización de histogramas. Sin embargo, para aplicar las técnicas de reducción de varianza, las puntuaciones unimodales deben tener sus medias normalizadas mediante JMN. Por este motivo, en nuestras pruebas se ha realizado la combinación de normalizaciones no afines y afines. Debido a las propiedades de las transformaciones no afines, la normalización *min-max* no tiene ningún efecto sobre ellas. Por lo tanto, presentamos los resultados para QLQ, QLQ-ZS,

QLQ-JMN, QLQ-MSDSW, QLQ-MVSW, HEQ, HEQ-ZS, HEQ-JMN, HEQ-MSDSW y HEQ-MVSW para las técnicas de fusión combinatorias.

	SS	MW
QLQ	0.796 %	1.092 %
QLQ-ZS	0.746 %	0.789 %
QLQ-JMN	0.819 %	0.796 %
QLQ-MSDSW	0.789 %	-
QLQ-MVSW	0.789 %	-
HEQ	0.802 %	0.716 %
HEQ-ZS	0.759 %	0.739 %
HEQ-JMN	0.759 %	0.739 %
HEQ-MSDSW	**0.636 %**	-
HEQ-MVSW	**0.636 %**	-

Tabla 6-3: Resultados para las normalizaciones no afines para la fusión de puntuaciones de MFCC20 y caras.

La combinación de HEQ y las normalizaciones basadas minimización de las varianzas, MSDSW y MVSW, obtienen el mejor resultado, que incluso mejora el resultado obtenido por la SVM, y superan en un 15% el rendimiento obtenido por las técnicas afines convencionales y en un 10% los resultados obtenidos por todas las técnicas combinatorias de fusión probadas. Cabe remarcar también que con la aplicación de la normalización QLQ, se consigue una reducción del EER para todas las combinaciones excepto para QLQ-ZS-SS. En el caso de HEQ, la mejora se obtiene en todos los casos excepto para HEQ-SS y HEQ-ZS-SS, y ha alcanzado el 19% para sus mejores resultados, HEQ-MSDSW-SS y HEQ-MVSW-SS. La mejora obtenida respecto a la biometría unimodal con mejor rendimiento (caras) es del 69%.

En este punto, se van a comparar algunos de los resultados de reconocimiento obtenidos con la suma de las varianzas de las puntuaciones de clientes e impostores. En particular, se van a comparar los resultados y las varianzas de las puntuaciones multimodales basadas en JMN, MSDSW y MVSW y su combinación con HEQ.

En la tabla 6-4 se puede observar una clara relación entre la suma de las varianzas de las puntuaciones de clientes e impostores y la tasa de reconocimiento de las biometrías multimodales. Sólo en el caso de HEQ-JMN-SS, una menor suma de varianzas no conlleva un mejor rendimiento. Para el resto de técnicas, una menor suma de varianzas supone un menor EER. Esta relación se muestra en la figura 6-3.

	SS	MW
JMN	0.849 %	0.789 %
JMN – Suma de varianzas	0.5149	0.4813
MSDSW	0.789 %	-
MSDSW – Suma de varianzas	0.4822	-
MVSW	0.782 %	-
MVSW – Suma de varianzas	0.4793	-
HEQ – JMN	0.759 %	0.739 %
HEQ – JMN – Suma de varianzas	0.4000	0.4496
HEQ – MSDSW	0.636 %	-
HEQ – MSDSW – Suma de varianzas	0.3992	-
HEQ – MVSW	0.636 %	-
HEQ – MVSW – Suma de varianzas	0.3992	-

Tabla 6-4: EER y suma de varianzas para JMN, MSDSW y MVSW y su combinación con HEQ (MFCC20 y caras).

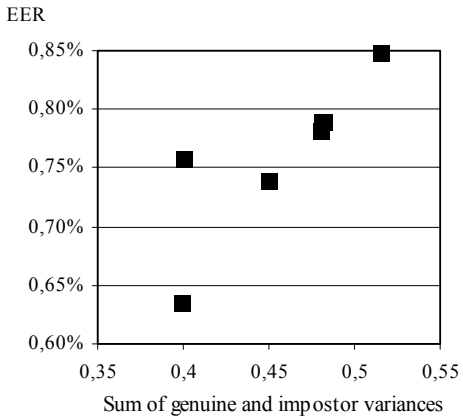


Figura 6-3: Relación entre suma de varianzas y EER para JMN, MSDSW y MVSW y su combinación con HEQ (MFCC20 y caras).

Cuando se utilizan 60 parámetros para el sistema de reconocimiento de locutor, se puede realizar un análisis similar de los resultados obtenidos, teniendo en cuenta que, en este caso, los dos sistemas de verificación unimodal tienen un rendimiento similar.

Los resultados obtenidos por las técnicas de normalización afines en combinación con las técnicas de fusión se muestras en la tabla 6-5.

El mejor resultado lo obtiene la fusión mediante SVM. Para los sistemas de fusión combinatorios, MSDSW obtiene el mejor resultado de reconocimiento combinado con SS. Sin embargo, las normalizaciones MVSW y MM obtienes resultados similares.

	SS	MW	SVM
MM	0.383 %	0.393 %	**0.359 %**
ZS	0.426 %	0.413 %	-
JMN	0.429 %	0.419 %	-
MSDSW	0.376 %	-	-
MVSW	0.383 %	-	-

Tabla 6-5: Resultados de fusión para MFCC60 y caras para las normalizaciones afines.

Los resultados obtenidos con la combinación de técnicas de normalización no afines y los métodos de fusión combinatorios son los mostrados en la tabla 6-6.

	SS	MW
QLQ	0.353 %	0.363 %
QLQ-ZS	0.390 %	0.370 %
QLQ-JMN	0.399 %	0.390 %
QLQ-MSDSW	0.353 %	-
QLQ-MVSW	0.353 %	-
HEQ	0.369 %	0.360 %
HEQ-ZS	0.363 %	**0.346 %**
HEQ-JMN	0.363 %	**0.346 %**
HEQ-MSDSW	0.353 %	-
HEQ-MVSW	0.353 %	-

Tabla 6-6: Resultados para las normalizaciones no afines para la fusión de puntuaciones de MFCC60 y caras.

Mediante el uso de normalizaciones no afines, el mejor resultado lo obtiene la utilización de ecualización de histograma junto con *z-score* y normalización conjunta de medias, en combinación con la fusión *matcher weighting*, que mejoran los obtenidos mediante SVM. Para el mejor resultado, la reducción de la tasa de error equivalente respecto a los métodos de normalización afines es de un 8%. Todos los resultados obtenidos mediante la utilización de normalizaciones no afines han mejorado los obtenidos con las correspondientes normalizaciones afines aunque, es este caso, hay una menor diferencia entre dichos resultados que en el caso anterior. La mejora obtenida respecto a la biometría unimodal con mejor rendimiento (caras) es del 83%.

Cabe resaltar que los rendimientos de los sistemas de fusión son más similares cuando se utilizan 60 parámetros en el sistema de reconocimiento de locutor, es decir, cuando los dos sistemas

unimodales tienen una menor diferencia entre sus resultados de reconocimiento. En este caso, la normalización y los factores de ponderación son más similares para los diferentes sistemas de normalización y fusión que cuando los sistemas biométricos unimodales tienen mayor diferencia en su rendimiento.

La tabla 6.7 resume los resultados y las sumas de las varianzas de las técnicas basadas en JMN, MSDSW y MVSW y su combinación con HEQ.

	SS	MW
JMN	0.429 %	0.419 %
JMN – Suma de varianzas	0.3655	0.3650
MSDSW	0.376 %	-
MSDSW – Suma de varianzas	0.3662	-
MVSW	0.383 %	-
MVSW – Suma de varianzas	0.3656	-
HEQ – JMN	0.363 %	0.346 %
HEQ – JMN – Suma de varianzas	0.3676	0.3673
HEQ – MSDSW	0.353 %	-
HEQ – MSDSW – Suma de varianzas	0.3673	-
HEQ – MVSW	0.353 %	-
HEQ – MVSW – Suma de varianzas	0.3673	-

Tabla 6-7: EER y suma de varianzas para JMN, MSDSW y MVSW y su combinación con HEQ (MFCC60 y caras).

En este caso, no se puede establecer una relación directa entre la suma de las varianzas y las tasas de verificación obtenidas. Sin embargo, la diferencia entre la suma de las varianzas de los diferentes métodos es muy pequeña y los resultados de reconocimiento cercanos. Por lo tanto, se puede concluir que de forma general existe una relación entre las varianzas y los resultados de reconocimiento, pero esta conclusión no es aplicable cuando hay diferencias pequeñas entre las varianzas.

Más allá de los resultados obtenidos por cada técnica de fusión, hay otras cuestiones que se deben tener en cuenta a la hora de escoger las técnicas de fusión. Para finalizar este apartado realizaremos una comparación de las ventajas y desventajas de cada técnica.

La aplicación de la normalización MM es simple debido a que solo depende de los valores extremos de las puntuaciones. Sin embargo, esta técnica puede verse afectada por la presencia de *outliers*. ZS, JMN, MSDSW y MVSW dependen de las medias y varianzas de las puntuaciones y, por este motivo, se ven menos afectadas por los *outliers*. Además, MSDSW y MVSW, tienen en

cuenta la reducción de las varianzas separadas de clientes e impostores de las puntuaciones multimodales. Los mejores resultados obtenidos por todas estas técnicas son similares.

Los parámetros necesarios para la aplicación de las normalizaciones QQ y QLQ son el centro y la anchura de la región de solapamiento de las puntuaciones de entrenamiento de clientes e impostores. Aunque se obtiene cierta mejora mediante la aplicación de estas técnicas, la presencia de *outliers* en las puntuaciones máximas de impostores o mínimas de clientes puede afectar gravemente a estas técnicas. HEQ toma como referencia una aproximación a la totalidad de las propiedades estadísticas de las puntuaciones unimodales. En consecuencia, esta técnica es más robusta a la presencia de *outliers* aunque los requerimientos computacionales son mayores en la fase de entrenamiento. Esta técnica ha obtenido los mejores resultados de reconocimiento.

En cuanto a las técnicas de fusión, SS es la técnica más directa y simple y MW se basa en los resultados de reconocimiento individuales obtenidos por cada sistema biométrico unimodal. Ambos métodos son fáciles de entrenar y de aplicar en la fase de test. Las máquinas de aprendizaje, como las SVM, requieren mayores recursos computacionales y de memoria, tanto en la fase de entrenamiento como en la fase de test. En esta sección se ha demostrado que los resultados obtenidos por una máquina de aprendizaje pueden ser mejorados mediante el uso del conocimiento de las estadísticas y el rendimiento de las biometrías unimodales.

Como conclusión de este apartado cabe destacar que, en el ámbito de la fusión a nivel de puntuaciones, se han utilizado diversas normalizaciones basadas en el tratamiento por separado de las estadísticas de las puntuaciones de clientes e impostores. En concreto, se ha utilizado la normalización conjunta de medias y métodos de normalización para la minimización de las varianzas: minimización de la suma de las desviaciones estándar y minimización de la suma de las varianzas. Este último método conlleva que la biometría multimodal tenga una menor suma de las varianzas de puntuaciones de clientes e impostores que las de las modalidades biométricas originales.

Además, se ha utilizado la ecualización de histogramas como método de normalización. Dicha ecualización toma como referencia el histograma de la modalidad biométrica con mejor resultado de reconocimiento, que es de esperar que tenga menores varianzas separadas, para obtener una mayor reducción de las varianzas multimodales.

JMN, MSDSW y MVSW mejoran o obtienen resultados similares que las técnicas convencionales mediante la utilización de métodos de fusión combinatorios. Además, el uso de ecualización de histogramas como método de normalización mejora los resultados obtenidos por las técnicas de normalización convencionales y otras técnicas de normalización no afines. En particular, la combinación de HEQ, MSDSW y MVSW con una técnica de fusión ponderada han obtenido los mejores resultados, incluso los obtenidos mediante el uso de una SVM.

Por otro lado, se ha probado una cierta relación entre las varianzas de las puntuaciones de clientes e impostores y la fiabilidad de una biometría. Esta relación se puede explotar para mejorar los resultados de las biometrías multimodales.

6.2 Estrategias para la fusión de espectro de voz, prosodia y caras

En este apartado se presentan la configuración de los experimentos y los resultados obtenidos por medio de diferentes técnicas y estrategia de fusión. En el siguiente subapartado, se detallan los sistemas de reconocimiento utilizados en los experimentos de fusión, basados en espectro de voz, tanto parámetros espectrales como prosodia, y caras. En el subapartado 6.2.2, se presentan los resultados experimentales obtenidos de acuerdo con los diferentes esquemas de fusión, combinando diferentes técnicas y estrategias.

6.2.1 Sistemas unimodales

En este apartado se revisan los sistemas unimodales de reconocimiento de locutor y caras. En este caso los sistemas de voz se basan en la información espectral, incluyendo parámetros espectrales y prosodia. La información facial, al igual que en el apartado 6.1 se basa en los parámetros NMFFaces.

6.2.1.1 *Información espectral de voz*

La señal de voz puede ser sujeto de variaciones causadas por diversos factores: las condiciones físicas del hablante (dimensiones de los pliegues vocales, tracto vocal, estado de salud, etc.), condiciones emocionales, características dialectales y sociolingüísticas, características del hablante o incluso por el ambiente en que la señal de voz se produce. Para un correcto funcionamiento de un sistema biométrico, los parámetros utilizados para caracterizar la señal de voz deben tener las siguientes capacidades: discriminar entre hablantes, ser transparentes a la variabilidad en el mismo hablante, no cambiar con el tiempo, no ser imitables, no ser afectadas por ruidos ambientales o de transmisión, ser fácilmente mensurables, y ocurrir de forma natural y frecuente en el habla (Peskin et al., 2003).

Los parámetros espectrales son aquellos que sólo tienen en cuenta el nivel acústico de la señal, como magnitudes espectrales, frecuencias de los formantes, etc. y están más relacionados con las características físicas del hablante.

Los parámetros prosódicos tienen en cuenta otros niveles lingüísticos más relacionados con hábitos adquiridos y estilo, como niveles léxicos, prosódicos o fonéticos, y analizan diversas características como entonación, duración de los sonidos, tipo de vocabulario, etc.

Parámetros espectrales

Los coeficientes cepstrales son la forma más habitual de representar la envolvente espectral de una trama de voz en los actuales sistemas de reconocimiento de locutor. Estos parámetros son la representación más prevalente de la señal de voz y contienen un gran nivel de especifidad del hablante. Los parámetros mel-cepstrum convencionales provienen de un conjunto de energías de un banco de filtros con escala mel (LFBE) $S(k)$, $k=1,...,Q$. La secuencia de coeficientes cepstrales es casi incorrelada y una representación compacta del espectro de voz.

Sin embargo, los coeficientes cepstrales tienen al menos tres desventajas: 1) no tienen un significado físico claro y útil como las energías de un FB; 2) requieren una transformación lineal a partir de las energías de un FB o de los coeficientes LP; y 3) en los modelos ocultos de Markov (HMM) con funciones de probabilidad gaussianas de matriz de covarianza diagonal, la forma de la ventana no tiene ningún efecto y sólo su longitud, es decir, el número de parámetros, es una variable de control. Para superar estas desventajas, (Nadeu et. al, 1995) presenta una alternativa al uso de cepstrum en el reconocimiento de voz que consiste en un simple procesado lineal en el dominio LFBE. La transformación de la secuencia $S(k)$ a coeficientes cepstrales se evita filtrando esta secuencia. Se ha denominado a esta operación filtrado frecuencial (*FF: Frequency Filtering*) para indicar que la convolución se realiza en el domino frecuencial. FF produce dos efectos, decorrelación y ponderación, en un solo paso utilizando un simple filtro FIR paso alto de primer o segundo orden. Además, el filtrado frecuencial produce una ponderación cepstral de forma implícita en modelos ocultos de Markov con funciones de probabilidad gaussianas de matriz de covarianza diagonal.

Parámetros prosódicos

Los humanos utilizamos habitualmente diferentes niveles de información para reconocer a los demás únicamente a partir de su voz: timbre de la voz, una risa característica, una expresión repetida con asiduidad, etc. (Campbell et al., 2003; Reynolds et al., 2003). La información prosódica se utiliza, por ejemplo, cuando se escucha una conversación a través de un muro, o para discernir a los hablantes objeto de una imitación humorística (Reynolds et al., 2003). En consecuencia, aunque las características prosódicas no obtienen buenos resultados de reconocimiento utilizados de forma aislada, proporcionan una información complementaria y mejoran los resultados cuando se fusionan con sistemas de reconocimiento basados en el espectro de voz.

Además, algunas de estas características tienen la ventaja de que son más robustas al ruido que las de bajo nivel (Kajarekar et al., 2003). Los patrones espectrales pueden verse afectados por las características frecuenciales del canal de transmisión y la información espectral depende también del volumen de la voz y la distancia entre el hablante y los micrófonos, mientras que, por ejemplo, la frecuencia fundamental no se ve afectada por estas variaciones (Atal, 1972).

El sistema de reconocimiento prosódico utilizado para obtener los resultados de este apartado ha sido constituido por un total de 9 parámetros prosódicos ya utilizados en (Carey et al., 1996) y que se pueden dividir en dos grandes grupos:

Parámetros relacionados con la duración de palabras y segmentos de voz:

- logaritmo del número de tramas por palabra promediado para todas las palabras
- longitud promedio de los segmentos de voz internos en las palabras
- longitud promedio de los segmentos sin voz internos en las palabras

Parámetros relacionados con el tono (*pitch*)

- media del logaritmo de F0
- máximo del logaritmo de F0
- logaritmo del rango de F0 (máximo F0 – mínimo F0)
- "pseudo pendiente" del tono calculado como: (último F0 – primer F0) / número de tramas en la palabra
- pendiente promedio sobre todos los segmentos de una estilización lineal por tramas de F0

6.2.1.2 *Información facial*

De igual manera que en el apartado anterior, el experto en reconocimiento facial está basado en el algoritmo NMFFaces (Zafeiriou et al., 2005a). Como se ha comentado anteriormente, los sistemas de reconocimiento facial se basan comúnmente en la descomposición de la cara en partes distribuidas de forma dispersa: ojos, nariz, boca, etc. Zafeiriou et al. utilizan la factorización no negativa de matrices (NMF: Non-negative matrix factorization) para conseguir que la representación distribuida de parámetros localizados representen las partes constituyentes de la cara en las imágenes faciales. La factorización no negativa de matrices es una técnica de reconocimiento facial basada en la apariencia que utiliza las técnicas convencionales de análisis de componentes.

6.2.2 Preparación de los experimentos

Para realizar los experimentos de fusión se ha creado una base de datos quimérica mediante la combinación de los registros de la base de datos de voz Switchboard-I (Campbell et al., 1999; Godfrey et al. 1990) y la base de datos de voz, caras y vídeo XM2VTS (Lüttin et al., 1998).

La base de datos Switchboard-I se ha utilizado para los experimentos de reconocimiento de locutor. Es un conjunto de 2430 fragmentos de conversaciones telefónicas entre 543 hablantes (302 hombres y 241 mujeres). Las puntuaciones de voz se han obtenido utilizando dos sistemas diferentes: un sistema de reconocimiento de locutor basado en el espectro de voz y un sistema de reconocimiento basado en prosodia. En ambos sistemas cada modelo de hablante se ha entrenado con 8 fragmentos de conversación y se ha testeado de acuerdo con la tarea NIST's 2001 Extended Data (NIST Speaker Recognition website).

El sistema de reconocimiento de locutor basado en el espectro de voz es un sistema basado en GMM (*Gaussian Mixture Model*) de 32 componentes con matriz de covarianza diagonal. Se han extraído 20 parámetros *Frequency Filtering* de tramas de 30 ms con un desplazamiento de 10 ms, y los 20 coeficientes correspondientes a delta y aceleración se han añadido al vector de parámetros. El UBM (*Universal Background Model*) se ha entrenado con 116 conversaciones.

En el sistema de reconocimiento basado en la prosodia se ha extraído un vector de 9 parámetros para cada secuencia de voz. Se ha calculado la media y la desviación estándar para cada parámetro individualmente. El sistema se ha testeado con 1 secuencia de voz, utilizando el método k-Nearest Neighbor con k=3 y la divergencia simetrizada de Kullback-Leibler.

Para los experimentos de reconocimiento facial se ha utilizado la base de datos XM2VTS. Las imágenes de caras (cuatro imágenes frontales para cada sujeto) de los 295 usuarios de la base de datos se han utilizado en estos experimentos.

Para evaluar los algoritmos de verificación facial en la base de datos se ha seguido el protocolo de Lausanne para la configuración II descrita en (Lüttin et al., 1998). El convencional criterio discriminante de Fisher se ha construido para descubrir las proyecciones lineales discriminantes y obtener las puntuaciones faciales (Lee et al., 2000).

Para los experimentos de fusión, se ha creado una base de datos quimérica combinando 179 usuarios de la base de datos Switchboard-I y 270 usuarios de la base de datos XM2VTS. Para el entrenamiento de los métodos de fusión, se ha construido un conjunto de vectores de puntuaciones multimodal mediante la asociación uno a uno de 930 puntuaciones de cada modalidad (336 clientes y 594 impostores). Para los experimentos de testeo, las puntuaciones de voz y caras se han combinado para obtener un total de 46500 experimentos (16800 clientes y 29700 impostores).

La base de datos Switchboard-I se ha elegido como proveedor de las señales de voz para estos experimentos porque contiene habla espontánea y, por lo tanto, permite extraer los parámetros prosódicos, cosa que no ocurre con las señales de voz de la base de datos XM2VTS.

Las puntuaciones unimodales se han fusionado por medio de estrategias de fusión de un paso, dos pasos y tres pasos que se explican en detalle en el apartado 6.2.3. La normalización de las puntuaciones se ha realizado por medio de las técnicas del estado del arte, *min-max* (MM), *z-score* (ZS), *tanh* (TANH) y por medio de la ecualización de histograma a una de las modalidades biométricas (HEQ). El proceso de fusión se ha realizado por medio de dos métodos de combinación lineal de puntuaciones, ponderación en función del resultado de reconocimiento de cada biometría unimodal o *matcher weighting* (MW) y un método de ponderación aprendido (LW: *learning weighting*) que obtiene los factores de pesado que minimizan el EER a partir de los datos de entrenamiento de la fusión ("fuerza bruta"), y máquinas de vector soporte (SVM) sin utilizar la normalización MM que suele aplicarse de forma previa en este método.

En el proceso de ecualización de histograma, se han utilizado como referencia los histogramas correspondientes a las puntuaciones de las diferentes modalidades biométricas. En este caso se ha aplicado una ecualización de histograma de 100 intervalos sobre las puntuaciones unimodales.

En la fusión mediante SVM se ha utilizado el kernel RBF (*radial basis function*), y se han utilizado las distancias de los datos al hiperplano óptimo como las puntuaciones finales de verificación. Para cada prueba de fusión con SVM se ha realizado un proceso de validación cruzada para determinar los parámetros del kernel. Los valores de σ utilizados en el proceso de entrenamiento son 1/2, 1/√2, 1, √2, 2, 2√2 y 4 y para el parámetro C se han utilizado los valores 1, 10, 100, y 200.

6.2.3 Resultados unimodales y bimodales

La tabla 6-8 muestra el EER obtenido por cada parámetro prosódico utilizado en el sistema de reconocimiento basado en la prosodia. Como se puede observar, las características basadas en medidas del tono obtienen los mejores resultados.

Parámetros	EER
log (#tramas/palabras)	30.3
Longitud promedio de los segmentos de voz internos en las palabras	31.5
Longitud promedio de los segmentos sin voz internos en las palabras	31.5
log (media F0)	19.2
log (máximo F0)	21.3
log (mínimo F0)	21.5
log (rango F0)	26.6
"pseudo pendiente" del tono	38.3
pendiente promedio sobre todos los segmentos de una estilización PWL de F0	28.7

Tabla 6-8: EER(%) para los parámetros prosódicos.

Los EER obtenidos por cada sistema de reconocimiento unimodal se muestran en la tabla 6-9. Cabe destacar que, en este caso, únicamente se utiliza fusión en el sistema prosódico, dado que hay 9 puntuaciones prosódicas a combinar. Para esta fusión, se han probado la combinación de tres técnicas de normalización con tres de fusión. Para el proceso de normalización, se utilizan las técnicas mencionadas más arriba: *z-score* (ZS), *tanh* (TANH) y ecualización de histograma (HEQ).

Origen	EER
Espectro de voz	9.52
Caras	2.50

Prosodia	MW	LW	SVM
ZS	15.66	14.88	14.82
TANH	16.83	14.65	13.97
HEQ – log (media F0)	15.48	15.48	13.80

Tabla 6-9: EER(%) para cada sistema unimodal.

La ecualización sobre las puntuaciones "log (media F0)" ha obtenido mejores resultados que la ecualización al resto de características por lo que sólo se presentan los resultados de esta ecualización. En el proceso de fusión las técnicas utilizadas son: ponderación en función del resultado de reconocimiento de cada biometría unimodal o *matcher weighting* (MW), un método de

ponderación aprendida (LW) que obtiene los pesos que minimizan el EER a partir de los datos de entrenamiento de la fusión ("fuerza bruta"), y una máquina de vector soporte (SVM). Estas son las técnicas que se utilizarán también para el resto de procesos de fusión presentadas en este apartado.

La tabla 6-10 muestra los resultados de fusión para dos sistemas bimodales: un sistema basado en prosodia fusionado con el sistema de reconocimiento basado en el espectro de voz y el mismo sistema basado en espectro de voz fusionado con el sistema de reconocimiento facial, utilizando los mismos métodos de normalización y fusión que en el caso anterior. En el sistema bimodal basado en voz, la prosodia (PS: *prosodic scores*) y las puntuaciones espectrales (SS: *speech spectral scores*) se fusionan en dos pasos según la figura 6-4.

En la fusión de prosodia y espectro de voz, la ecualización en el primer paso de fusión se ha realizado sobre las puntuaciones de "log (media F0)", y la segunda ecualización se ha realizado sobre las puntuaciones de espectro de voz y sobre las puntuaciones del sistema prosódico fusionado. En la fusión de caras y espectro de voz, se ha probado la ecualización sobre ambas modalidades biométricas.

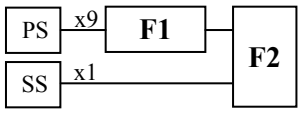


Figura 6-4: Fusión en dos pasos de los sistemas prosódico y de espectro de voz.

Prosodia + Espectro de voz	MW	LW	SVM
ZS	7.142	6.548	6.250
TANH	6.845	6.548	5.654
HEQ – espectro de voz	7.406	5.654	6.733
HEQ – prosodia	8.333	8.036	7.071

Espectro de voz + Caras	MW	LW	SVM
ZS	1.953	1.792	1.065
TANH	1.523	1.869	1.077
HEQ – caras	1.506	1.131	1.065
HEQ – espectro de voz	1.524	1.744	1.352

Tabla 6-10: EER(%) para los sistemas de reconocimiento bimodales.

En ambos procesos de fusión bimodal la ecualización sobre la biometría que tiene un mejor resultado de reconocimiento unimodal, el espectro de voz en el primer caso y las caras en el segundo, obtiene mejores resultados que la ecualización sobre la otra biometría.

Cuando se quieren fusionar las tres modalidades, se pueden aplicar diferentes estrategias en función de cómo se agrupan las puntuaciones para fusionarlas.

6.2.4 Estrategias y técnicas de fusión

1) Fusión en un paso

La fusión en un paso (Fig. 6-5) consiste en combinar de una sola vez todas las puntuaciones obtenidas a partir de las 11 características: puntuaciones relacionadas con 9 parámetros prosódicos (PS: *prosodic scores*), puntuaciones de espectro de voz (SS: *speech spectral scores*) y puntuaciones de caras (FS: *facial scores*). El EER para la fusión en un paso se muestra en la tabla 6-11.

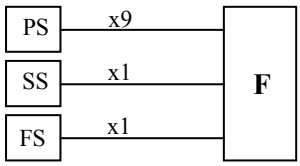


Figura 6-5: Fusión en un paso.

Fusión en un paso	MW	LW	SVM
ZS	1.417	1.441	1.077
TANH	1.690	2.596	1.030
HEQ – caras	1.011	1.172	0.888
HEQ – espectro de voz	1.517	1.662	1.381
HEQ – log (media F0)	5.269	6.107	6.013

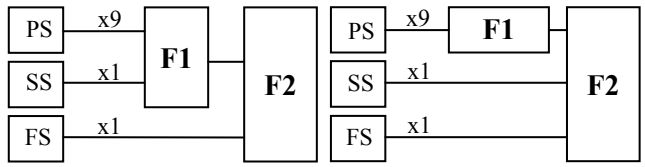
Tabla 6-11: EER(%) para fusión en un paso.

En este caso, la ecualización se ha realizado sobre los tres conjuntos de puntuaciones que obtienen los mejores resultados de reconocimiento unimodal, las caras, el espectro de voz y log(media F0). Una vez más, se confirma que el mejor resultado se obtiene cuando se ecualiza utilizando como referencia el histograma de la modalidad biométrica que obtiene los mejores resultados unimodales.

Además, en este caso, la ecualización de las puntuaciones a la biometría facial es decisiva para la clasificación de las 11 puntuaciones en la fusión en un solo paso con SVM, que obtiene una mejora relativa del 13.8% respecto al resto de técnicas.

2) Fusión en dos pasos

La fusión en dos pasos consiste en combinar todas las puntuaciones obtenidas de las 11 características en dos pasos consecutivos. En este tipo de fusión se han considerado dos estrategias diferentes según la figura 6-6. En la primera configuración (configuración A), las puntuaciones de todas las características de voz (9 parámetros prosódicos y 1 característica espectral) se fusionan en primer lugar y los resultados obtenidos se fusionan de nuevo con las puntuaciones faciales. En la segunda configuración (configuración B) las puntuaciones de las 9 características prosódicas se fusionan en primer lugar y los resultados obtenidos son posteriormente fusionados con las puntuaciones de espectro de voz y faciales.



(a) Configuración A (b) Configuración B

Figura 6-6: Dos configuraciones para la fusión en dos pasos.

La tabla 6-12 muestra los EER de las dos configuraciones propuestas para la fusión en dos pasos. Se han utilizado las mismas técnicas de normalización y fusión en ambos pasos. La ecualización de histograma se ha realizado en cada paso sobre el conjunto de puntuaciones con el mejor resultado de reconocimiento.

Para la mayor parte de las técnicas de fusión, la configuración B obtiene mejor resultado que la configuración A. Este hecho se podría explicar porque en la configuración A, el primer paso fusiona información no homogénea mientras que en la configuración B, el primer paso de fusión obtiene una puntuación global para la información prosódica, que está fuertemente correlada entre sí, y en el segundo paso de fusión se combinan los tres orígenes de información, que están incorrelados.

Fusión en dos pasos – A	MW	LW	SVM
ZS	1.994	1.262	0.761
TANH	1.476	1.013	0.785
HEQ	1.482	0.964	0.845

Fusión en dos pasos – B	MW	LW	SVM
ZS	1.673	1.357	0.874
TANH	1.073	0.983	0.703
HEQ	1.084	0.828	0.673

Tabla 6-12: EER(%) para fusión en dos pasos.

El mejor resultado ha sido obtenido por la combinación de ecualización de histograma con SVM en la configuración B. De hecho, este es el mejor resultado obtenido en las tres estrategias de fusión, con un 4.27% de mejora con respecto al resto de sistemas de fusión.

3) Fusión en tres pasos

En la configuración para la fusión en tres pasos, en primer lugar, se fusionan las puntuaciones prosódicas obtenidas a partir de las 9 características prosódicas. Los resultados obtenidos se combinan con las puntuaciones espectrales de voz obtenidas de los parámetros espectrales, y los nuevos resultados son, una vez más, fusionados con las puntuaciones faciales obtenidas de los parámetros de las imágenes faciales, tal y como se muestra en la figura 6-7.

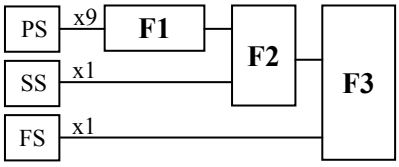


Figura 6-7: Fusión en tres pasos.

Los EER obtenidos para los métodos probados se muestran en la tabla 6-13. El método basado en SVM obtiene los mejores resultados para todas las técnicas de normalización con resultados similares para todas ellas.

Fusión en tres pasos	MW	LW	SVM
ZS	1.994	1.226	0.720
TANH	1.703	0.953	0.749
HEQ	1.638	0.959	0.737

Tabla 6-13: EER(%) para la fusión en tres pasos.

La comparación de los resultados de las diferentes estrategias de fusión muestra que el sistema bimodal basado en los sistemas de reconocimiento de espectro de voz y de caras se mejora con la introducción de los parámetros prosódicos. Además, las SVM mejoran a la fusión convencional *matcher weighting* y al método de ponderación aprendido. Por otro lado, una ecualización de histograma de las puntuaciones previa al proceso de fusión mejora los resultados obtenidos por las técnicas de clasificación basadas en SVM para la configuración de fusión que obtiene los mejores resultados, una fusión en dos pasos donde los 9 parámetros prosódicos se combinan en primer lugar, y la puntuación prosódica obtenida se fusiona después con las puntaciones espectrales de voz y las puntuaciones de caras (configuración B).

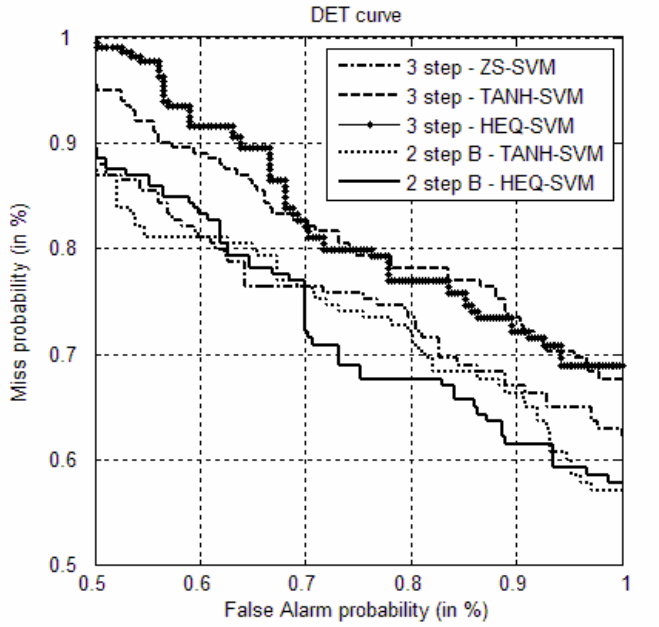


Figura 6-8: Curva DET para diferentes estrategias de fusión.

En la figura 6-7 se muestra la curva DET para las cinco técnicas que han obtenido mejores resultados de reconocimiento: configuración B de fusión en dos pasos utilizando HEQ-SVM y TANH-SVM, y fusión en tres pasos utilizando ZS-SVM, TANH-SVM y HEQ-SVM.

Los métodos de fusión en dos pasos y la fusión en tres pasos con ZS-SVM obtienen los mejores resultados para todo el rango de FAR y FRR. La configuración B para la fusión en dos pasos utilizando HEQ-SVM mejora al resto de técnicas para los valores alrededor del EER.

Como conclusión se puede resaltar que el rendimiento de un sistema bimodal basado en información facial y de espectro de voz mejora cuando se añade información prosódica al sistema. En los experimentos, el uso de ecualización de histograma y máquinas de vector soporte superan los mejores resultados obtenidos por el resto de combinaciones de técnicas de normalización como *z-score*, *tanh* y ecualización de histograma y técnicas de fusión como *matcher weighting*, un método de ponderación basado en el aprendizaje y SVM. La estrategia de fusión de las puntuaciones es relevante para el rendimiento del sistema. En concreto, los mejores resultados se obtienen por medio de una fusión en dos pasos de todas las puntuaciones con la técnica HEQ-SVM donde las 9 características prosódicas se fusionan en primer lugar. También se ha observado que la fusión de la información de voz (puntuaciones espectrales y prosódicas) en un único paso previo a la fusión con las puntuaciones de caras no contribuye a la mejora del sistema.

6.3 Normalización en fusión de espectro de voz, prosodia y caras.

De la misma forma que en el apartado 6.2, las puntuaciones unimodales utilizadas en estos experimentos has sido obtenidas por medio de tres expertos: la fusión mediante SVM de 9 características prosódicas de voz, un sistema de reconocimiento de locutor basado en espectro de voz y GMM y un sistema de reconocimiento facial basado en el algoritmo NMFFaces (Zafeiriou et al., 2005a).

6.3.1 Sistemas unimodales.

Los detalles de los sistemas unimodales utilizados en este caso son los mismos que en apartado 6.2 aunque aquí se hace un pequeño resumen de ellos.

En el sistema de reconocimiento basado en prosodia se ha extraído un vector de 9 características prosódicas para cada fragmento de conversación (Wolf, 1972). El sistema se ha testeado con un fragmento de conversación, utilizando el método k-Nearest Neighbor con k=3 y la divergencia simetrizada de Kullback-Leibler. Los vectores prosódicos se han fusionado mediante un sistema de clasificación con SVM y kernel RBF para obtener una única puntuación global.

El sistema de reconocimiento de locutor basado en espectro de voz está basado en GMM de 32 componentes con matriz de covarianzas diagonal. Se ha utilizado un vector de 20 parámetros Frequency Filtering (Nadeu et al., 1995) calculado cada 30 ms con un desplazamiento de 10 ms, y

se han añadido los correspondientes 20 parámetros delta y de aceleración. El UBM se ha entrenado con 116 conversaciones. Se han utilizado parámetros Frequency Filtering dado que obtienen resultados comparables o mejores que los coeficientes mel-cepstrum.

El sistema experto en reconocimiento facial se basa en el algoritmo NMFFaces. Los sistemas de reconocimiento facial se basan comúnmente en la descomposición de la cara en partes distribuidas de forma dispersa: ojos, nariz, boca, etc. Zafeiriou et al. utilizan la factorización no negativa de matrices (*NMF: Non-negative matrix factorization*) para conseguir que la representación distribuida de parámetros localizados representen las partes constituyentes de la cara en las imágenes faciales. La factorización no negativa de matrices es una técnica de reconocimiento facial basada en la apariencia que utiliza las técnicas convencionales de análisis de componentes.

6.3.2 Preparación de los experimentos.

Para realizar los experimentos se ha creado una base de datos multimodal quimérica utilizando muestras de voz de la base de datos Switchboard-I (Godfrey et al. 1990) e imágenes fijas de la base de datos XM2VTS (Lüttin et al., 1998). La base de datos Switchboard-I es un conjunto de 2430 conservaciones telefónicas conteniendo los fragmentos de cada uno de los dos hablantes y contiene información de 543 usuarios de Estados Unidos. La base de datos XM2VTS contiene imágenes de 295 personas. A partir de los expertos de reconocimiento de voz y caras, se disponía de 1860 experimentos de voz (con información prosódica y de espectro) y 33361 experimentos de reconocimiento facial. Las puntuaciones de todos los usuarios se han dividido en dos grupos, para el entrenamiento y las pruebas. Las puntuaciones de entrenamiento y pruebas de las dos biometrías se han combinado de forma independiente para obtener un total de 5000 vectores de puntuaciones para el entrenamiento de los modelos de fusión y 46500 vectores de puntuaciones que se han utilizado en la fase de pruebas.

Como ya se ha comentado en el apartado anterior, la base de datos Switchboard-I se ha elegido como proveedor de las señales de voz para estos experimentos porque contiene habla espontánea y, por lo tanto, permite extraer los parámetros prosódicos, cosa que no ocurre con las señales de voz de la base de datos XM2VTS.

En los experimentos multimodales, se han aplicado sobre las puntuaciones unimodales algunos de los métodos de normalización presentados en las secciones anteriores: *min-max* (MM), *z-score* utilizando la media y varianza del total de puntuaciones (ZS), *z-score* utilizando la media y varianza de las puntuaciones de impostores (ZS_i), una técnica basada en tanh (TANH), normalización de rango (RANK), un método de estimación de la densidad basado en la ventana de Parzen (DEST), ecualización de histograma al histograma de las puntuaciones de caras porque, en general, la mejor elección para el histograma de referencia es el de la modalidad biométrica con el

mejor resultado de reconocimiento unimodal (HEQ), ecualización a gaussiana (GEQ) y ecualización de doble gaussiana, tomando EER (BGEQ-EER) y HTER (BGEQ-HTER) para calcular la desviación estándar de las gaussianas que componen la distribución de referencia, o manteniendo las desviaciones estándar originales (BGEQ-STD).

Posteriormente, las puntuaciones normalizadas se han fusionado por medio de dos métodos combinatorios, una ponderación en función del resultado de reconocimiento de cada biometría unimodal o *matcher weighting* (MW) y un método aprendido de ponderación (LW: *learning weighting*), y un sistema de clasificación basado en SVM. Para MW cada puntuación unimodal se ha ponderado por la inversa del EER de la modalidad correspondiente. Para LW, los pesos se han calculado para minimizar el EER en la fase de entrenamiento.

6.3.3 Resultados unimodales.

En la tabla 6-14 se muestran la tasa de error equivalente (EER) y la mínima tasa de error total media (HTER) para los sistemas biométricos unimodales. Estos resultados se han obtenido a partir del conjunto de puntuaciones de test para cada modalidad biométrica de forma previa a la creación de la base de datos multimodal, y van a ser la base para la comparación con los resultados unimodales. El experto facial obtiene los mejores resultados con mucha diferencia respecto a los sistemas de locutor.

	EER (%)	HTER (%)
Caras	2.50	2.00
Espectro de voz	9.52	8.50
Prosodia	14.65	14.27

Tabla 6-14: Resultados unimodales.

6.3.4 Fusión mediante combinación aritmética de puntuaciones

El EER y el HTER obtenidos para cada técnica de normalización para los dos sistemas de fusión basados en la ponderación y suma de las puntaciones se muestran en la tabla 6-15. También se han incluido los intervalos de confianza al 90%. La ecualización de histograma ha obtenido los mejores resultados para los dos métodos de ponderación. Los resultados obtenidos con el método basado en el aprendizaje mejoran los obtenidos por MW utilizando HEQ y obtiene una diferencia mayor del 8% respecto al resto de técnicas.

	MW		LW	
	EER	HTER	EER	HTER
MM	1.363 ± 0.056	1.342 ± 0.047	1.090 ± 0.283	1.013 ± 0.279
ZS	1.125 ± 0.145	1.077 ± 0.146	1.101 ± 0.279	1.018 ± 0.281
ZS_i	1.701 ± 0.047	1.598 ± 0.061	1.101 ± 0.286	1.013 ± 0.277
TANH	0.875 ± 0.119	0.842 ± 0.131	0.869 ± 0.155	0.862 ± 0.175
RANK	0.927 ± 0.088	0.820 ± 0.107	0.899 ± 0.119	0.822 ± 0.110
DEST	0.899 ± 0.095	0.833 ± 0.096	0.862 ± 0.123	0.811 ± 0.137
HEQ	**0.815 ± 0.094**	**0.793 ± 0.095**	**0.725 ± 0.097**	**0.680 ± 0.086**
GEQ	0.923 ± 0.046	0.868 ± 0.049	0.994 ± 0.178	0.863 ± 0.174
BGEQ-EER	0.987 ± 0.193	0.936 ± 0.209	1.060 ± 0.230	0.969 ± 0.243
BGEQ-HTER	0.929 ± 0.052	0.913 ± 0.066	0.858 ± 0.161	0.813 ±0.163
BGEQ-STD	0.970 ± 0.057	0.937 ± 0.049	0.791 ± 0.118	0.761 ± 0.123

Tabla 6-15: Resultados multimodales para sumas ponderadas.

	MW	LW
MM	192.563	110.702
ZS	111.846	115.118
ZS_i	325.729	115.118
TANH	36.068	35.415
RANK	55.442	46.043
DEST	40.252	28.350
HEQ	9.888	-
GEQ	45.500	83.114
BGEQ-EER	37.283	30.008
BGEQ-HTER	48.737	11.688
BGEQ-STD	63.657	97.663

Tabla 6-16: Resultados del test de McNemar para fusión mediante sumas ponderadas.

Los intervalos de confianza de las diferentes técnicas se solapan, por lo que no se podría concluir con un 90% de confianza que los resultados obtenidos con HEQ-LW superan a los obtenidos mediante el resto de técnicas. Sin embargo, el test de McNemar refleja que, para los métodos de fusión basados en la ponderación, este método supera al resto con un margen superior al 95%, dado que se supera el valor 3.841 al aplicar la ecuación (3.6) sobre los errores de HEQ-LW y del resto de técnicas en el punto de trabajo del EER. Los resultados obtenidos se muestran en la tabla 6-16.

6.3.5 Fusión mediante máquinas de vector soporte

Para comparar el efecto de cada método de normalización sobre el sistema de fusión con SVM, se han probado diferentes configuraciones de dicha SVM, con kernel RBF y polinomial. En concreto para el kernel RBF,

$$k_{RBF}\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{6.8}$$

se han probado diferentes valores de la desviación estándar de la gaussiana σ: 1/2, $\sqrt{1/2}$, 1, $\sqrt{2}$ y 2. Para el kernel polinomial,

$$k_{poly}\left(x_i, x_j\right) = \left(\left\langle x_i, x_j \right\rangle + 1\right)^{\alpha} \tag{6.9}$$

se han utilizado valores entre 1 y 4 para el parámetro α. Además, para el parámetro regulador C, que controla el compromiso entre permitir errores de entrenamiento y forzar márgenes rígidos, se han utilizado los valores 1, 10, 100 y 200. El número de intervalos para los procesos de ecualización se ha establecido en 1000. Los parámetros óptimos para cada fusión se han establecido mediante una validación cruzada sobre los parámetros de entrenamiento (Devijver et al., 1982).

Los EER y HTER obtenidos por el sistema de verificación con SVM para cada kernel y técnica de normalización se presentan en la tabla 6-17. Las desviaciones sobre estos resultados que son necesarias para calcular los intervalos de confianza al 90% se han incluido también en esta tabla.

Las técnicas HEQ y BGEQ obtienen las menores medidas de error para ambos kernels. Tomando BGEQ-EER como técnica de referencia, dado que es la que ha obtenido menor EER, las diferencias relativas son mayor del 20% con respecto a MM y del 7% con respecto a ZS y ZS_i, las técnicas más convencionales de normalización. Los resultados obtenidos por TANH, RANK y DEST son más de un 1.8% mayores que los obtenidos por BGEQ-EER. En este caso los intervalos de confianza están considerablemente solapados y no se puede obtener una conclusión clara a partir

de los resultados. Las tasas de error obtenidas mediante GEQ se ven reducidas en más de un 8% por BGEQ-EER.

	RBF		Polinomial	
	EER	HTER	EER	HTER
MM	1.005 ± 0.150	0.948 ± 0.151	0.869 ± 0.137	0.827 ± 0.127
ZS	0.852 ± 0.362	0.686 ± 0.248	0.959 ± 0.148	0.957 ± 0.141
ZS_i	0.815 ± 0.232	0.690 ± 0.112	0.923 ± 0.307	0.918 ± 0.324
TANH	0.714 ± 0.118	0.671 ± 0.085	0.703 ± 0.146	0.657 ± 0.103
RANK	0.739 ± 0.120	0.662 ± 0.098	0.697 ± 0.128	0.627 ± 0.107
DEST	0.744 ± 0.134	0.683 ± 0.096	0.750 ± 0.110	0.676 ± 0.101
HEQ	**0.690 ± 0.107**	0.652 ± 0.101	0.708 ± 0.121	0.647 ± 0.051
GEQ	0.953 ± 0.152	0.832 ± 0.167	0.785 ± 0.102	0.677 ± 0.087
BGEQ-EER	0.701 ± 0.140	0.636 ± 0.107	**0.667 ± 0.115**	0.612 ± 0.094
BGEQ-HTER	**0.690 ± 0.141**	0.650 ± 0.142	0.684 ± 0.145	**0.591 ± 0.120**
BGEQ-STD	**0.690 ± 0.137**	**0.629 ± 0.108**	0.727 ± 0.151	0.679 ± 0.111

Tabla 6-17: Resultados multimodales mediante SVM para los kernel RBF y polinomial (EER y HTER en %).

	RBF	Polinomial
MM	100.979	52.738
ZS	135.206	6127.284
ZS_i	25.268	4518.383
TANH	5.011	2.943
RANK	11.112	2.641
DEST	10.937	12.779
HEQ	1.124	4.32
GEQ	93.176	24.923
BGEQ-EER	4.167	-
BGEQ-HTER	2.128	0.98
BGEQ-STD	0.991	7.01

Tabla 6-18: Resultados del test de McNemar para fusión mediante SVM para los kernel RBF y polinomial.

Los resultados de comparar BGEQ-EER con kernel polinomial con el resto de técnicas según el test de McNemar se muestran en la tabla 6-18. De estos resultados se puede concluir que, con un 95% de fiabilidad, BGEQ-EER con kernel polinomial supera al resto de técnicas convencionales exceptuando TANH y RANK, también con kernel polinomial, que serían superadas con una confianza de más del 89% según la distribución χ^2 aplicada a los resultados del test de McNemar.

El sistema basado en SVM mejora los resultados obtenidos por los sistemas basados en sumas ponderadas para todas las técnicas de normalización. Los mejores resultados globales los obtienen las técnicas BGEQ-EER y BGEQ-HTER con kernel polinomial, y las diferencias van del 0.87% al 9.08% con respecto a las mismas técnicas utilizando el kernel RBF y mayor de un 8% respecto a los mejores resultados obtenidos por los sistemas de suma ponderada.

En la figura 6, se muestra la curva DET para las diferentes técnicas de normalización utilizando SVM con kernel polinomial. Se han incluido las curvas para las normalizaciones MM, TANH, RANK, HEQ y BGEQ-EER. Se puede observar que la normalización MM se ve mejorada por el resto de las técnicas mientras que TANH y RANK son siempre superadas por BGEQ-EER o HEQ, que obtienen los mejores resultados de reconocimiento para todos los valores de tasa de falsa aceptación (*FAR: false acceptance rate*) y tasa de falso rechazo (*FRR: false rejection rate*).

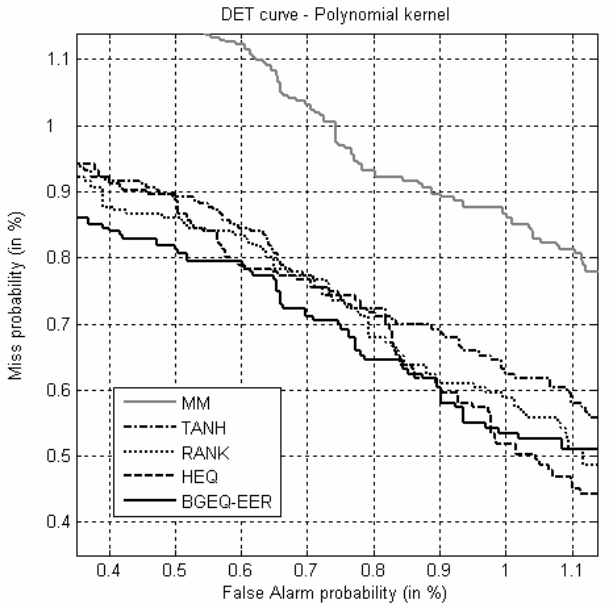


Figura 6-9: Curva DET para fusión mediante SVM con kernel polinomial.

Como conclusión de este apartado cabe destacar que, en experimentos multimodales de verificación de personas donde se han fusionado puntuaciones de prosodia, espectro de voz y caras, la introducción de técnicas de ecualización de histograma han reducido las tasas de error

conseguidas por el resto de las técnicas de normalización utilizadas, tanto en la aproximación combinatoria como en la clasificatoria. Además, la utilización de máquinas de vector soporte ha mejorado los resultados obtenidos por los métodos de suma ponderada de puntuaciones.

En concreto, la ecualización a doble gaussiana, una nueva técnica propuesta en esta tesis, y que tiene en cuenta la distribución separada de clientes e impostores, ha obtenido los mejores resultados en combinación con un sistema de fusión basado en SVM con kernel polinomial. De hecho, las técnicas de ecualización han superado a las técnicas convencionales de normalización para todos los valores de FAR y FRR tanto para el kernel RBF como para el polinomial.

Finalmente, se puede concluir que, en un sistema biométrico multimodal de fusión de puntuaciones la selección de la técnica de normalización es un tema fundamental y puede resultar tan decisivo como la selección del método de fusión, en las dos aproximaciones de fusión, la combinatoria y la clasificatoria.

7 Fusión multinivel en el entorno del proyecto Agatha

El objetivo principal del proyecto Agatha: "Plataforma de identificación de personas en bases de datos mediante biometría multimodal" consiste en la implementación de un sistema de identificación de usuarios mediante el reconocimiento de voz, de cara o la combinación de ambos. El entorno de dicho sistema será generalmente criminalístico o de seguridad. Dicho proyecto pretende dar solución al reconocimiento de un individuo de entre los usuarios que componen una base de datos, normalmente muy grande, permitiendo que dicho reconocimiento se realice a partir de características biométricas individuales o con una combinación de las mismas. Para ello el sistema, dado un individuo, devolverá los N usuarios más probables ordenados según probabilidad. Dicha identificación no es necesario que se realice en tiempo real aunque sí que la respuesta se produzca en un tiempo razonable en relación a la tecnología disponible.

Por lo que respecta a la voz, la identificación de locutores comprende una tarea compleja y que no se suele realizar en tiempos razonables con grandes bases de datos. En cuanto al reconocimiento facial, sí existen productos comerciales que realizan el reconocimiento prácticamente en tiempo real, ahora bien, con unas tasas de fiabilidad que dependen mucho de las condiciones de adquisición de las imágenes.

La motivación de este proyecto es disponer de la identificación de locutores, un campo donde las posibilidades de investigación son muy amplias y donde aún queda mucho terreno por recorrer; y de la identificación de caras, en una sola herramienta. De esta forma se podría realizar la fusión de las dos biometrías en el caso de que tuviésemos ambos datos del usuario.

Dada la velocidad con la que se puede realizar el reconocimiento facial éste podría complementar perfectamente al sistema de reconocimiento de voz. El sistema de reconocimiento facial podría, en un primer reconocimiento, devolver los N candidatos más probables para que posteriormente el sistema de reconocimiento de voz sólo tuviera que realizar el reconocimiento entre estos N candidatos. Por lo tanto la combinación de dichas biometrías no sólo se ciñe a aspectos

relacionados con la mejora de las tasas de reconocimiento, por el hecho de combinar diferentes fuentes de información, sino también el permitir reconocimientos a altas velocidades en sistemas que hasta la fecha no podían implementar dichos tiempos de respuesta, sistemas de reconocimiento de locutor en grandes bases de datos.

Como conclusión destacar que el atractivo de la solución presentada es que reúne en un único sistema ambos tipos de reconocimiento, al mismo tiempo que permite utilizar individualmente las biometrías de voz y caras para la identificación de personas.

7.1 Líneas de investigación del proyecto Agatha

Las líneas de investigación en las que se centra el proyecto Agatha son las siguientes:

Identificación de locutores

La identificación de locutores consiste en determinar a quién pertenece un fragmento de voz de entre los modelos de locutores existentes en una base de datos.

En esta línea de investigación se incluye la detección de voz / silencio, la búsqueda de la mejor parametrización para el reconocimiento de voz independiente del texto y la implementación de modelos estadísticos del locutor.

Resultará determinante la elección de la tecnología de modelo de patrones. Entre las tecnologías que se estudiarán se encuentran los *Gaussian Mixture Models* (GMM), las *Support Vector Machines* (SVM), los *Nearest Neighbours* (NN) y la Cuantificación Vectorial (VQ: *Vector Quantization*).

Identificación de caras

La identificación de caras consiste en determinar a quién pertenece una fotografía de entre los modelos de caras existentes en una base de datos.

En esta línea de investigación se incluye la detección de cara, la búsqueda de parámetros óptimos para combatir problemas de iluminación, brillo e intensidad y la implementación de modelos estadísticos de caras.

Del mismo modo que en la línea de investigación anterior, resultará determinante la elección de la tecnología de modelado de patrones. El estudio se centrará fundamentalmente en la tecnología de Características Locales (LF: *Local Features*) y los *Nearest Neighbours* (NN).

Algorítmica de bases de datos

Esta línea se basa en la determinación de algoritmos que sean capaces de realizar búsquedas cercanas al tiempo real en grandes de bases de datos.

Para ello es crucial haber trabajado en colaboración con las dos primeras líneas de investigación pues su diseño influye decisivamente en qué tipo de algoritmos se podrán utilizar y en las prestaciones finales del sistema.

Biometría multimodal.

La biometría multimodal consiste en utilizar dos o más tecnologías biométricas para reducir el error individual que tendría cada una de ellas si se utilizasen por separado.

En este caso, hay que decidir a qué nivel se aplica la fusión multimodal e investigar cómo afecta a la tasa de igual error del sistema (EER).

Los participantes en el proyecto Agatha han sido el ITI (Instituto Tecnológico de Informática), Biometric Technologies, S.L. y la UPC (Universitat Politècnica de Catalunya) que ha liderado la actividad de fusión multimodal de biometrías.

En esta tesis, se presentan los experimentos de fusión multimodal correspondientes a la participación de la UPC en el proyecto Agatha liderando la actividad de fusión multimodal de biometrías.

7.2 Fusión de puntuaciones de voz y caras en una base de datos quimérica.

En esta fase del proyecto hemos utilizado diferentes métodos en la tarea de fusión a nivel de puntuaciones de sistemas de reconocimiento de voz y caras.

7.2.1 Preparación de los experimentos.

Para realizar la fusión se ha seguido la configuración I del protocolo Laussane para la base de datos XM2VTS (Lüttin et al., 1998), que define un conjunto de datos de desarrollo para entrenar las técnicas de fusión y un conjunto de evaluación con el que se obtienen los resultados para comparar las diferentes técnicas.

Las puntuaciones de caras han sido proporcionadas por el ITI y han sido obtenidas mediante los algoritmos *eigenfaces* y *fisherfaces*, técnicas que se explicarán en más detalle en el apartado 7.3. Se han utilizado las puntuaciones de voz disponibles en la base de datos de fusión para autenticación biométrica que se encuentra en

http://personal.ee.surrey.ac.uk/Personal/Norman.Poh/web/fusion/main.php (Poh et al., 2004). Los conjuntos de puntuaciones utilizados son (PAC, GMM) y (SSC, GMM).

7.2.2 Resultados unimodales.

Los resultados mostrados para esta tarea, en todos los casos, corresponden con la tasa de igual error (*EER: Equal Error Rate*) y la mínima tasa de error total media (*HTER: Half Total Error Rate*) calculados sobre los experimentos de evaluación.

Los resultados obtenidos por las técnicas unimodales utilizadas para los experimentos de fusión son los siguientes:

	EER (%)	HTER (%)
Voz		
(PAC, GMM)	6,500	6,144
(SSC, GMM)	4,500	4,298
Caras		
Eigenfaces	3,000	2,790
Fisherfaces	1,500	1,440

Tabla 7-1: Resultados unimodales de voz y caras.

7.2.3 Resultados multimodales.

Existen dos aproximaciones principales para la fusión de biometrías a nivel de puntaciones: la fusión mediante combinación aritmética o lógica de las puntaciones y la fusión mediante clasificadores más complejos. En ambos casos, de manera previa a la fusión propiamente dicha, se debe realizar una normalización de los conjuntos de puntaciones para que sus rangos de valores sean comparables.

En primer lugar, se han aplicado técnicas de fusión basadas en la combinación aritmética de las puntaciones sobre los 4 grupos de puntuaciones unimodales. Para comparar los resultados obtenidos por las técnicas desarrolladas en la UPC se han utilizado diversas técnicas del estado del arte. En concreto, para la normalización se han utilizado *min-max* (MM), *z-score* (ZS) y una técnica basada en la tangente hiperbólica (TANH). Las técnicas desarrolladas por la UPC utilizadas para la normalización son la normalización conjunta de medias (JMN) y la normalización para la minimización de la suma de varianzas (MVSW: *Minimum Variance Sum Weighting*). Para la fusión

se han utilizado la suma directa o simple (SS) y la ponderación en función del resultado o *matcher weighting* (MW). Los resultados obtenidos con estas técnicas son los siguientes:

	SS		MW	
	EER (%)	HTER (%)	EER (%)	HTER (%)
MM	0,512	0,439	0,558	0,451
ZS	0,500	0,438	0,750	0,644
TANH	1,000	0,822	0,500	0,412
JMN	0,750	0,573	0,500	0,404
MVSW	0,539	0,425	-	-

Tabla 7-2: Resultados multimodales con técnicas afines de normalización.

Si comparamos los resultados obtenidos por las técnicas desarrolladas en la UPC con los obtenidos mediante las técnicas del estado del arte, podemos observar que la combinación JMN-MW mejora los resultados obtenidos por el resto de técnicas mientras que la combinación JMN-MVSW obtiene resultados comparables con los del estado del arte.

Otra de las opciones que se ha planteado para la realización de la tarea de fusión de biometrías es la utilización de técnicas de normalización convencionales en el ámbito de las biometrías unimodales para su aplicación a la fusión multimodal. Este es el caso de la ecualización de histograma (*HEQ: Histogram Equalization*), que ha sido ampliamente utilizada, entre otras aplicaciones, para el tratamiento de imágenes y para corregir el efecto de transformaciones no lineales sobre la señal de voz. La ecualización de histograma realiza un mapeo no lineal y monótonamente creciente entre el histograma de una señal y el histograma de una señal de referencia, de manera que se reducen las diferencias estadísticas entre las dos señales. En los resultados que se presentarán se ha realizado la ecualización de histograma utilizando como referencia el histograma de la biometría que ofrece el mejor resultado de reconocimiento, en este caso, el algoritmo *fisherfaces* de reconocimiento de caras.

Además, se presentan los resultados obtenidos mediante la ecualización de doble gaussiana (*BGEQ: Bi-Gaussian Equalization*), una técnica de ecualización de histograma en que se tienen en cuenta las estadísticas separadas de clientes e impostores. Para hacerlo, se han definido de forma independiente las distribuciones de clientes e impostores de la distribución de referencia como dos gaussianas. La gaussiana de la distribución de clientes se ha centrado en 1 mientras que la gaussiana de la distribución de impostores se ha centrado en -1. Las varianzas de las dos distribuciones se han definido iguales y con un valor tal que la distribución global tenga el mismo

EER que la distribución unimodal original. De esta manera, cada conjunto de puntuaciones unimodal se debe ecualizar sobre una distribución de la forma:

$$f_{ref}(x) = \frac{1}{2\sigma\sqrt{2\pi}}\left[e^{-\frac{(x+1)^2}{2\sigma^2}} + e^{-\frac{(x-1)^2}{2\sigma^2}} \right] \tag{7.1}$$

En este caso, la distribución de referencia para la ecualización es diferente para cada biometría ya que la varianza se calcula en función del EER de cada biometría.

Los resultados obtenidos por HEQ y BGEQ con las técnicas de fusión lineales se presentan en la tabla 7-3. Estos resultados no mejoran los obtenidos por las técnicas del estado del arte aunque obtienen resultados comparables a los obtenidos por el resto de técnicas.

	SS		MW	
	EER (%)	HTER (%)	EER (%)	HTER (%)
HEQ	1,000	0,810	0,553	0,553
BGEQ	0,750	0,599	0,559	0,463

Tabla 7-3: Resultados multimodales con HEQ y BGEQ.

Hasta el momento, se han presentado resultados obtenidos mediante técnicas de fusión que combinan aritméticamente las puntuaciones de las biometrías unimodales. Sin embargo, la utilización de clasificadores permite también la fusión a nivel de puntuaciones definiendo, en el caso de la verificación, dos clases: clientes e impostores.

Las máquinas de vector soporte (*SVM: Support Vector Machines*) (Cristianini et al., 2000) son un tipo de clasificadores que forman ya parte del estado del arte para realizar fusión multimodal. Las SVM permiten definir un hiperplano que separa los datos en dos clases. Además, mediante la utilización de kernels se puede realizar la definición del hiperplano en un espacio transformado donde la separación de los datos entre clientes e impostores se puede realizar de forma más eficiente. En este tipo de sistemas, además, se puede definir un parámetro *C* que establece el compromiso entre los datos que quedan mal clasificados y la rigidez de los márgenes de separación.

Los kernels más utilizados son el RBF (*radial basis function*) y el polinomial, que se formulan de la siguiente manera.

$$k_{RBF}\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{7.2}$$

$$k_{poly}\left(x_i, x_j\right) = \left(\left\langle x_i, x_j \right\rangle + 1\right)^{\alpha} \tag{7.3}$$

La tabla 7-4 muestra los resultados obtenidos por las diferentes técnicas de normalización en una fusión mediante SVM con kernel RBF y polinomial. Para la obtención de los resultados se ha realizado una validación cruzada (*cross-validation*) de los resultados de entrenamiento con valores ½, 1, 2 y 4 para σ y con valores 1, 10, 100 y 200 para C en el caso del kernel RBF y con valores 1, 2, 3 y 4 para α y con valores 1, 10, 100 y 200 para C en el caso del kernel polinomial.

Para el kernel RBF, la ecualización de doble gaussiana obtiene los mejores resultados y supera, también, los resultados obtenidos por las técnicas lineales de fusión. Sin embargo, los mejores resultados son obtenidos por BGEQ con kernel polinomial, que consigue una mejora relativa de al menos un 36% en el EER y un 12% en el HTER sobre todos los resultados obtenidos con las técnicas convencionales. La normalización JMN obtiene los segundos mejores resultados con el kernel polinomial mientras que HEQ obtiene resultados equiparables a los obtenidos por las técnicas convencionales.

	SVM – RBF		SVM – Polinomial	
	EER (%)	HTER (%)	EER (%)	HTER (%)
MM	0,500	0,396	0,500	0,376
ZS	0,750	0,557	0,468	0,343
TANH	0,488	0,363	0,496	0,371
JMN	0,552	0,433	0,466	0,341
HEQ	0,502	0,416	0,500	0,386
BGEQ	0,453	0,328	0,298	0,298

Tabla 7-4: Resultados multimodales con fusión SVM.

Si se compara el resultado obtenido mediante BGEQ con kernel polinomial con el mejor de los resultados unimodales, la mejora relativa es del 80.10% para el EER y del 79.27% para el HTER, demostrando de esta manera la mejora proporcionada por el proceso de fusión.

7.3 Fusión multinivel de voz y caras en la base de datos XM2VTS.

7.3.1 Preparación de los experimentos

En esta fase del proyecto hemos utilizado diferentes métodos en las tareas de fusión a nivel de parámetros, a nivel de puntuaciones y a nivel de decisión de sistemas de reconocimiento de voz y caras. Para realizar la fusión se ha seguido la configuración I del protocolo Laussane para la base de datos XM2VTS (Lüttin et al., 1998), que define un conjunto de datos de entrenamiento, para crear los modelos de los usuarios, un conjunto de desarrollo que se puede utilizar para entrenar las técnicas de fusión de puntuaciones y un conjunto de evaluación con el que se obtienen los resultados para comparar las diferentes técnicas.

Los parámetros de caras han sido proporcionados por el ITI y han sido obtenidos mediante los algoritmos *eigenfaces* y *fisherfaces*. Los parámetros de voz han sido proporcionados por Biometric y consisten en supervectores GMM, *eigenvoices* y *fishervoices*, tal y como se describe en el apartado de reconocimiento de locutor.

Para poder realizar en paralelo las diversas tareas del proyecto, se ha decidido obtener los resultados necesarios para realizar las fusiones a nivel de puntuaciones y a nivel de decisión a partir de los anteriores parámetros.

7.3.1.1 *Reconocimiento de imagen mediante eigenfaces y fisherfaces*

Los métodos de reconocimiento facial, por ejemplo el clasificador de vecino más cercano en el espacio de la imagen, son computacionalmente costosos y requieren gran cantidad de espacio de almacenamiento. Por este motivo, es natural proponer esquemas que permitan reducir la dimensionalidad del espacio de parámetros (Belhumeur et al., 1997).

Una técnica utilizada habitualmente para reducir la dimensionalidad en el reconocimiento facial es el análisis de componentes principales (PCA: principal components análisis), que selecciona una proyección lineal de menor dimensión que maximiza la dispersión de las muestras proyectadas. Los parámetros obtenidos por la aplicación de la técnica PCA a la imagen se denominan ***eigenfaces***.

Por otro lado, bajo ciertas condiciones óptimas, la variación intraclase se encuentra en un subespacio lineal de espacio imagen. Se puede realizar una reducción de la dimensionalidad utilizando una proyección lineal y manteniendo todavía la separabilidad lineal. Por medio de una transformación de discriminante lineal (Jonsson et al., 2002; Belhumeur et al., 1997) se puede conseguir una reducción lineal y los parámetros resultantes se denominan ***fisherfaces***.

Para la clasificación de los vectores de parámetros, los sistemas de reconocimiento de caras utilizan habitualmente técnicas basadas en la distancia euclídea (Turk et al., 1991). Recientemente, se han

introducido las máquinas de vector soporte (SVM: Support Vector Machines) para realizar la tarea de clasificación en sistemas de reconocimiento facial con resultados satisfactorios (Jonsson et al., 2002).

En el primer caso, la distancia euclídea se calcula entre los vectores de pruebas y el vector modelo para cada ocurrencia. En el segundo caso, la técnica basada en SVM realiza la clasificación mediante un hiperplano separador entrenado mediante aprendizaje (Cristianini et al., 2000; Burges, 1998). Se consiguen límites no lineales utilizando una función específica denominada kernel que transporta los datos del espacio de entrada en un espacio de mayor dimensión. En estos resultados, únicamente se han utilizado SVM lineales.

En diversos trabajos, entre ellos (Ross et al., 2006), se ha demostrado la importancia de la normalización de los datos previamente al proceso de fusión mediante SVM. Para la normalización de los parámetros faciales se ha utilizado un método de normalización estándar, que normaliza la media y la varianza de los parámetros de la biometría unimodal. Los parámetros normalizados x_{SN} se calculan como

$$x_{SN} = \frac{a - mean(a)}{std(a)} \tag{7.4}$$

donde *mean(a)* es la media estadística de conjunto de parámetros *a* y *std(a)* es su desviación estándar.

La ecualización de histograma (HEQ) iguala la función de distribución acumulada de un cierto conjunto de datos a una distribución de referencia. Esta técnica se puede ver como una extensión de la normalización estadística realizada por la técnica anterior a toda la estadística de una modalidad biométrica y no sólo a su media y su varianza. Para la normalización de los parámetros faciales se ha utilizado la ecualización a gaussiana (GEQ: Gaussian Equalization), donde la referencia es una distribución normal, en este caso con varianza unitaria.

7.3.1.2 *Reconocimiento de locutor con supervectores y Support Vector Machines (SVM)*

Los sistemas de verificación de locutor realizan tres tareas principales: una extracción de características, un reconocimiento de patrones y una decisión final o clasificación de la señal de entrada. En este caso se utilizan parámetros mel-cepstrum (MFCC) y se modela estadísticamente la forma de hablar de cada locutor.

Los Modelos Ocultos de Markov, o HMM de sus siglas en inglés, son una herramienta potente que nos permite caracterizar la voz humana. Un HMM consiste en una serie de estados, interconectados entre sí, y a los que se puede llegar desde otro estado con una determinada probabilidad. Además, y de ahí el significado de Oculto, cada estado tiene asociada una determinada distribución de

probabilidad gaussiana. Dicha función de probabilidad modela cómo se distribuyen los vectores de características asociados a un determinado estado.

El uso de distribuciones gaussianas se justifica con el Teorema Central del Límite. Según éste, cualquier distribución de probabilidad se puede conseguir mediante una combinación ponderada de múltiples distribuciones, cada una de ellas con una media y una varianza determinada, como se puede ver en la expresión siguiente:

$$g(x) = \sum_{i=1}^{N} \lambda_i N(x; \mu_i, \Sigma_i) \tag{7.5}$$

donde λ_i son los pesos que ponderan cada mezcla, *N(.)* es una distribución de probabilidad gaussiana, μ_i y Σ_i son la media y la matriz de covarianza de las gaussianas, y N es el número total de componentes. Estos parámetros se relacionan de la forma siguiente:

$$N(x; m_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} e^{-(1/2)(\vec{x}-\vec{\mu}_i)^t \Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)} \tag{7.6}$$

Habitualmente se utilizarán matrices de covarianza diagonales. Esto se hace por tres razones. En primer lugar, el modelado de un GMM con una matriz de covarianza completa se puede conseguir igualmente con una matriz de covarianza diagonal de orden superior. En segundo lugar, las matrices diagonales son computacionalmente más eficientes. Y en tercer lugar, se ha observado empíricamente que los GMMs de matrices de covarianza diagonales superan a los GMMs de matrices completas.

Para aplicaciones de reconocimiento de locutor con texto independiente, la topología más común es un HMM con un único estado. Por eso también se le conoce como Gaussian Mixture Model (GMM).

Se puede obtener un GMM para cada realización de voz adaptando un modelo general, también conocido como Universal Background Model (UBM). Un UBM se ha de entrenar con voz suficientemente representativa del conjunto de hablantes que se desea reconocer (mismo idioma o dialecto, mismo canal, etc.), idealmente sin contener voz de los propios usuarios de la aplicación. Las ventajas de esta técnica son, por un lado, que los parámetros se pueden estimar con una cantidad relativamente pequeña de datos de entrenamiento y, por el otro, que se consigue minimizar la variabilidad no dependiente del hablante.

Para el caso del presente estudio, los GMMs se obtienen adaptando los vectores de medias del GMM global usando el criterio conocido como *Maximum A Posteriori* (MAP) (Reynolds et al., 2000; Thyes et al., 2000).

Este algoritmo es conocido como el algoritmo de *Expectation Maximization* (EM). En primer lugar, se calculan las probabilidades de ocupación de cada gaussiana del GMM a partir de la voz que se

usa para la adaptación. Los pesos de las mezclas y las matrices de covarianza se conservan por simplicidad y para una mayor robustez en la estimación de los parámetros. Finalmente, y a partir del modelo adaptado, se forma el supervector GMM con las medias del GMM. La figura 7-1 muestra como se construye un supervector GMM.

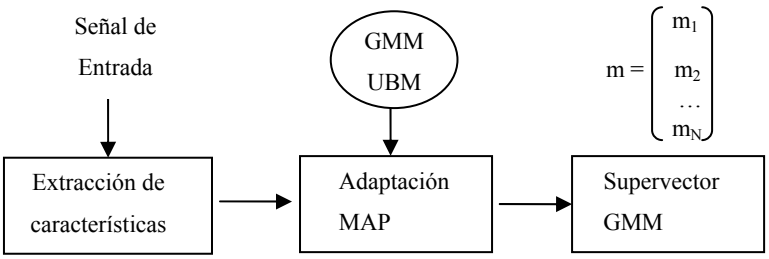


Figura 7-1: Producción de un supervector GMM.

Cuando se utiliza el supervector GMM como vector de parámetros de voz, la dimensión del patrón de voz es fija y no depende de la duración de la ocurrencia. Esta característica permite el uso de estos parámetros como técnicas de clasificación como las SVM (Campbell et al., 2006) y es una ventaja para incluir la información de voz en la fusión multimodal a nivel de parámetros.

Para el reconocimiento de locutor se construye un modelo general de voz o UBM con los datos de voz de la base de datos BANCA. Cada señal de entrada se preprocesa con Detector de Actividad Vocal o VAD de sus siglas en inglés. De esta forma se descartan los segmentos que no contienen voz, es decir, aquellos correspondientes a silencio, ruidos, etc. Después, las tramas de voz son preenfatizadas y procesadas mediante ventanas de Hamming de 25 ms con un desplazamiento temporal de 10 ms. El conjunto de características está formado por 12 coeficientes cepstrales más el logaritmo de la energía normalizada. También se aplica la Substracción Cepstral de la Media o CMS de sus siglas en inglés. Finalmente, la voz de 208 locutores grabados a lo largo de 12 sesiones se utiliza para entrenar un GMM UBM con 64 mezclas gaussianas.

Por otro lado, tanto para crear los modelos de clientes como para simular los ataques verdaderos y de impostores se utiliza la base de datos XM2VTS, con un preprocesado de las señales como el descrito en el párrafo anterior. De acuerdo con el Protocolo de Lausana I (LP I) (Lüttin et al., 1998), para los supervectores GMM de los clientes, que posteriormente se utilizarán para entrenar sus modelos mediante SVM, se utilizan 3 señales para adaptar el modelo UBM mediante la técnica MAP. Con todo esto obtenemos un supervector GMM de 832 coeficientes.

Para evaluar el sistema que se está proponiendo se van a utilizar tres técnicas, distancia euclídea, SVM y SVM con una normalización GEQ previa. Los resultados para los supervectores GMM son los mostrados en la tabla 7-5.

	Distancia euclídea	SVM	GEQ-SVM
Supervector GMM	6.46 %	0.53 %	0.50 %

Tabla 7-5: Reconocimiento de locutor mediante supervector GMM.

Utilizando la técnica convencional de *Viterbi* se ha obtenido el resultado de la tabla 7-6.

	Viterbi
GMM	17.93 %

Tabla 7-6: Reconocimiento de locutor mediante Viterbi.

Como se puede ver en las tablas previas, con una configuración como la escogida en este trabajo se obtienen unos resultados pobres aplicando las técnicas convencionales (Viterbi). Esto se debe a que se están utilizando vectores de tan sólo 13 parámetros y 64 mezclas gaussianas para modelar las distribuciones de voz de cada hablante, mientras que en otros trabajos de referencia como en (Campell et al., 2006) se utilizan vectores de 38 coeficientes y 2048 mezclas gaussianas, con los que se obtienen valores para el EER alrededor del 5%.

Por otro lado, los resultados que se obtienen mediante la técnica SVM, bien sea normalizando los parámetros o no, son excelentes. Esto pone de manifiesto la potencia de la técnica que se está tratando en este trabajo. Partiendo de un sistema sencillo, se pueden obtener resultados que superan a los que se obtienen con otras técnicas convencionales.

7.3.1.3 *Eigenvoices y fishervoices*

Taking into account the high dimensionality that the GMM supervector can achieve, the concept of dimensionality reduction applied to eigenfaces and fisherfaces can be generalized to the GMM supervector case. In this way, principal components analysis (PCA) and linear discriminant analysis (LDA) can be applied upon the GMM supervector to reduce the feature vector length.

El Análisis de Componente Principal es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos. Esta técnica busca la proyección según la cual los datos queden mejor

representados en términos de mínimos cuadrados. El uso de PCA en el caso concreto del reconocimiento facial se conoce como la técnica Eigenfaces (Turk et al., 1991).

El PCA busca una transformación lineal que escoge un nuevo sistema de coordenadas en el cual la varianza de mayor tamaño del conjunto de datos es capturada en el primer eje (Primer Componente Principal), la segunda varianza más grande es el segundo eje, y así sucesivamente.

Para construir esta transformación lineal en primer lugar debe construirse la matriz de covarianza o matriz de coeficientes de correlación. Debido a la simetría de esta matriz existe una base completa de vectores propios de la misma. La transformación que lleva de las antiguas coordenadas a las coordenadas de la nueva base es precisamente la transformación lineal necesaria para reducir la dimensionalidad de datos.

Debido al orden de los vectores de la base de proyección según varianzas, si al transformar los vectores a un nuevo espacio utilizando PCA en lugar de tomar todas las componentes tomamos únicamente las n primeras, reducimos la dimensionalidad de los vectores abarcando la mayor variabilidad posible, y desechando las componentes que no aporten mucha variabilidad a los datos.

El Análisis Discriminante Lineal y el relacionado Discriminante Lineal de Fisher son métodos utilizados en estadística y aprendizaje para buscar la combinación lineal de características que mejor separa dos o más clases de objetos o eventos. La combinación resultante puede utilizarse como un clasificador lineal o más comúnmente para reducción de dimensionalidad previo a la clasificación. Este método aplicado en el caso concreto de reconocimiento facial es conocido con el nombre de *fisherfaces* (Belhumeur et al., 1997).

LDA utiliza información intra-clase y entre clases, mientras que PCA sólo utiliza información entre clases. Esto nos aporta mayor información para una mejor distinción entre clases con la desventaja de que las muestras de entrenamiento deben estar etiquetadas.

Si se aplica PCA o LDA al supersector GMM se pueden obtener nuevos grupos de características con menor dimensionalidad denominados *eigenvoices* y *fishervoices*.

Una característica común a los tres conjuntos de parámetros, supervector GMM, *eigenvoices* y *fishervoices* es que la longitud del vector de parámetros es fija. Esto es una ventaja para la clasificación por medio de técnicas de reconocimiento de patrones como SVM y para la fusión a nivel de parámetros de la información de voz con la proveniente de otras modalidades.

En el proceso de clasificación para reconocimiento de locutor se ha utilizado la distancia euclídea con normalización estándar y SVM con las normalizaciones estándar y GEQ.

7.3.2 Resultados unimodales

Para el reconocimiento de caras, se han utilizado como parámetros tanto *eigenfaces* como *fisherfaces*. Para la clasificación unimodal, la mayor parte de sistemas de reconocimiento de caras utilizan técnicas basadas en la distancia euclídea (Turk et al., 1991). Más recientemente, las máquinas de vector soporte (*SVM: Support Vector Machines*) se han introducido con éxito para realizar esta tarea de clasificación (Jonsson et al., 2002). Estas son las dos técnicas que se han utilizado en esta tarea.

En el primer caso, la distancia euclídea se calcula entre los vectores de evaluación y un vector modelo. En el segundo caso, una SVM realiza la clasificación mediante un hiperplano separador que se tiene que entrenar para cada cliente mediante técnicas de aprendizaje de máquinas (*machine learning*) (Burges, 1998; Cristianini et al., 2000). Para esta tarea se han utilizado SVM lineales.

En diversos trabajos se ha demostrado la importancia de la normalización de los datos antes de la fusión mediante SVM (Ross et al., 2006). Para la normalización de los parámetros de caras, se ha utilizado una normalización estándar y un método basado en la ecualización de histogramas, la ecualización a gaussiana. La normalización estándar transforma la media y la varianza de los parámetros de la biometría unimodal para que tengan valores 0 y 1 respectivamente.

De esta forma, los métodos de clasificación utilizados para el reconocimiento unimodal de caras son: una distancia euclídea sin normalización de los parámetros, para preservar los *eigenfaces* y *fisherfaces* originales, y una SVM lineal precedida por una normalización estándar (SVM) y por una ecualización a gaussiana (GEQ-SVM). El EER obtenido por cada una de las combinaciones de parámetros y técnicas de clasificación para el reconocimiento de caras se presenta en la siguiente tabla:

	Distancia euclídea	SVM	GEQ-SVM
Eigenfaces	4.10 %	1.50 %	1.50 %
Fisherfaces	2.93 %	1.25 %	1.25 %

Tabla 7-7: Resultados unimodales de reconocimiento facial.

En el caso del reconocimiento unimodal de locutores se han utilizado como parámetros supervectores GMM. Además, de forma similar a como se hace para *eigenfaces* y *fisherfaces*, la dimensionalidad de los supervectores GMM se han reducido utilizando PCA (*eigenvoices*) y LDA (*fisherfaces*). En ambos casos la matriz de transformación se ha estimado con los datos de impostores del conjunto de desarrollo del LP1.

Para la evaluación del reconocimiento de locutor se han utilizado la mismas técnicas de clasificación que en la caso del reconocimiento mediante caras excepto porque los parámetros basados en supervectores GMM se han normalizado de forma estándar antes del sistema de verificación basado en distancia euclídea. Los resultados (EER) se muestran en la siguiente tabla:

	Distancia euclídea	SVM	GEQ-SVM
Supervector GMM	6.46 %	0.53 %	0.50 %
Eigenvoices	6.59 %	1.25 %	1.23 %
Fishervoices	7.40 %	1.50 %	1.00 %

Tabla 7-8: Resultados unimodales de reconocimiento de locutor.

En este punto, se han revisado los resultados unimodales obtenidos por los sistemas de voz y caras que nos servirán como base para valorar los resultados multimodales y hemos presentado alguno de los métodos de normalización y fusión que también se utilizarán para la fusión, como pueden ser GEQ y SVM.

7.3.3 Fusión a nivel de parámetros

Para la fusión a nivel de parámetros, las características extraídas a partir de las señales de caras y voz se concatenan para obtener un vector conjunto de parámetros. Teniendo en cuenta los resultados obtenidos en los experimentos unimodales se ha elegido como técnica de clasificación una SVM lineal con normalización GEQ. La longitud fija para una sentencia de voz de los vectores de parámetros basados en supervectores GMM hace que la longitud del vector multimodal también sea de longitud fija y sea factible la clasificación mediante SVM.

Como ya se ha comentado anteriormente, la ecualización a gaussiana es una técnica basada en la ecualización de histogramas. Las técnicas de ecualización de histogramas (*HEQ: Histogram Equalization*) transforman la distribución estadística de una serie de datos para que se asemeje a una distribución de referencia. Esta técnica se puede ver como una extensión de la normalización estadística realizada por la normalización estándar a la estadística global de los datos y no solo a su media y su varianza.

En el caso de la ecualización a gaussiana (*GEQ: Gaussian Equalization*) la referencia es una distribución normal con varianza 1.

Para obtener resultados de fusión de parámetros se han combinado tanto *eigenfaces* como *fisherfaces* con los parámetros de los supervectores GMM y, además, los eigenparámetros, es decir,

eigenfaces con *eigenvoices*, y los fisherparámetros, *fisherfaces* con *fishervoices*. Los resultados obtenidos se muestran en la siguiente tabla:

	GEQ-SVM
Eigenfaces + supervector GMM	0.008 %
Eigenfaces + Eigenvoices	0.25 %
Fisherfaces + supervector GMM	0.026 %
Fisherfaces + Fishervoices	0.25 %

Tabla 7-9: Fusión a nivel de parámetros.

El mejor resultado lo ha obtenido la utilización conjunta de *eigenfaces* y los parámetros del supervector GMM y es el menor EER obtenido en los experimentos realizados para esta tarea. Además, la utilización del supervector GMM completo obtiene mejores resultados que *eigenvoices* y *fishervoices*.

7.3.4 Fusión a nivel de puntuaciones

En el caso de la fusión a nivel de puntuaciones también se ha demostrado la importancia de la normalización de los datos de forma previa al proceso de fusión (Jain et al., 2005; Ross et al., 2006). Por esta razón, la ecualización de doble gaussiana propuesta (*BGEQ: Bi-Gaussian Equalization*) se ha utilizado para normalizar las puntuaciones antes de la clasificación mediante SVM, ya que esta técnica ha demostrado ser superior a otras normalizaciones en los resultados del apartado 7.2.

En la ecualización mediante BGEQ la distribución de referencia es la suma de dos funciones gaussianas, cuyas desviaciones estándar σ modelan el solape entre los lóbulos de las puntaciones correspondientes a clientes e impostores de la distribución original.

Las puntuaciones para realizar la fusión se han obtenido de los sistemas unimodales. En concreto se han utilizado los sistemas basados en GEQ-SVM. Como ya se ha dicho, las puntaciones procedentes de estos sistemas se han normalizado mediante BGEQ y los vectores de puntuaciones multimodales se han clasificado mediante una SVM lineal. Los resultados obtenidos son los mostrados en la tabla 7-10.

Las puntuaciones de los sistemas basados en *eigenfaces* y *fisherfaces* en combinación con las puntuaciones del sistema basado en supervectores GMM obtienen los mejores resultados y superan a los sistemas que incluyen reducción de la dimensionalidad del supervector GMM.

	BGEQ-SVM
Eigenfaces + supervector GMM	0.019 %
Eigenfaces + Eigenvoices	0.25 %
Fisherfaces + supervector GMM	0.017 %
Fisherfaces + Fishervoices	0.25 %

Tabla 7-10: Fusión a nivel de puntuaciones.

Los resultados obtenidos con fusión a nivel de puntuaciones con "fisherfaces + supervector GMM" mejoran los obtenidos en el caso de fusión de parámetros, pero es al revés para el caso "eigenvoices + supervector GMM", que, como ya se ha comentado, obtiene el mejor resultado de la tarea con fusión de parámetros.

7.3.5 Fusión a nivel de decisión

Para la fusión a nivel de decisión se ha decidido utilizar un votador de mayoría. Como, en ese caso, es necesario que el número de expertos utilizado sea impar, se han probado las combinaciones coherentes de decisiones de los eigensistemas y fishersistemas añadiendo en ambos casos la decisión del sistema GMM supervector, que es el que obtiene el mejor resultado unimodal. También se ha calculado el resultado obtenido mediante la votación de las cinco técnicas.

Para obtener los resultados, se ha hecho trabajar cada uno de los expertos en el punto en que obtienen su EER y se han calculado las tasas falsa aceptación (FAR) y falso rechazo (FRR) de los sistemas multimodales. También se incluye en los resultados la tasa media de error total (HTER) de cada sistema de votación de mayoría. Los resultados se muestras en la tabla 7-11.

	FAR	FRR	HTER
Eigenfaces + Eigenvoices + supervector GMM	0.185 %	0.25 %	0.217 %
Fisherfaces + Fishervoices + supervector GMM	0.292 %	0.5 %	0.396 %
Eigenfaces + Eigenvoices + Fisherfaces + Fishervoices + supervector GMM	0.125 %	0.25 %	0.187 %

Tabla 7-11: Fusión a nivel de decisión.

La combinación de todos los expertos obtiene los mejores resultados, superando los mejores resultados unimodales, pero sin superar los obtenidos mediante fusión de parámetros o puntuaciones.

Como conclusión, en este apartado, se ha fusionado la información espectral de voz y de imágenes faciales a tres niveles diferentes: parámetros, puntuaciones y decisión, a partir del trabajo realizado en el entorno del proyecto Agatha.

La utilización de supervectores GMM ha permitido realizar de manera simple la fusión de parámetros de voz con otras biometrías dado que la longitud de los vectores de parámetros es constante. Además, la reducción de la dimensión de dicho supervector mediante PCA o LDA no ha mejorado los resultados obtenidos con el supervector completo.

La aplicación de técnicas de fusión, tanto a nivel de parámetros como a nivel de puntuaciones, ha mejorado ostensiblemente los resultados de verificación unimodales llegando en este caso a una reducción del EER de un 98% entre los mejores resultados unimodal y multimodal. La aplicación de técnicas de normalización basadas en ecualización de histograma junto con SVM ha sido clave en estas mejoras.

8 Conclusiones

Diversos trabajos demuestran la importancia de la normalización de parámetros y puntuaciones en los procesos de fusión biométrica multimodal. En esta tesis se han presentado diferentes técnicas de normalización basadas en la adaptación de la estadística de dichas características biométricas.

En el ámbito de la fusión a nivel de puntuaciones se han presentado diversas normalizaciones de la media y la varianza de las puntuaciones de clientes e impostores para reducir las varianzas multimodales. En un sistema bimodal de fusión de puntuaciones de espectro de voz y caras, presentado en el apartado 6.1, estas técnicas han mejorado o han obtenido resultados similares a las técnicas convencionales mediante la utilización de métodos de fusión combinatorios.

En el mismo sistema bimodal, se ha utilizado también como método de normalización la ecualización de histogramas a una de las modalidades biométricas involucradas en el proceso de fusión. La utilización de esta técnica mejora los resultados obtenidos por las técnicas de normalización convencionales y otras técnicas de normalización no afines. En particular, la combinación de ecualización de histogramas, normalización conjunta de medias y las técnicas de fusión ponderadas han obtenido para estos experimentos bimodales los mejores resultados, incluso los obtenidos mediante el uso de una SVM.

Por otro lado, se ha probado una cierta relación entre las varianzas de las puntuaciones de clientes e impostores y la fiabilidad de una biometría. Esta relación se puede explotar para mejorar los resultados de las biometrías multimodales.

En el apartado 6.2 se exploran diferentes estrategias para la fusión de puntuaciones de espectro de voz, prosodia y caras. En este caso, el rendimiento del sistema basado en la información facial y de espectro de voz mejora cuando la información prosódica se añade al sistema. El mejor resultado se obtiene mediante la fusión en dos pasos donde las características prosódicas se fusionan en primer lugar obteniendo una puntuación conjunta para, posteriormente, fusionarla con las puntuaciones de

espectro de voz y caras. La utilización de ecualización de histograma como normalización previa a una fusión basada en máquinas de vector soporte ha permitido superar los resultados obtenidos por el resto de técnicas utilizadas, incluyendo técnicas convencionales de normalización y técnicas de fusión basadas en la ponderación de las puntuaciones y en el aprendizaje.

Por otro lado, en experimentos multimodales de verificación de personas (apartado 6.3), donde se han fusionado puntuaciones de prosodia, espectro de voz y caras, la introducción de técnicas de ecualización de histograma han reducido las tasas de error proporcionadas por el resto de las técnicas de normalización utilizadas, tanto en la aproximación combinatoria como en la clasificatoria. Además, la utilización de máquinas de vector soporte ha mejorado los resultados obtenidos por los métodos de suma ponderada de las puntuaciones.

En concreto, la ecualización de doble gaussiana, una técnica de ecualización que tiene en cuenta la distribución separada de clientes e impostores, ha obtenido los mejores resultados en combinación con un sistema de fusión basado en SVM con kernel polinomial. De hecho, las técnicas de normalización basadas en ecualización han superado a las técnicas convencionales de normalización para todos los valores de FAR y FRR tanto para el kernel RBF como para el polinomial.

También en el ámbito de esta tesis, se ha colaborado en el proyecto Agatha, cuyo objetivo era crear una plataforma de identificación de personas mediante biometría multimodal. La aportación, presentada en el apartado 7, ha consistido en la fusión multimodal de información espectral de voz y de imágenes faciales a tres niveles diferentes: parámetros, puntuaciones y decisión.

Para realizar la fusión a nivel de parámetros, se han utilizado supervectores GMM como parámetros de voz, lo que ha permitido realizar de forma simple la fusión de dichos parámetros con los de otras biometrías, dado que la longitud de los vectores de parámetros de voz es constante en este caso. La utilización en este caso de técnicas de reducción de parámetros como PCA o LDA no ha mejorado los resultados obtenidos con el supervector completo.

Para la normalización de los parámetros y las puntuaciones unimodales se han utilizado técnicas de ecualización de histograma, en concreto, ecualización a gaussiana para la normalización de los parámetros y ecualización de doble gaussiana para la normalización de las puntuaciones. La aplicación de estas técnicas de normalización junto con una fusión basada en SVM lineal ha mejorado ostensiblemente los resultados de verificación unimodales llegando en este caso a una reducción del EER en un 98% entre los mejores resultados unimodal y multimodal.

De todas estas consideraciones, se puede concluir que, en un sistema biométrico multimodal de fusión de parámetros o puntuaciones la selección de una técnica de normalización es un tema fundamental y puede resultar tan decisivo como la selección del método de fusión. En este ámbito, las técnicas de normalización estadística presentadas en esta tesis han obtenido buenos resultados y

creemos que deben tomarse en consideración en el diseño de sistemas biométricos multimodales de fusión de parámetros o puntuaciones.

Como posibles líneas de trabajo futuro cabe destacar la aplicación de las técnicas presentadas en esta tesis sobre bases de datos multimodales que contengan registros de otras biometrías, además de la voz y las caras, y que tengan una extensión suficiente para garantizar resultados estadísticamente fiables. También podría resultar de interés comprobar la eficacia de los métodos basados en ecualización de histograma sobre características obtenidas en ambientes ruidosos o con interferencias. Parece razonable pensar que la transformación de los intervalos de las señales ruidosas a los intervalos de referencia pueda aumentar la robustez de los sistemas de reconocimiento. Otra tarea de interés sería investigar técnicas de normalización no lineales que reduzcan las varianzas de las puntuaciones multimodales de clientes e impostores, así como determinar el efecto de los kernel de las máquinas de vector soporte sobre dichas varianzas.

Referencias

R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, 10(1), pages 42-54, 2000.

B.S. Atal. Automatic speaker recognition based on pitch contours. *Journal of the Acoustical Society of America*, 52, pages 1687-1697, 1972.

M. Awad and L. Khan. Applications and Limitations of Support Vector Machines. *Knowledge Discovery and Data Mining*, 2(2), pages 121-167, 1998.

E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, The BANCA Database and Evaluation Protocol, *Springer LNCS-2688, 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA'03*, Springer-Verlag, 2003.

W. Baker, A. Evans, L. Jordan, and S. Pethe. User Verification System. *In proceedings of MASPLAS'02 The Mid-Atlantic Student Workshop on Programming Languages and Systems Pace University*, April, 2002.

R. Balchandran and R. Mammone. Non parametric estimation and correction of non-linear distortion in speech systems. *In proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, II*, Seattle, WA, pages 749–752, 1998.

P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7), pages 711-720, 1997.

S. Bengio. Multimodal Authentication Using Asynchronous HMMs. *In proceedings of AVBPA'2003, Audio- and Video-Based Biometric Person Authentication, 4th International Conference*, pages 770-777, Guilford, UK, June 2003.

S. Ben-Yacoub. Multi-Modal Data Fusion for Person Authentication Using SVM. *In proceedings of International Conference on Audio- and Video-Based Person Authentication*, pages 25-30, 1999.

S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz. Fusion of face and speech data for person identity verification. *IEEE Trans. on Neural Networks*, 10(5), pages 1065-1074, 1999.

E. S. Bigun, J. Bigun, B. Duc, and S. Fisher. Expert conciliation for multi modal person authentication systems by Bayesian statistics. *In proceedings of the 1st International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, Crans-Montana, Switzerland, pages 291-300, March 1997a.

J. Bigun, B. Duc, S. Fischer, A. Makarov, and F. Smeraldi. Multi modal person authentication. *Nato-Asi advanced study on face recogniton, H. Wechsler et. al.*, F-163, pages 26–50, Springer, 1997b.

F. Bimbot, J.F. Bonastre , C. Fredouille , G. Gravier , I. Magrin-Chagnolleau , S. Meignier , T. Merlin , J. Ortega-García , D. Petrovska-Delacrétaz , D.A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 2004(1), pages 430-451, January 2004.

C. Bishop. Neural Networks for Pattern Recognition. *Oxford University Press*, Oxford, 1995.

W.W. Boles and B. Boashash. A human Identification Technique Using Images of the Iris and Wavelet Transform. *IEEE Transactions On Signal Processing*, 46(4), pages 1185-1188, April 1998.

R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, A.W. Senior. Guide to Biometrics. *Springer-Verlag New York*, Inc. 2004

R.M. Bolle, S. Pankanti, and N.K. Ratha. Evaluation techniques for biometrics-based authentication systems (FRR). *In proceedings of ICPR 2000, 15th International Conference on Pattern Recognition*, 2, pages 831 -837, September 2000.

C.J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Knowledge Discovery and Data Mining*, 2(2), pages121-167 1998.

C. Burges, P. Knirsch, and R. Haratsh, Support vector web page: http://svm.research.belllabs.com. Technical Report. *Lucent Technologies*, 1996.

W.M. Campbell, D.E. Sturim, and D.A. Reynolds. Support Vector Machines Using GMM Supervectors for Speaker Verification. *IEEE Signal Processing Letters*, 13(5), pages 308-311, 2006.

W.M. Campbell, J.P. Campbell, D.A. Reynolds, D.A. Jones, and T.R. Leek. High-Level Speaker Verification with Support Vector Machines. *In proceedings of the International Conference on Acoustics, Speech, and Signal Processing, IEEE*, I, pages 73-76, 2004.

J.P. Campbell, D.A. Reynolds, and R.B. Dunn. Fusing high- and low-level features for speaker recognition. *In proceedings Eurospeech*, pages 2665-2668, 2003a.

W.M. Campbell, J.P. Campbell, D.A. Reynolds, D.A. Jones, and T.R. Leek. Phonetic Speaker Recognition with Support Vector Machines. *In proceedings of the Neural Information Processing Systems Conference*, pages 1377-1384, 2003b.

J.P. Campbell and D.A. Reynolds. Corpora for the evaluation of speaker recognition systems. *Presented at ICASSP*, Phoenix, Arizona, 1999.

J.P. Campbell. Speaker recognition: A tutorial. *In Proceedings of the IEEE*, 85, pages 1437–1462, 1997.

S. Carcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux les Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacrtaz. BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities. *Springer LNCS-2688, 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA'03)*, pages 845–853, Guildford, 2003.

M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett. Robust prosodic features for speaker identification. *Presented at ICSLP*, Philadelphia, 1996.

U.V. Chaudhari, G.N. Ramaswamy, G. Potamianos, and C. Neti. Information fusion and decision cascading for audiovisual speaker recognition based on time-varying stream reliability prediction. *In proceedings of the International Conference on Multimedia & Expo 2003 (ICME2003)*, 3, pages 9-12, July 2003a.

U.V. Chaudhari, G.N. Ramaswamy, G. Potamianos, and C. Neti. Audio-visual speaker recognition using time-varying stream reliability prediction. *In proceedings of the International Conference on Acoustics, Speech and Signal Processing 2003 (ICASSP 2003)*, V, pages 712-715, 2003b.

V. Chatzis, A.G. Bors, and I. Pitas. Multimodal decision-level fusion for person authentication. *IEEE Trans. On System, Man, and Cybernetics*, part A, 29(6), pages 674-680, 1999.

T. Chen. Audiovisual Speech Processing. *IEEE Signal Processing Magazine*, 18, pages 9-21, January 2001.

M. Cheung, M. Mak, and S. Kung. A Two-Level Fusion Approach to Multimodal Biometric Verification. *In proceedings of the International Conference on Acoustics, Speech, and Signal Processing, IEEE*, pages 485-488, 2005.

R. Clarke. Human identification in information systems: Management challenges and public policy issues. *Information Technology & People*, 7(4), pages 6-37, December 1994.

N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines (and other kernel-based learning methods). *Cambridge University Press*, 2000.

N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On Kernel-Target Alignment. *Neural Information Processing Systems, NIPS*, 2001.

J. Czyz, M. Sadeghi, J. Kittler, and L. Vandendorpe. Decision fusion for face authentication. *International Conference on Biometric Authentication, LNCS proceedings*, HongKong, 2004a.

J. Czyz, J. Kittler, and L. Vandendorpe. Multiple classifier combination for face-based identity verification. *Pattern Recognition*, 34(7), 2004b.

J. Czyz, S. Bengio, C. Marcel, and L. Vandendorpe. Scalability Analysis of Audio-Visual person identity verification. *In proceedings of International Conference on Audio- and Video-based Person Authentication*, 2003.

J. Czyz, J. Kittler, and L. Vandendorpe. Combining face verification experts. *16th International Conference on Pattern Recognition (ICPR 2002)*, 2, pages 28-31, August 2002.

J. Daugman. Biometric Decision Landscapes. *Technical Report 482, The Computer Laboratory*, University of Cambridge, 1999.

P.A. Devijver and J. Kittler. Pattern Recognition: A Statistical Approach. *Prentice-Hall*, London, 1982.

T.G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 7(10), pages 1895–1924, 1998.

G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheeps, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *In proceedings of ICSLD*, Sydney, Australia, November 1998.

B. Duc, E.S. Bigun, J. Bigun, G. Maitre, and S. Fischer. Fusion of audio and video information for multi modal person authentication. *Pattern Recognition Letters*, 18, pages 835-843, 1997.

R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. *John Wiley & Sons*, 2001

R.P.W. Duin. The combining classifier: to train or not to train? *In proceedings of the IAPR Intl. Conf. on Pattern Recognition, ICPR*, pages 765-770. IEEE CS Press, 2002.

G. Feng, K. Dong, D. Hu1, and D. Zhang. When Faces Are Combined with Palmprints: A Novel Biometric Fusion Strategy. *In proceedings of International Conference on Biometric Authentication, ICBA*, pages 701-707, 2004.

J. Fierrez, J. Ortega-Garcia, D. Torre-Toledano and J. Gonzalez-Rodriguez. BioSec baseline corpus: A multimodal biometric database. *Pattern Recognition*, 40(4), pages 1389-1392, April 2007.

J.A. Fierrez, J.G. Ortega, J.R. Gonzalez, J. Bigun. Kernel-based multimodal biometric verification using quality signals. *Defense and Security Symposium, Biometric Technologies for Human Identification, BTHI, Proc. SPIE*, 5404, pages 544-554, Orlando, USA, April 2004.

J.A. Fierrez, J.G. Ortega, D.R. Garcia, J.R. Gonzalez. A comparative evaluation of fusion strategies for multimodal biometric verification. *In proceedings of IAPR International Conference on Audio- and Video-based Person Authentication, AVBPA*, pages 830-836, Springer, 2003a.

J.A. Fierrez, J.G. Ortega, J.R. Gonzalez. Fusion Strategies in Multimodal Biometric Verification. *In proceedings of the IEEE International Conference on Multimedia and Expo*, 3, pages 5-8, 2003b.

N.A. Fox, R. Gross, J.F. Cohn, and R.B. Reilly. Robust Automatic Human Identification Using Face, Mouth, and Acoustic Information. *Analysis and Modelling of Faces and Gestures (AMFG)*, pages 264-278, Beijing, China, October 16, 2005.

N.A. Fox, R.B. Reilly. Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features. *In proceedings of the 4th International Conference on Audio and Video Based Biometric Person Authentication*, June 2003a.

N.A. Fox, R. Gross, P. de Chazal, J.F. Cohn, and R.B. Reilly. Person Identification Using Automatic Integration of Speech, Lip, and Face Experts. *In proceedings of the ACM SIGMM Multimedia Biometrics Methods and Applications Workshop (WBMA)*, pages 25-32, Berkeley, CA., November 2003b.

T. Fu, X.X. Liu, L.H. Liang, X. Pi, and A.V. Nefian. A Audio-Visual Speaker Identification using Coupled Hidden Markov Models. *In proceedings of the International Conference on Image Processing*, 3, pages 29-32, September 2003.

S. Furui. Recent advances in speaker recognition. *In J. Bigün, G. Chollet, and G. Borgefors, editors, Audio- and Video-based Biometric Person Authentication, volume 1206 of Lecture Notes in Computer Science*, pages 237-252, Springer-Verlag, Heidelberg, Germany, 1997.

D.R. Garcia, J.A. Fierrez, J.R. Gonzalez, J.G. Ortega. Using quality measures for multilevel speaker recognition. *Computer Speech and Language, Special Issue on Odyssey-04: The Speaker and Language Recognition Workshop*, 20(2-3), pages 192-209, April-July 2004.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. *Presented at ICASSP*, 1990.

M. Gurban and J. Thiran. Audio-Visual Speech Recognition with a Hybrid SVM-HMM System. *13th European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, 2005.

B. Gutschoven and P. Verlinde. Multi-Modal Identity Verification using Support Vector Machines (SVM). *In proceedings of the International Conference on Information Fusion*, pages 3-8, 2000.

S. Gutta, J.R.J. Huang, P. Jonathon, and H. Wechsler. Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *IEEE Transactions on Neural Networks 11*, pages 948-960, July 2000.

F.R. Hampel, P.J. Rousseeuw, E.M. Ronchetti, W.A. Stahel. Robust Statistics: The Approach Based on Influence Functions. *John Wiley & Sons*, 1986.

A.O. Hatch, A. Stolcke, and B. Peskin. Combining feature sets with support vector machines: Application to speaker recognition. *In proceedings of IEEE Speech Recognition and Understanding Workshop*, pages 75-79, San Juan, Puerto Rico, November 2005.

M.A. Hearst. Trends and Controversies: Support Vector Machines. *IEEE Intelligent Systems*, 13, pages18-28, 1998.

J. Hernando and C. Nadeu. Speaker Verification on the POLYCOST Database Using Frequency Filtered Spectral Energies. *In proceedings of the ICSLP*, 1998.

A. Higgins, L. Bhaler, and J. Porter. Voice Identification using Nearest Neighbor Distance Measure. *In proceedings of the ICASSP*, pages 375-378, 1993.

F. Hilger and H. Ney. Quantile based histogram equalization for noise robust speech recognition. *In proceedings of Eurospeech*, pages 1135-1138, Aalborg, Dinamarca, September 2001.

IBG BioPrivacy Initiative website. http://www.bioprivacy.org/.

M. Indovina, U. Uludag, R. Snelick, A. Mink, and A. Jain. Multimodal Biometric Authentication Methods: A COTS Approach. *In proceedings of MMUA, Workshop on Multimodal User Authentication*, pages 99-106, Santa Barbara, CA, December 2003.

A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12), pages 2270-2285, December 2005.

A. Jain and A. Ross. Multibiometric systems. *Comm. of the ACM 47*, pages 34-40, 2004a.

A. Jain, S. Dass, and K. Nandakumar. Can soft biometric traits assist user recognition? *In proceedings of SPIE, Biometric Technology for Human Identification*, 5404, August 2004b.

A. Jain and A. Ross. Learning User-Specific Parameters in Multibiometric System. *In proceedings of International Conference on Image Processing, ICIP*, pages 57-60, Rochester, NY, September 2002.

A. Jain, S. Prabhakar, and S. Chen. Combining multiple matchers for a high security fingerprint verification system. *Pattern Recognition Letters*, 20, pages 1371-1379, 1999.

A.K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, 14(1), pages 4–20, January 1994.

A. Jain. Fundamentals of Digital Image Processing. *Prentice-Hall*, pages 241-243, 1986.

K. Jonsson, J. Kittler, Y.P. Li, and J. Matas. Support vector machines for face authentication. *Image and Vision Computing*, 20(5-6), pages 369-375, 2002.

S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sönmez, E. Shriberg, A. Stolcke, H. Bratt, and R.R. Grade. Speaker recognition using prosodic and lexical features. *Presented at IEEE Speech Recognition and Understanding Workshop*, 2003.

kernel machines website. http://www.kernel-machines.org/.

J. Kittler and F.M. Alkoot. Sum versus vote fusion in multiple classifier systems. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 25(1), pages 110-115, 2003.

J. Kittler and K. Messer. Fusion of Multiple Experts in Multimodal Biometric Personal Identity Verification Systems. *In proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pages 3-12, 2002.

J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3), pages 226-239, 1998.

J. Kittler, G. Matas, K. Jonsson, and M. Sánchez. Combining evidence in personal identity verification systems. *Pattern Recognition Letters*, 18(9), pages 845-852, September 1997a.

J. Kittler, Y. Li, J. Matas, and M.U. Sanchez. Combining Evidence in Multimodal Personal Identity Recognition Systems. *In proceedings of the 1st International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, Crans-Montana, Switzerland, pages 327-334 1997b.

L.I. Kuncheva. Combining Pattern Classifiers: Methods and Algorithms. *Wiley*, July 2004.

International Biometric Industry Association website. http://www.ibia.org/.

P. Langley, W. Iba, and K. Thompson. An analysis of Bayesian classifiers. *In proceedings of the Tenth National Conference on Artificial Intelligence*, pages 223-228, San Jose, CA, AAAI Press, 1992.

H.C. Lee, R.E. Gaensslen. Advances in Fingerprint Technology. *CRC Press*, Boca Raton, FL, 1994.

D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Presented at Advances in Neural Information Processing Systems: Proceedings of the 2000 Conference*, 2001.

A. Leon-Garcia. Probability and Random Processes for Electrical Engineering (2nd Edition). *Addison-Wesley Pub Co.*, July, 1993.

S. Lucey and T. Chen. Improved audio-visual speaker recognition via the Use of a hybrid combination strategy. *The 4th International Conference on Audio- and Video- Based Biometric Person Authentication (AVBPA)*, Guildford, U.K., June 2003.

J. Lüttin and G. Maître. Evaluation Protocol for the Extended M2VTS Database (XM2VTSDB). *IDIAP Communication 98-05*, Martigny, Switzerland, 1998.

B. Maison, C. Neti, and A. Senior. Audio-visual speaker recognition for video broadcast news: Some fusion techniques. *Invited paper in Journal of VLSI Signal Processing special issue on Multimedia*, 29(1/2), 2001.

D. Maltoni, D. Maio, A.K. Jain, and S. Prabhakar. Handbook of Fingerprint Recognition. *Springer*, New York, 2003.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. *In Eurospeech*, 4, pages 1895-1898, Rhodos, Greece, 1997.

H. Melin and L. Lindberg. Guidelines for experiments on the POLYCOST database. *In proceedings of COST 250 workshop on application of speaker recognition technologies in telephony*, Vigo, Spain,1996.

B. Miller. Vital signs of identity. *IEEE Spectrum*, 31(2), pages 22-30, February 1994.

N. Mirghafori, A.O. Hatch, S. Stafford, K. Boakye, D. Gillick, and B. Peskin. ICSI's 2005 Speaker Recognition System. *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 23-28, San Juan, Puerto Rico, November 2005.

C. Nadeu, J. Hernando, and M. Gorricho. On the decorrelation of filter bank energies in speech recognition. *Presented at Eurospeech*, 1995.

A.V. Nefian, L.H. Liang, T. Fu, and X.X. Liu. A Bayesian Approach to Audio-Visual Speaker Identification. *In proceedings AVBPA, Audio- and Video-Based Biometric Person Authentication, 4th International Conference*, pages 761-769, Guilford, UK, June 2003.

NIST website. http://www.nist.gov/.

NIST Speaker Recognition website. http://www.nist.gov/speech/tests/spk/2001/.

J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J.J. Igarza, C. Vivaracho, D. Escudero, and Q.I. Moro. Biometric on the Internet MCYT Baseline Corpus: a Bimodal Biometric Database. *IEE Proc. Visual Image Signal Processing*, 150(6), pages 395-401, December 2003.

J. Pelenacos and S. Sridharan. Feature warping for robust speaker verification. *In proceedings of ISCA Workshop on Speaker Recognition: A Speaker Odyssey*, pages 213-218, June 2001.

D. Petrovska, J. Hennebert, H. Melin, and D. Genoud. POLYCOST: a Telephone-Speech Database for Speaker Recognition. *RLA2C*, pages 211-214, France, 1998.

B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D.A. Reynolds, and B. Xiang. Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02. *Presented at ICASSP*, 2003.

N. Poh and S. Bengio. Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments? *IDIAP, Research Report 05-20*, 2005a.

N. Poh and S. Bengio. How do correlation and variance of base classifiers affect fusion in biometric authentication tasks? *IEEE Trans. on Signal Processing*, 53(11), pages 4384-4396, 2005b.

N. Poh and S. Bengio. Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication. *Research Report 04-44, IDIAP*, Martigny, Switzerland, 2004.

N. Poh and J. Kittler. Incorporating Model-Specific Score Distribution in Speaker Verification Systems. *IEEE transactions on audio, speech, and language*, 16(3), pages 594-606, 2008.

G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: an overview. *Issues in audio-visual speech processing (G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, eds.), MIT Press*, 2004.

L.R. Rabiner and B.H. Juang. Fundamentals of Speech Recognition, *Prentice Hall*, 1993.

L.A. Rabiner. A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *In proceedings of the IEEE*, 77(2), pages 257-286, 1989.

A. Rattani and M. Tistarelli. Robust Multi-modal and Multi-unit Feature Level Fusion of Face and Iris Biometrics. *Lecture Notes in Computer Science*, pages 960-969, Springer Berlin / Heidelberg, 2009.

D.A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. *Presented at ICASSP*, 2003.

D.A. Reynolds, T.F. Quatieri, and R. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3), pages 19-41, 2000.

F. Roli, J. Kittler, G. Fumera, and D.Muntoni. An experimental comparison of classifier fusion rules for multimodal personal identity verification systems. *In proceedings of Third International Workshop on Multiple Classifier Systems, MCS*, pages 252-261, 2002a.

F. Roli, G. Fumera, and J. Kittler. Fixed and trained combiners for fusion of imbalanced pattern classifiers. *In proceedings of the International Conference on Information Fusion*, pages 278-284, 2002b.

A. Ross, A. Jain, and J. Qian. Information Fusion in Biometrics. *In proceedings of the 3rd Audio and Video-Based Person Authentication*, pages 354-359, 2001.

A. Ross, K. Nandakumar, and A. Jain. Handbook of Multibiometrics. *International Series on Biometrics, Springer-Verlag*, New York, 2006.

C. Sanderson. Biometric Person Recognition: Face, Speech and Fusion. *VDM Verlag*, 2008.

D. Shah, K.J. Han, and S.S. Narayanan. A Low-Complexity Dynamic Face-Voice Feature Fusion Approach to Multimodal Person Recognition. *In proceedings of the 11th IEEE International Symposium on Multimedia*, pages 24-31, 2009.

R. Snelick, M. Indovina, J. Yen, and A. Mink. Multimodal Biometrics: Issues in Design and Testing. *In proceedings of The 5th International Conference on Multimodal Interfaces (IMCI)*, Vancouver, British Columbia, Canada, November 2003.

R. Snelick, U. Uludag, and A. Mink. Large scale evaluation of multimodal biometric authentication using state-of the-art systems. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 27(3), pages 450-455, March 2005.

A. Stolcke, L. Ferrer, S. Kajarekar, E. Shrigerg, and A. Venkataraman. MLLR Transforms as Features in Speaker Recognition. *In proceedings of Eurospeech*, pages 2425-2428, Lisbon, 2005.

Q. Tao and R. Veldhuisa. Threshold-optimized decision-level fusion and its application to biometrics. *Pattern Recognition*, 42(5), pages 823-836, May 2009.

S. Theodoridis and K. Koutroumbas. Pattern Recognition. *Academia Press*, 2003.

O. Thyes, R. Kuhn, P. Nguyen, and J.C. Junqua. Speaker identification and verification using eigenvoices. *Presented at ICSLP*, 2, pages 242-245, Beijing, China, 2000.

M.A. Turk, A.P. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), pages 71-86, 1991.

J.R. Vacca. Biometric Technologies and Verification Systems. *Butterworth-Heinemann*, 2007.

V.N. Vapnik. The Nature of Statistical Learning Theory. *Springer*, 2000.

P. Verlinde, G. Chollet, and M. Acheroy. Multi-modal identity verification using expert fusion. *Information Fusion*, 1(1), pages 17-33, 2000.

V. Wan and S. Renals. Speaker Verification Using Sequence Discriminant Support Vector Machines. *IEEE Transactions on Speech and Audio Processing*, 13, pages 203-210, 2005.

Z. Wang, Q. Han, X. Niu, and C. Busch. Feature-Level Fusion of Iris and Face for Personal Identification. *Lecture Notes in Computer Science*, pages 356-364, Springer Berlin / Heidelberg, 2009.

Y. Wang, Y. Wang, and T. Tan. Combining Fingerprint and Voiceprint Biometrics for Identity Verification: an Experimental Comparison. *In proceedings of International Conference on Biometric Authentication, ICBA*, pages 663-670, Hong Kong, China, July 2004.

J.L. Wayman, A.K. Jain, D. Maltoni, and D. Maio. Biometric Systems: Technology, Design and Performance Evaluation. *Springer*, New York, 2005.

J. Wayman. A path forward for multi-biometrics. *In proceedings of International Conference on Analytical Sciences and Spectroscopy*, 5, pages VV, 2006.

H. Wechsler. Reliable Face Recognition Methods: System Design, Implementation and Evaluation. *Springer*, 2006.

R.P. Wildes. Iris recognition: An emerging biometric technology. *In proceedings of the IEEE*, 85(9), pages 1348-1363, September 1997.

J.J. Wolf. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, 51, pages 2044-2056, 1972.

J.D. Woodward. Biometrics: Privacy's Foe or Privacy's Friend? *In proceedings of the IEEE*, 85(9), page 1487, September 1997.

S. Zafeiriou, A. Tefas, and I. Pitas. Discriminant NMFfaces for frontal face verification. *In proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Mystic, Connecticut, September 2005a.

S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas. Exploiting discriminant information in non-negative matrix factorization with application to face verification. *IEEE Transactions on Neural Networks*, 2005b.

W. Zhao, R. Chellapa, A. Rosenfeld, and P.J. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399-458, 2003.

Y. Zuev, S. Ivanon. The voting as a way to increase the decision reliability. *Foundations of Information/Decision Fusion with Applications to Engineering Problems*, pages 206-210, Washington D.C., USA, 1996.

R. Zunkel. Hand geometry based authentication. *In A.K. Jain, R.M. Bolle, and S. Pankante, editors, Biometrics: Personal Identification in Networked Society*, Kluwer Academic Press, pages 87-102, Boston, MA, 1999.

Publicaciones del autor

- P. Ejarque and J. Hernando. Variance reduction by using separate genuine- impostor statistics in multimodal biometrics. In *Proceedings of the Interspeech*, pages 785-788, Lisbon, Portugal, September 2005.
- J. Hernando, M. Farrús, P. Ejarque, A. Garde, and J. Luque. Person verification by fusion of prosodic, voice spectral and facial parameters. In *Proceedings of the International Conference on Security and Cryptography*, pages 17-23, Setúbal, Portugal, August 2006.
- M. Farrús, A. Garde, P. Ejarque, J. Luque, and J. Hernando. On the fusion of prosody, voice spectrum and face features for multimodal person verification. In *Proceedings of the Interspeech*, pages 2106-2109, Pittsburgh.USA, September 2006.
- P. Ejarque, A. Garde, J. Anguita, and J. Hernando. On the use of genuine-impostor statistical information for score fusion in multimodal biometrics. *Multimodal Biometrics in Annals of Telecommunications 62(1-2)*, pages 109-129, January, 2007.
- P. Ejarque and J. Hernando. On the effect of score equalization in SVM multimodal biometric systems. In *Proceedings of the International Conference on Security and Cryptography*, pages 33-38, Barcelona, Spain, July 2007.
- M. Farrús, J. Hernando, and P. Ejarque. Jitter and Shimmer Measurements for Speaker Recognition. In *Proceedings of the Interspeech*, pages 778-781, Antwerp, Belgium, August 2007.
- M. Farrús, P. Ejarque, A. Temko, and J. Hernando. Histogram Equalization in SVM Multimodal Person Verification. In volume 4642 of *Lecture Notes in Computer Science*, pages 819-827, International Conference on Biometrics, Seoul, Korea, August 2007.
- P. Ejarque and J. Hernando. Bi-Gaussian Score Equalization in an Audio-Visual SVM-based Person Verification System. In *Proceedings of the Interspeech*, pages 2663-2666, Brisbane, Australia, Sept. 2008.

- D. Hernando, D. Gómez, J. R .Saeta, P. Ejarque, and J. Hernando. Agatha: Multimodal Biometric Authentication Platform in Large-Scale Databases. *Securing Electronic Business Processes, ISSE 2008*, pages 186-193, Viewe+Teubner, October, 2008.
- P. Ejarque and J. Hernando. Score Equalization in SVM Multimodal Fusion for Person Recognition. E-business and Telecommunications, *Communications in Computer and Information Science*, 23, pages 152-161. ISBN 978-3-540-88652-5. Springer Berlin Heidelberg, 2009.
- P. Ejarque and J. Hernando. Score Bi-Gaussian Equalization for Multimodal Person Verification. *IET Signal Processing, Special Issue on Biometric Recognition*, 3(4), pages 322-332, July, 2009.
- P. Ejarque, J. Hernado, D. Hernando, and D. Gómez. Eigenfeatures and Supervectors in Feature and Score Fusion for SVM Face and Speaker Verification. In volume 5707 of *Lecture Notes in Computer Science*, pages 81-88, Springer Berlin Heidelberg, 2009.