UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

-PROYECTO FIN DE CARRERA-

CÁLCULO DEL PESO DE LA EVIDENCIA EN CASOS FORENSES DE RECONOCIMIENTO AUTOMÁTICO DE LOCUTOR EN LOS QUE EXISTEN VARIAS TOMAS DE VOZ DE PROCEDENCIA DESCONOCIDA

Eva Barriel Guitián

Diciembre 2011

Cálculo del peso de la evidencia en casos forenses de reconocimiento automático de locutor en los que existen varias tomas de voz de procedencia desconocida

AUTOR: Eva Barriel Guitián
TUTOR: Daniel Ramos Castro

Área de Tratamiento de Voz y Señales (ATVS)
Dpto. de Tecnología Electrónica y Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Diciembre de 2011

Resumen

Este Proyecto presenta el estudio, implementación y evaluación de diversas técnicas de combinación de evidencias procedentes de un sistema de reconocimiento de locutor forense cuando se introducen diferentes muestras de audio de procedencia desconocida. El objetivo es evaluar la estrategia más adecuada de todas las propuestas en el trabajo de cara a calcular una medida del peso de la evidencia forense final.

Para ello, se han desarrollado diferentes propuestas algorítmicas, la mayoría de ellas originales, divididas en dos bloques principales. El primero de ellos consiste en la combinación de evidencias mediante diferentes métodos a nivel tecnológico: Bayes ingenuo y concatenación de muestras de audio. El segundo consiste en la adaptación de redes Bayesianas para calibración y combinación de *scores* de salida de un sistema de reconocimiento automático de locutor, haciendo uso de la herramienta Hugin Expert.

Todas las estrategias propuestas se han probado sobre una base de datos forense real en español, AHUMADA III, cuyas grabaciones provienen de terminales GSM y presentan una gran variabilidad en cuanto a entornos, estados emocionales, ruido ambiental, etc.

Por último se realiza un análisis de los resultados para la obtención de conclusiones finales y el planteamiento del posible trabajo futuro relacionado con este proyecto.

Palabras Clave

Reconocimiento automático de locutor, ciencia forense, evidencia, relaciones de verosimilitud, calibración, combinación, red Bayesiana, Hugin Expert.

Abstract

This M.Sc. Thesis presents the study, implementation and evaluation of different evidence combination techniques that come from a forensic speaker recognition system when entering different audio samples of unknown origin. The goal of this project is to evaluate the best way of combining the evidence in order to obtain its final weigth.

In order to achieve this goal, different algorithmic proposals have been carried out, separted in two main parts. The first one consists of combining evidence through different methods in a technological way: Naive Bayes and the combination of audio samples. The second one consists of the adaptation of Bayesian networks to calibrate and combine scores from an automatic speaker recognition system, using Hugin Expert tool.

All the methods proposed have been tested on a real forensic database in Spanish, AHUMADA III, which contains recordings from landlines and GSM terminals and has a great variability in environments, emotional states, noise, etc.

Finally, a general analysis of the results is carried out, in order to obtain final conclusions and define possible future work related with this project.

Key words

Automatic speaker recognition, forensic science, evidence, likelihood ratio, calibration, combination, Bayesian network, Hugin Expert.

Agradecimientos

En primer lugar quisiera agradecer a Daniel Ramos la oportunidad que me ha brindado para realizar este proyecto en el ATVS y aprender de él. Muchas gracias por su tiempo, dedicación, apoyo y sobre todo, paciencia.

Mi mayor agradecimiento, como no podía ser de otra manera, a mis padres por su incansable esfuerzo y apoyo durante estos años, por animarme a seguir adelante y aguantarme en los malos momentos, por estar siempre tan orgullosos de mí, porque sin su ayuda todo habría resultado mucho más difícil. Os quiero.

A mi hermano por estar siempre, en cada situación, cuando lloro, cuando río, cuando lo necesito, por ser el mejor hermano del mundo. También a Laura, por transmitirme esa alegría y buen rollo en todo momento.

A mi Tata, mi "hermana" mayor, por esos momentos geniales en los que no podemos parar de reír y por el cariño y apoyo incondicional que junto a Javi me demuestran.

A Eugenio, por esas risas en el laboratorio, momentos de quejas y gabinetes de crisis, por ser un amigo de verdad. A Pi, por su buen humor y disposición.

A todos mis compañeros y amigos de la universidad, y por supuesto, posteriores y grandísimos fichajes. Gracias por las risas en los buenos momentos, y el apoyo en los malos. Sin dudarlo, sois lo mejor que me llevo de estos años.

A Tono, por ser de las pocas personas con las que sé con certeza que siempre puedo contar.

A Dupa, una de las personas más importantes de mi vida, por todos estos años, por estar ahí siempre que lo necesito, por cada día de apoyo, por su gran paciencia para aguantar mis momentos de agobio y mal humor y por la nueva vida que comenzamos juntos.

.

Índice de Contenidos

Índice de Figuras

Índice de Tablas

1

Introducción

1.1 Motivación

El reconocimiento automático de locutor puede definirse como el uso de máquinas con el objetivo de reconocer personas a partir de sus voces [1]. Debido a la accesibilidad y aceptabilidad de la voz como rasgo biométrico, el rango de posibles aplicaciones de los sistemas de reconocimiento automático de locutor es más amplio comparado con otros rasgos. Algunas de las aplicaciones más comunes de este tipo de sistemas son: control de acceso, aplicaciones telefónicas, comercio electrónico, domótica o sistemas forenses. Esta última aplicación es sobre la que versa el presente proyecto.

En investigaciones criminales que involucran grabaciones de voz, se suelen tener dos tipos de muestras: grabaciones incriminatorias (dubitadas) procedentes de micrófonos ocultos, pinchazos telefónicos, llamadas anónimas etc. y grabaciones del sospechoso (indubitadas), tomadas en dependencias policiales o de las que se conoce su autoría. El grado de similitud entre las muestras de voz dubitada e indubitada representa la **evidencia forense**.

Para medir esta similitud entre ambas grabaciones, se puede hacer uso de sistemas automáticos de reconocimiento de locutor, cuya salida será un *score* o puntuación. Sin embargo, la respuesta a la pregunta sobre si ambas muestras pertenecen a la misma fuente no puede ser SÍ o NO como en sistemas comerciales, sino que debe ser un dato probabilístico extraído a partir de toda la información disponible sobre el caso. Para ello el sistema forense debe convertir los *scores* en valores interpretables por los Tribunales.

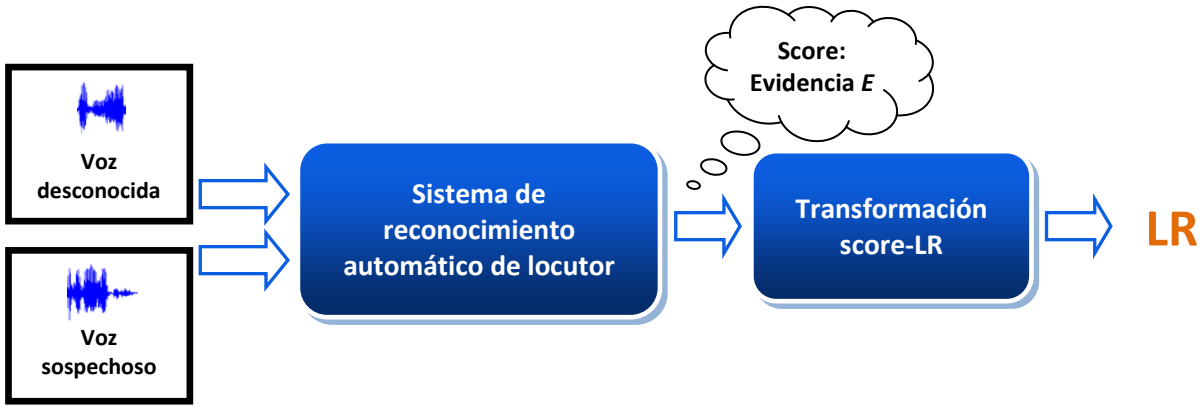


Figura 1.1: *Sistema de reconocimiento de locutor forense.*

El cálculo de relaciones de verosimilitud (*likelihood ratio* – LR), propuesto para sistemas de reconocimiento de locutor por [2], hace uso de la teoría bayesiana de la probabilidad y está fuertemente establecida como marco teórico aplicable a cualquier disciplina forense [3]. Con el cálculo de LRs, se determina cuánto más probable es el hecho de observar la evidencia si se asume que la muestra de voz incriminatoria pertenece al individuo sospechoso que si se asume la hipótesis contraria, por tanto es una medida del peso de la evidencia[1] forense.

En este marco bayesiano, los papeles del científico y del Tribunal están claramente separados. El científico desarrolla los análisis de los datos e interpretación de los resultados para proporcionar el peso de la evidencia a través de LRs. Este valor, junto con las demás circunstancias del caso, ayudará al Tribunal a realizar su papel: la toma de decisión.

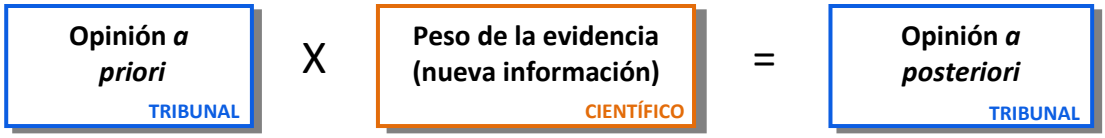


Figura 1.2: *El peso de la evidencia modifica la apuesta a priori de la hipótesis inicial, sabiendo que "el transcurso de las probabilidades a priori hacia las probabilidades a posteriori significa pasar de una valoración subjetiva de probabilidad a otra. Lo que sucede es un simple cambio de opinión como respuesta a disponer de nueva información" [4].*

Uno de los mayores problemas a la hora de evaluar el peso de la evidencia forense utilizando sistemas automáticos de reconocimiento de locutor aparece en los casos en los que diversos fragmentos de voz cuya procedencia es desconocida son objeto de análisis. La aportación total de cada uno de esos fragmentos al peso total de la evidencia de voz constituye un tema abierto de investigación muy importante en casos forenses reales. Este es el tema principal del proyecto desarrollado y en él se estudian diferentes métodos de combinación de evidencias para una posterior evaluación de los resultados.

1.2. Objetivos y metodología

Los objetivos principales que se persigue con este proyecto son:

- Estudio del estado del arte de los sistemas de reconocimiento automático de locutor basados en GMM (*Gaussian Mixture Models*) con compensación basada en JFA (*Joint Factor Analysis*), y la aplicación de dichos sistemas en entornos forenses.
- Estudio de la interpretación de la evidencia en entornos forenses mediante la teoría Bayesiana de probabilidad en forma de LRs.
- Estudio de diferentes técnicas de cálculo de LRs a partir de los *scores* generados por el sistema de reconocimiento automático de locutor.

[1] Algunos autores utilizan la expresión "peso de la evidencia" para referirse al valor del log-LR y "valor de la evidencia" para referirse al valor del LR en unidades naturales.

- Estudio, implementación e integración de diferentes métodos de combinación de la evidencia, así como el estudio de su funcionamiento en entornos forenses reales y evaluación del mejor método de combinación de la información aportada por las diferentes tomas dubitadas de cara a generar el LR final.

El diagrama de Gantt de la figura 1.2 muestra la planificación temporal seguida para la realización del proyecto. Dicho proyecto se puede dividir en 2 partes: la combinación de evidencias a nivel tecnológico en el que se incluye una formación general previa, y combinación de evidencias mediante redes Bayesianas. Cada una de ellas tiene en común las siguientes fases:

1. **Documentación ( ):** antes de comenzar a trabajar con los sistemas de reconocimiento de locutor se ha realizado una formación general al respecto mediante la bibliografía básica (publicaciones científicas y libros) sobre el estado del arte actual en biometría, técnicas de reconocimiento de locutor independientes de texto, GMMs, interpretación de la evidencia forense; y más específica sobre redes Bayesianas para la segunda parte. También se ha realizado una documentación sobre las bases de datos utilizadas (NIST y Ahumada III).
2. **Estudio del software ( ):** se ha realizado una formación sobre el entorno de experimentos implementado por el ATVS. En él, se encuentran ya desarrolladas diferentes funciones para el manejo de datos, generación de scores, compensación de variabilidad y evaluación de resultados utilizadas en este proyecto, de ahí que la primera parte se haya realizado en un tiempo considerablemente menor que la segunda. En la segunda parte, la fase de estudio del software se divide en dos (figura 1.2). La primera se refiere al estudio de características y utilidad del software de implementación de redes Bayesianas Hugin Expert, de cara a resolver el problema planteado en el proyecto. Una vez evaluadas dichas características se ha hecho un estudio tanto de su entorno gráfico como de su API para C++.
3. **Investigación e implementación ( ):** se ha realizado un estudio e implementación de combinación de evidencias forenses mediante diferentes métodos, la mayor parte de ellos originales, prestando especial interés en la combinación mediante redes Bayesianas. Las fases 2 y 3 referentes a la parte de redes Bayesianas son, sin duda, las más laboriosas y complejas de todo el proyecto debido a que no había sido desarrollado anteriormente por el grupo. En la figura 1.2 esta fase aparece dividida en dos para las cada una de las partes del proyecto, se refiere a los dos tipos de experimentos realizados en cada una de ellas.
4. **Experimentos y conclusiones ( ):** en esta fase se lleva a cabo la ejecución de los experimentos. Cabe destacar que en la figura 1.2 se considera el tiempo de ejecución de los experimentos de la primera parte debido al alto tiempo de computación requerido. Posteriormente se ha realizado un análisis de los resultados obtenidos en las pruebas llevadas a cabo así como una comparativa entre los diferentes métodos utilizados según los resultados arrojados por las mismas. En la figura 1.2 la primera barra representa la generación de gráficas y la segunda la generación de un informe parcial con las conclusiones extraídas.
5. **Elaboración de la memoria ( ):** a partir de los resultados, junto con la revisión del estado del arte y un estudio completo del proyecto llevado a cabo, se ha elaborado la memoria con la que se finaliza el proyecto fin de carrera.

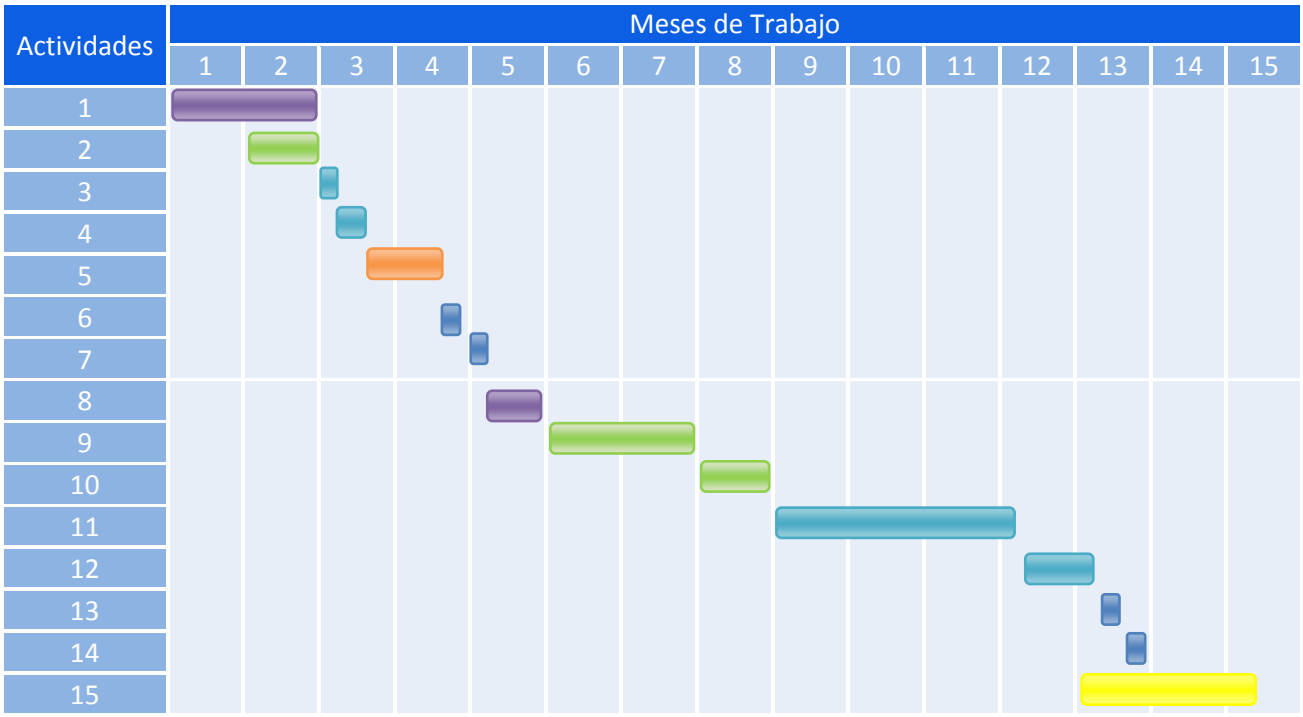


Figura 1.3: *Diagrama de Gantt del proyecto.*

1.3 Organización de la memoria

Este proyecto está organizado según se describe a continuación:

Capítulo 1. Introducción
Este capítulo presenta la motivación para la realización de este proyecto y los objetivos que se persiguen durante el desarrollo del mismo.

Capítulo 2. Introducción a la biometría
Este capítulo empieza con una introducción a la biometría donde se revisa la literatura al respecto y se comparan diferentes rasgos biométricos. A continuación se describen los fundamentos de los sistemas biométricos genéricos y sus aplicaciones.

Capítulo 3. Reconocimiento de locutor
En este capítulo se comienza describiendo el funcionamiento de un sistema de reconocimiento de locutor y su posible clasificación, posteriormente se definen las características de la voz que hacen que se pueda utilizar como rasgo biométrico, las técnicas de extracción y modelado de dichas características y los factores que la limitan, dando especial importancia a los modelos de mezclas Gaussianas (GMM) con compensación JFA. Para finalizar, se introducen los sistemas de reconocimiento de locutor en entornos forenses.

Capítulo 4. Evaluación de evidencias forenses
Este capítulo comienza con una descripción de la estadística bayesiana y su aplicación a la evaluación de la evidencia forense. A continuación se dan los conceptos teóricos sobre teoría bayesiana y relaciones de verosimilitud, así como la forma de calcularlos automáticamente por distintos métodos y la evaluación de su rendimiento.

Capítulo 5. Redes Bayesianas
En este capítulo se presenta el estado del arte de las redes Bayesianas. Comienza con una breve explicación de los fundamentos básicos, se sigue con una descripción de sus posibles aplicaciones, posteriormente se presenta la importancia del uso de esta herramienta en entornos forenses y finalmente se explica su utilización para la combinación de evidencias.

Capítulo 6. Marco Experimental
Este capítulo describe los procedimientos y protocolos utilizados para el desarrollo de los experimentos en este proyecto, así como los sistemas y bases de datos utilizados.

Capítulo 7. Experimentos y resultados
En este capítulo se exponen los diferentes experimentos realizados sobre la combinación de evidencias forenses en reconocimiento forense de locutor y se exponen las diferentes conclusiones teniendo en cuenta los resultados obtenidos.

Capítulo 8. Conclusiones y trabajo futuro
Muestra las conclusiones extraídas a partir de los resultados así como las líneas de trabajo futuro relacionadas con la investigación de este proyecto.

1.4 Contribuciones

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

Contribuciones originales:

1. Métodos de combinación de evidencias:

Combinación de audio: combinación de evidencias mediante concatenación de archivos de audio procedentes de tomas dubitadas. Para ello se ha generado:

- Software en MATLAB para la combinación de evidencias forenses mediante concatenación de audio para cualquier número de combinaciones.

Redes Bayesianas: combinación de 2 evidencias mediante una red entrenada con modelado Gaussiano. Para ello se ha generado:

- Software en MATLAB para la estimación de la función de distribución bivariada necesaria para entrenar la red Bayesiana.
- Estudio del programa Hugin Expert para la utilización de redes Bayesianas en reconocimiento de locutor en entornos forenses.
- Tutorial de utilización de Hugin GUI.
- Software en C++ necesario para implementar y configurar redes Bayesianas con la herramienta de estudio Hugin Expert.

2. Métodos de calibración mediante redes Bayesianas:

PAV, GMM y modelado Gaussiano: los algoritmos utilizados en esta red Bayesiana simple son originales porque no estaban implementados aún, como es el caso de PAV o porque no estaban adaptados para su utilización en reconocimiento de locutor como es el caso del modelado Gaussiano o GMM.

- Software en MATLAB para la estimación de funciones de distribución necesarias para entrenar la red Bayesiana con la que se generan LRs a partir de *scores* de salida de un sistema de reconocimiento de locutor.

Otras contribuciones:

Bayes ingenuo: combinación de evidencias mediante suma de log-LR suponiendo independencia entre muestras. Para ello se ha generado:

- Software en MATLAB para la combinación de evidencias forenses mediante suma para cualquier número de combinaciones.

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

- Función en Matlab para realizar validación cruzada.

2

Introducción a la biometría

La identificación personal se ha basado tradicionalmente en la posesión de llaves, tarjetas, etc. o en el conocimiento de claves de palabras o números. Sin embargo, este tipo de autenticación posee algunas limitaciones: lo que se posee puede perderse o ser sustraído y lo que se conoce se puede olvidar, confundir o asociar a datos externos. Por esto, no se puede distinguir al usuario del impostor con posesión y/o conocimiento del medio.
La biometría surge con el objetivo de autenticar de forma segura la identidad de las personas, donde se utilizan características que son propias de cada individuo, como la voz, huella dactilar, rostro, etc.

2.1. ¿Qué es la biometría?

La biometría se puede definir como la ciencia que se ocupa de reconocer a un individuo basado en sus características físicas o conductuales [5]. Estas características se conocen como rasgos biométricos.

En función de los rasgos biométricos utilizados para la identificación, se pueden establecer dos tipos de biometría diferentes:

- **Biometría estática:** se refiere al estudio del conjunto de características físicas. Dentro de este tipo se encuentran entre otros: huellas dactilares, geometría de la mano/dedo, iris, ADN, etc.
- **Biometría dinámica:** se refiere al estudio del conjunto de características conductuales. Dentro de este tipo se tiene: Voz, firma, modo de teclear, modo de andar, etc.

2.2. Propiedades de los rasgos biométricos

Según diversos autores [6][7], para que los rasgos biométricos puedan ser utilizados como elementos de identificación deben cumplir una serie de características.

- **Universalidad:** toda persona debe poseer dicho rasgo biométrico.
- **Unicidad:** personas distintas deben poseer rasgos distintos.
- **Permanencia:** el rasgo debe ser invariante a lo largo de un periodo de tiempo aceptable.

- **Perennidad:** el rasgo debe ser perpetuo, es decir, invariante con el tiempo a lo largo de la vida de la persona.
- **Mensurabilidad:** el rasgo debe poder ser caracterizado cuantitativamente.
- **Aceptabilidad:** grado de aceptación personal y social.
- **Rendimiento:** el nivel de precisión y rapidez en la identificación debe ser elevado.
- **Evitabilidad:** es necesario que sea robusto frente a posibles ataques externos.

Estas características son ideales, ya que no es posible que se cumplan todos los requisitos en cada uno de los rasgos, por lo tanto, la selección de un rasgo específico para una aplicación particular está condicionada por las características concretas del mismo y los requisitos de la aplicación.

2.3. Comparación de varios rasgos biométricos

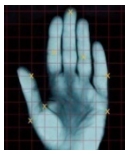
ADN: el ADN (ácido desoxirribonucleico) está presente en toda célula viva y es único para cada individuo excepto para gemelos monocigóticos, lo que le otorga un alto poder discriminante. De hecho, es el método más utilizado en aplicaciones de reconocimiento en ámbitos forenses. Sin embargo, su principal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

Cara: el rostro es probablemente el rasgo biométrico más usado en el reconocimiento humano entre individuos y supone un método de reconocimiento no invasivo. Las aproximaciones para el reconocimiento facial se basan bien en la localización y forma de los atributos faciales como ojos, nariz, labios y barbilla junto con su relación espacial (análisis local), o bien en un análisis global de la imagen de la cara. Las mayores limitaciones consisten en la forma de adquisición de las imágenes, requiriendo a veces un fondo fijo y simple o una iluminación especial, y en los problemas de reconocimiento de imágenes capturadas desde diferentes ángulos y bajo diferentes condiciones de iluminación [5].

Huella dactilar: una huella dactilar es el patrón de crestas y valles de la superficie de los dedos. Es el rasgo biométrico de mayor utilización debido a su gran poder discriminativo. Su uso está muy extendido en aplicaciones comerciales, pero también en ámbito forense, en el que se trata de identificar criminales que dejaron sus huellas en la escena del crimen. Algunas ventajas sobre otros rasgos biométricos son la invariabilidad con la edad, la regeneración y la facilidad de adquisición.

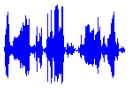
Iris: es la membrana coloreada y circular del ojo que rodea la pupila. La estructura del iris de cada ojo muestra alto grado de unicidad y estabilidad con el tiempo. El patrón se mantiene prácticamente invariante desde la infancia del individuo. Su captura se realiza mediante imágenes, donde la iluminación y la cooperación del usuario son determinantes. Por ello se considera un método intrusivo pero con un alto potencial debido a la rapidez de los sistemas y al alto poder de discriminación que ofrece.

Firma: la forma de firmar de cada persona es característica. Aunque requiere contacto con una superficie y la cooperación del usuario, es un rasgo muy aceptado como método de autenticación ya que se usa ampliamente en cantidad de transacciones. La firma varía a lo largo del tiempo para un mismo individuo y está influenciado por su estado físico y emocional. Además existen sujetos cuya firma varía muy significativamente en cada realización, por lo que su identificación es compleja.



Geometría de la mano: Los sistemas de reconocimiento para este rasgo se basan en un conjunto de medidas físicas como la forma de la mano, el tamaño de la palma y la longitud y el ancho de los dedos. Los factores ambientales no suponen un problema pero la geometría de la mano es un rasgo de baja distintividad de cada individuo y está sujeto a cambios a lo largo de la vida de una persona.

Termogramas: El patrón de calor radiado por el cuerpo humano es característico de cada individuo. Puede ser capturado por una cámara de infrarrojos de forma no intrusiva o incluso oculta. La mayor desventaja de esta clase de sistemas es el coste de los sensores y su vulnerabilidad ante otras fuentes de calor no controlables. Los termogramas también se emplean para captar la estructura de las venas de la mano.



Voz: la voz es una combinación de características físicas y conductuales. Las características físicas del habla de cada individuo permanecen invariantes, pero las características de conducta cambian a lo largo del tiempo y se ven influenciadas por la edad, las afecciones médicas o el estado de ánimo de la persona. Las principales desventajas de este rasgo son su baja distintividad y la facilidad con la que puede ser imitado. Por el contrario, la voz es un rasgo biométrico muy aceptado y fácil de obtener. La voz es el rasgo biométrico en que se centra este proyecto, por lo que se hablará de ella con más detalle en el capítulo 3.

Además de éstas, existen otras modalidades biométricas emergentes como la forma de la oreja, patrón vascular de la retina, modo de andar, dinámica de tecleo, olor o radiografías dentales. Más información sobre estos rasgos se puede encontrar en [9][10].

2.4. Sistemas biométricos

Un sistema biométrico es esencialmente un sistema de reconocimiento de patrones. Funciona mediante la adquisición de información biométrica de un individuo (rasgos biométricos), la extracción de un conjunto de características a partir de la información adquirida, y por último la comparación con un conjunto de características almacenadas en una base de datos [5].

En esta sección se presenta la estructura general de un sistema de reconocimiento, acompañado de una breve explicación sobre su funcionamiento, seguido de las aplicaciones principales de los sistemas biométricos. Se reservará un capítulo completo a un estudio más exhaustivo sobre este tipo de sistemas aplicados a reconocimiento de locutor, objeto del presente proyecto.

2.4.1. Estructura general de un sistema automático de reconocimiento biométrico

Todos los sistemas de reconocimiento automático de patrones poseen una estructura funcional común formada por varias fases cuya forma de proceder depende de la naturaleza del patrón o señal a reconocer. La figura 2.1 muestra un diagrama de bloques representativo de las diferentes etapas que constituyen un sistema biométrico, a continuación se describe cada una de ellas.

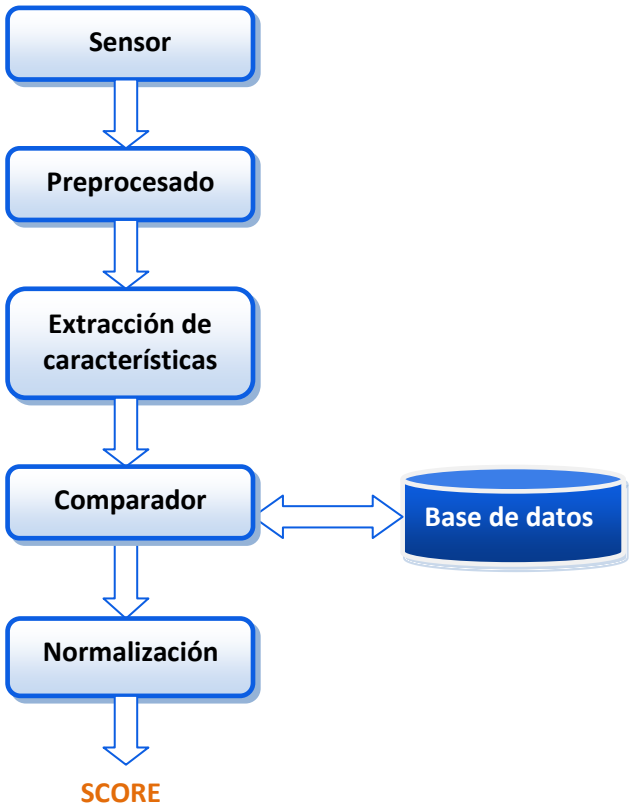


Figura 2.1: *Arquitectura general de un sistema biométrico.*

- Adquisición de datos
En esta etapa, se adquiere la información biométrica de un individuo a través de un sensor. Este proceso es determinante ya que de él depende la cantidad y la calidad de la información adquirida, la implementación de las siguientes fases, y por tanto, el resultado final que se obtiene.

- Preprocesado
En algunos casos es necesario acondicionar la información capturada para eliminar posibles ruidos o distorsiones producidas en la etapa de adquisición, o para normalizar la información a unos rasgos específicos para tener una mayor efectividad en el reconocimiento posterior.

- Extracción de características
La información adquirida en la etapa anterior se procesa para extraer un conjunto de características discriminantes.

- Generación de modelos y cálculo de similitud
Una vez extraídas las características más significativas, se elabora un modelo que represente a cada individuo y se almacena en una base de datos. Posteriormente se evalúa la similitud entre

éstos y los patrones de entrada y se calcula un *score* que supone una medida cuantitativa del parecido entre ambas muestras.

2.4.2. Aplicaciones de los sistemas biométricos

Las aplicaciones de los sistemas biométricos se pueden dividir en los siguientes tres grandes grupos [11]:

- **Aplicaciones comerciales:** protección de datos electrónicos, protección en red, e-comercio, cajeros automáticos, control de acceso físico, etc.
- **Aplicaciones gubernamentales:** DNI, carné de conducir, pasaporte, control en fronteras, etc.
- **Aplicaciones forenses:** identificación de cadáveres, investigación criminal, identificación de terroristas, determinación de parentesco, etc.

El estudio de este proyecto se basa en este último tipo de aplicaciones, en concreto, la investigación criminal a partir de un sistema de reconocimiento de locutor.

3

Reconocimiento automático de locutor

El reconocimiento automático de locutor representa actualmente una tarea clave dentro del reconocimiento biométrico [12]. Esto es debido principalmente a dos factores:

- La autenticación de la voz es una tecnología versátil, sencilla de usar y no intrusiva, es decir, fácilmente aceptada por los usuarios.
- No requiere el uso de aparatos específicos, basta con un teléfono. Además gracias a la red telefónica se puede hacer uso del mismo desde prácticamente cualquier punto del planeta.

En este capítulo se comenzará describiendo el funcionamiento de un sistema de reconocimiento de locutor y su posible clasificación, posteriormente se definirán las características de la voz que hacen que se pueda utilizar como rasgo biométrico, las técnicas de extracción y modelado de dichas características y los factores que la limitan, para acabar introduciendo los sistemas de reconocimiento de locutor en entornos forenses.

3.1. Estructura y funcionamiento de un sistema genérico

El objetivo en los sistemas de reconocimiento automático de locutor es conseguir métodos y procedimientos fiables que sin supervisión humana, sean capaces de tomar decisiones acerca de la identidad hablada. Para ello, dichos sistemas normalmente trabajan en dos fases:

1. **Fase de entrenamiento:**

   En esta fase se obtiene la información necesaria en forma de patrones, plantillas o modelos, que se usarán como valor de referencia correspondiente a cada uno de los usuarios del sistema. En la figura 3.1 se puede ver un diagrama de bloques correspondiente a esta fase.

2. **Fase de cálculo de la similitud:**

   Esta será la fase del sistema donde, a partir de nuevas señales habladas, se genera un modelo estadístico y se compara con modelos registrados en la base de datos. A partir de esta comparación se calcula un *score* o puntuación con el que sistema tomará decisiones acerca de la identidad del hablante. En la figura 3.2 se presenta el diagrama

de bloques de un sistema de reconocimiento genérico de locutores en esta fase de funcionamiento.

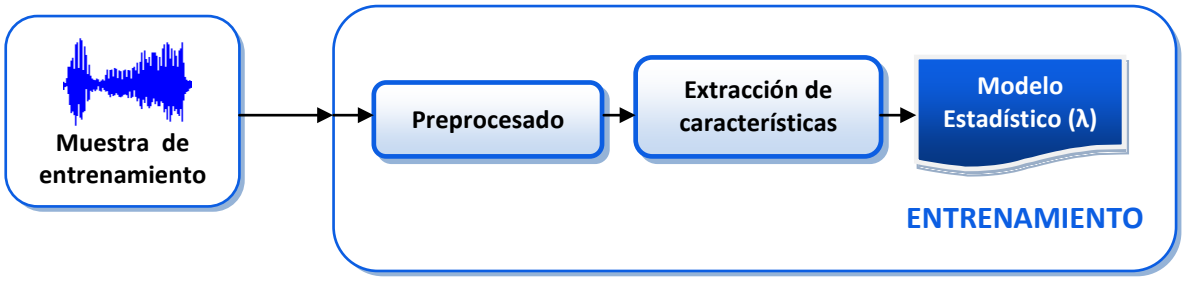


Figura 3.1: *Esquema general de un sistema de reconocimiento de locutor en modo entrenamiento.*

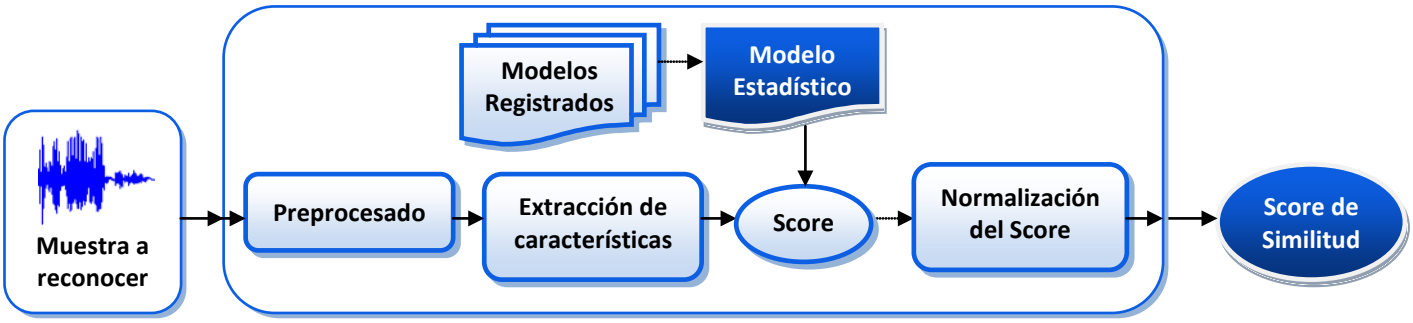


Figura 3.2: *Esquema general de un sistema de reconocimiento de locutor en fase de cálculo del score.*

3.2. Clasificación de sistemas

Los sistemas de reconocimiento de locutor pueden ordenarse en base a diferentes criterios, siendo los dos más importantes: la tarea específica a realizar por el sistema y la dependencia del sistema respecto al texto pronunciado.

Por un lado, según la tarea a realizar se tiene [13]:

- **Modo verificación ¿es esa persona X?**

  Una persona reclama tener una determinada identidad y el sistema debe verificar que es cierto. En los sistemas de verificación se dispone de una realización de voz a verificar y una solicitud de identidad que puede ser realizada de diversas formas: lectura de tarjeta magnética individual, introducción de un código mediante teclado o voz etc. Por lo tanto, se compara únicamente el modelo de la voz capturada, con el modelo de la persona quien dice ser almacenada en la base de datos. De este modo, las dos únicas salidas posibles son la aceptación o rechazo del individuo, dependiendo de la comparación con el umbral o regla de decisión.

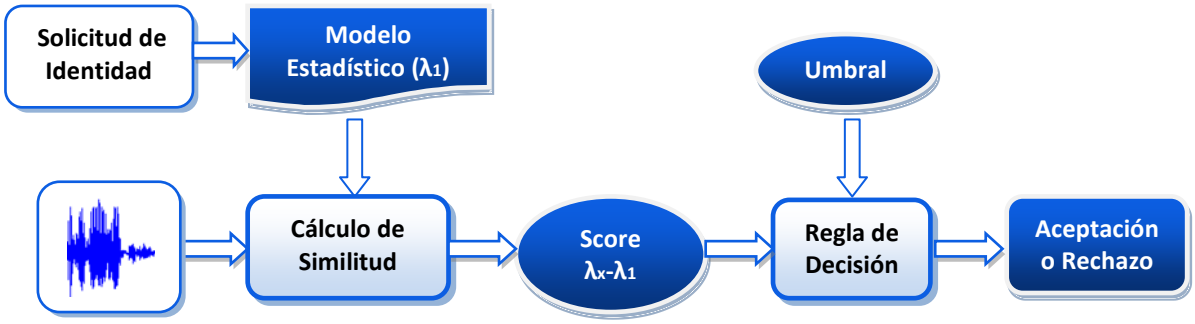


Figura 3.3: *Esquema general de un sistema de reconocimiento de locutor en modo verificación.*

- **Modo identificación ¿quién es la persona X?**

  Se tiene información sobre una persona de la que se desconoce su identidad. Para proceder a la identificación de una persona se debe comparar el modelo de la persona a identificar con los N modelos de las personas almacenadas en la base de datos [10]. Dentro de estos sistemas se pueden distinguir dos posibles casos:

  - **Identificación en conjunto cerrado:** el resultado del proceso es una asignación de identidad a uno de los individuos modelados por el sistema, y conocidos como usuarios. Existen tantas decisiones de salidas diferentes como usuarios registrados en el sistema.

  - **Identificación en conjunto abierto:** aquí se debe considerar una posibilidad adicional a las del caso anterior: que el individuo que pretende ser identificado no pertenezca al grupo de usuarios, con lo que el sistema de identificación debería contemplar la posibilidad de no clasificar la realización de entrada.

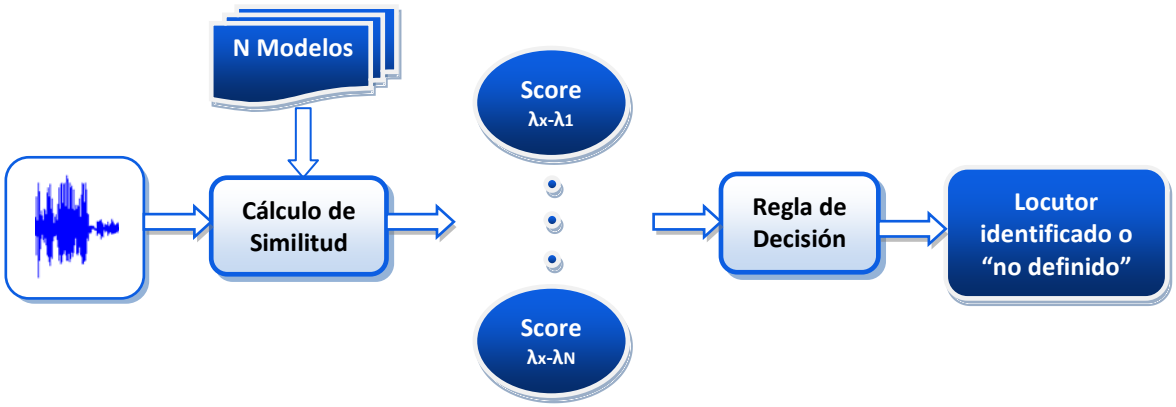


Figura 3.4: *Esquema general de un sistema de reconocimiento de locutor en modo identificación.*

- **Sistemas forenses**

  En este caso, se posee una muestra de voz de procedencia desconocida y unas muestras de referencia de un sospechoso obtenidas de manera controlada con las que se generará su modelo. Un Tribunal requiere la opinión de un experto para comparar las muestras. Para ello se calcula la similitud entre ambas muestras dando lugar a un *score* o puntuación que representará la evidencia forense. Para que los resultados sean

interpretables por el Tribunal, la puntuación obtenida debe transformarse en un LR, que representa el peso de la evidencia forense. Este método será descrito con detalle en el capítulo 4.

Por otro lado, según la dependencia con el texto pronunciado se tiene:

- **Sistemas dependientes de texto**

  La locución de entrenamiento y la de prueba son idénticas, de forma que lo que tiene que hacer el sistema es establecer una comparación entre realizaciones diferentes de una misma palabra o frase, compensando la variabilidad entre ambas mediante el alineamiento temporal de las secuencias de características. Este alineamiento puede llevarse a cabo de varias formas; algunas de estas técnicas son: el alineamiento temporal dinámico (Dynamic Time Warping, DTW), basada en modelos de plantilla, y los modelos ocultos de Markov (Hidden Markov Models, HMM), basada en modelos estocásticos.

- **Sistemas independientes de texto**

  En los sistemas independientes de texto, la locución a pronunciar en la fase de entrenamiento y la de reconocimiento no coinciden. Este tipo de sistemas han sido los claros dominantes en el reconocimiento de locutor durante las últimas décadas, especialmente los basados en características espectrales a corto plazo (o sistemas acústicos), ya que permiten modelar una mayor cantidad de observaciones acústicas y por lo tanto recoger más variabilidad intra-locutor. En la última década se han desarrollado también sistemas basados en características de alto nivel, pero su rendimiento está aún bastante lejos del de los sistemas acústicos [15]. Sin embargo, han sido empleados con éxito en reconocimiento de idioma, principalmente mediante su fusión con sistemas acústicos. Algunas de las técnicas utilizadas son GMM (Gaussian Mixture Models) o SVM (Support Vector Machine).

  El proyecto se centra en el reconocimiento de locutor en entornos forenses, donde no se conoce el contenido lingüístico de las muestras por lo que se engloba dentro de este tipo de sistemas. Por tanto los siguientes apartados se centrarán en las técnicas más empleadas en reconocimiento de locutor independiente de texto.

3.3. Identidad hablada y sistemas de reconocimiento

La principal función asociada con la transmisión de una señal de voz es transmitir un mensaje. Sin embargo, también se transmite otro tipo de información como son el género, el idioma, la identidad del locutor, estado emocional y de salud etc. Estas características de la voz vienen determinadas por la fisiología (la longitud del tracto vocal, su forma, y la configuración dinámica de los órganos articulatorios) y comportamiento (hábitos lingüísticos, entonación de las frases etc.) del hablante. El ámbito de reconocimiento de locutor se centrará, por tanto, en las características presentes en la señal de voz que individualizan al locutor [10]. Para modelar estas características, se tienen en cuenta diferentes niveles de información, representados en la figura 3.5. Ordenados de más bajo a más alto nivel se tiene:

- **Nivel acústico o espectral:** relativas a la realización de sonidos individuales. Aquí están incluidos el tamaño de las cavidades fonatorias que darán lugar a frecuencias de resonancia y anchos de banda de formantes característicos, el tamaño de las cuerdas

vocales que afectarán a los valores de frecuencia fundamental generados, y las peculiaridades articulatorias de los distintos sonidos.

- **Nivel fonético:** relativas al uso diferente que hace cada persona de los fonemas y sílabas.
- **Nivel prosódico:** la prosodia es una combinación de energía, entonación, duración y tono de los fonemas. Es responsable de dotar a la voz de naturalidad y sentido.
- **Nivel lingüístico:** características que describen la forma en que el locutor hace uso del lenguaje, que se ven influenciadas por aspectos como la educación, el origen y las condiciones sociológicas del hablante.

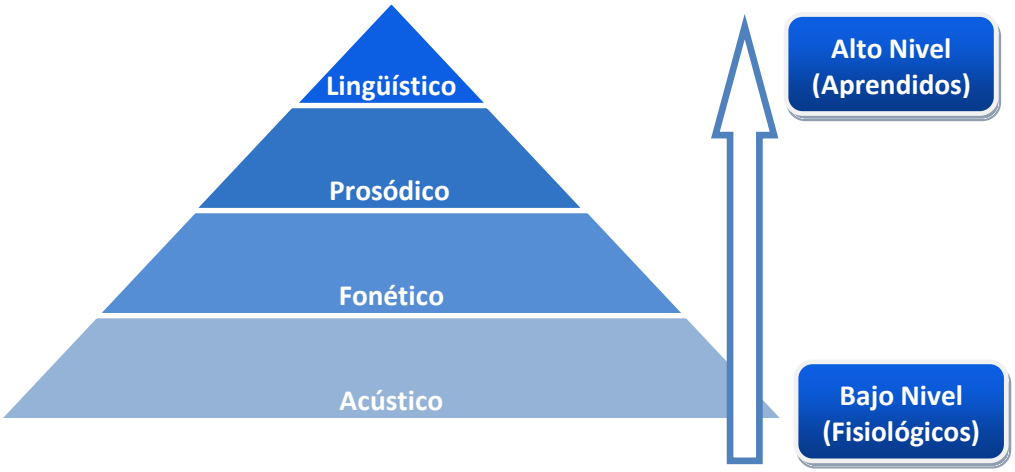


Figura 3.5: *Niveles de identidad en la señal de voz.*

3.3. Parametrización

La parametrización de la señal del habla consiste en transformar esta señal en un conjunto de vectores de características. Esta transformación permite obtener una representación más compacta, menos redundante y más útil de cara al modelado estadístico y al cálculo de puntuaciones.

3.3.1. Sistemas acústicos:

Los sistemas de reconocimiento de locutor más comúnmente utilizados se basan en las características acústicas de la señal. El modelo de producción de voz que se emplea habitualmente contempla dos componentes en la señal de voz. Por un lado se encuentra la excitación, que es responsable de la estructura fina del espectro y cuya utilidad es limitada a la hora de diferenciar unos sonidos frente a otros. Por otro lado, se encuentra la componente que se relaciona con la articulación de los sonidos. La articulación determina la forma de la envolvente espectral de la señal de voz y, por tanto, una buena parametrización debe ser capaz de estimar dicha envolvente a la vez que no se ve afectada por los detalles que porta la excitación.

Los parámetros MFCC estiman la envolvente espectral desechando los parámetros cepstrales que provienen de la excitación de la señal de voz y, de este modo, únicamente retienen aquellos coeficientes que representan la envolvente espectral.

La figura 3.6 representa la metodología que se sigue para realizar el análisis de la señal de voz que está basada, habitualmente, en el análisis a corto plazo. En dicha figura se ve como, por medio de ventanas, se seleccionan segmentos temporales donde la señal puede suponerse cuasi-estacionaria. A partir de cada ventana se obtienen los parámetros de interés.

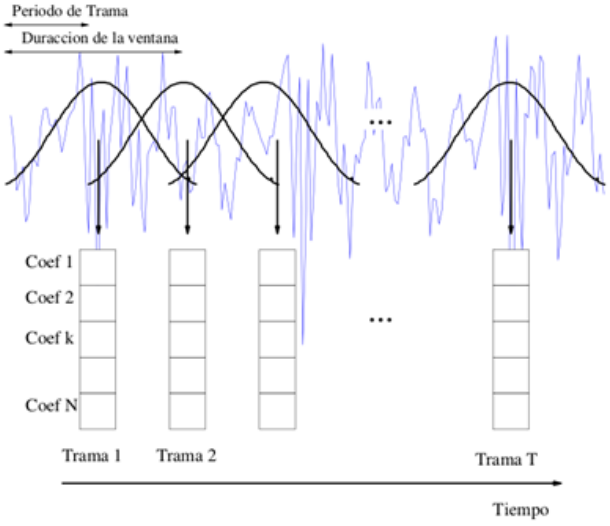


Figura 3.6: *Enventanado de la señal de voz para la posterior extracción de características.*

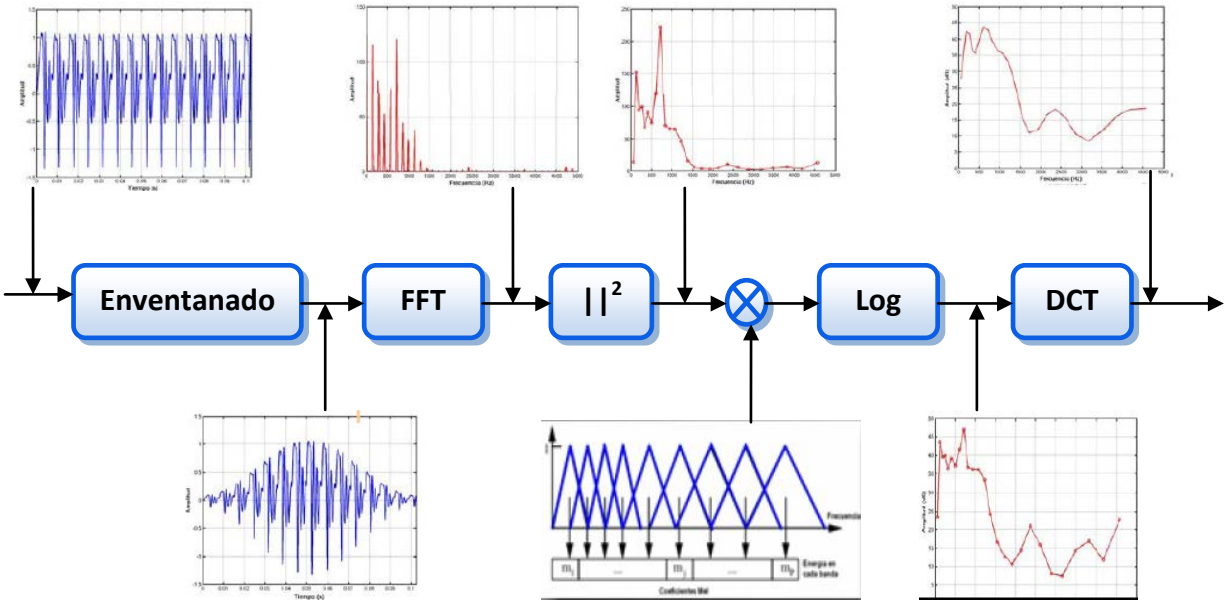


Figura 3.7: *Extracción de características MFCC*

Enventanado: típicamente se divide la locución en ventanas de 20 ms con solapamiento del 50% (10 ms) a través de ventanas de tipo hamming.

Análisis espectral: se calcula el espectro a corto plazo realizando la transformada de Fourier de la señal de voz enventanada. Además, ya que la fase no aporta información relevante para la discriminación de los sonidos, se calcula o bien el módulo del espectro o bien su densidad espectral de potencia.

Banco de filtros escala Mel: en esta etapa se aplica un banco de filtros al módulo (o a la potencia) del espectro obtenido en la etapa anterior. A través de este banco de filtros se simula el comportamiento del sistema auditivo, ya que estos filtros están distribuidos de forma no uniforme a lo largo del eje de frecuencias aportando mayor resolución a bajas frecuencias que a altas.

Log (Logaritmo): a través de este operador no lineal la señal de voz se divide en sus dos componentes principales: la excitación y la envolvente espectral. Además, este logaritmo permite simular el comportamiento del oído humano y su sensibilidad ante distintas intensidades de presión sonora.

DCT (Discrete Cosine Transform): las log-energías en banda obtenidas en la etapa anterior están altamente correladas y por medio de esta transformación conseguimos coeficientes más incorrelados. Además, la DCT concentra en los coeficientes bajos las componentes que provienen de la envolvente espectral y en los altos los que provienen de la excitación.

Una variante de esta técnica es LPCC, que utiliza codificación de predicción lineal en lugar de filtros MFCC [16].

3.3.2. Sistemas fonéticos:

Su finalidad es modelar el uso de los fonemas. Están compuestos de dos bloques: el primero se encarga de identificar los fonemas en cada locución. Una vez identificados los fonemas el segundo bloque se encargará de hacer un modelo estadístico del lenguaje del locutor en base al tipo de fonemas y la frecuencia con la cual los utiliza.

3.3.3. Sistemas prosódicos:

Se basan fundamentalmente en el análisis de dos factores:

- La frecuencia fundamental (pitch). Viene determinada por la frecuencia de la vibración de las cuerdas vocales, y es la responsable del tono del habla en cada individuo. En [17] se pueden encontrar distintos métodos para el cálculo del “pitch”.

- La energía y duración de los sonidos, fáciles de extraer por distintos métodos.

Al igual que los sistemas fonéticos, se componen de dos bloques: el primero se encarga de extraer los parámetros citados. El segundo realiza un modelado estadístico a partir de estos.

3.4. Construcción de modelos

Una vez extraídas las características más significativas, se elabora un modelo que represente a cada individuo y se almacena en una base de datos. Posteriormente se evalúa la similitud entre éstos y los modelos de entrada, obteniendo así la puntuación necesaria para la toma de decisiones. En esta sección se verán dos métodos, uno generativo en el que se modela la estructura o distribución estadística de los datos (GMM) y otro discriminativo en el que lo que se modela son fronteras entre regiones que representan diferentes clases (SVM).

3.4.1. GMM (Gaussian Mixture Models)

Los sistemas basados en modelos de gaussianas (GMM) consisten en modelar el habla a partir de distribuciones gaussianas, basándose en que la distribución de los parámetros de tipo perceptivo (como pueden ser en este caso los coeficientes cepstrales obtenidos en la etapa anterior) se aproxima a la de una mezcla de gaussianas. De modo que a cada usuario o población de interés le corresponderá un modelo λ, cuya función densidad de probabilidad viene dada por [18]:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i \, p_i(x) \tag{3.1}$$

donde M es el número de gaussianas componentes, x es el vector de características de dimensión d a observar, w_i son los pesos de cada una de las mezclas donde debe cumplirse $\sum_{i=1}^{M} w_i = 1$ y $p_i(x)$ son las M densidades componentes, que a su vez se descomponen en:

$$p_i(x) = \frac{1}{2\pi^{d/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \tag{3.2}$$

siendo μi los vectores de medias de dimensión dx1 y Σ_i las matrices de covarianzas de dimensión dxd. Debido a que estimar una matriz completa requiere, en general, un alto coste computacional y gran cantidad de datos de entrenamiento se suelen utilizar matrices de covarianza diagonales [19]. En la figura 3.8 está representado gráficamente un ejemplo GMM de 3 gaussianas.

El entrenamiento del GMM consistirá por tanto en estimar los parámetros del modelo λ $\{p_i(x), \mu_i, \Sigma_i\}$ a partir de un conjunto de vectores X del locutor a modelar. Suele realizarse mediante estimaciones de máxima verosimilitud (ML-Maximum Likelihood) a través del algoritmo estimación-maximización (EM-expectation maximitation). Para ello, el algoritmo EM modifica iterativamente los parámetros del GMM con respecto a los datos de entrenamiento, de manera que para la iteración k y $k+1$, $p(x|\lambda') > p(x|\lambda)$. El algoritmo se repite hasta que el valor de probabilidad converge o se alcanza un número de iteraciones máximo. Por otra parte, puede usarse el algoritmo K-means como método de inicialización del algoritmo EM, de forma que se necesiten menos iteraciones para su convergencia.

Para generar cualquier *score* de similitud se utilizarán dos modelos estadísticos. Por un lado se tendrá un modelo de habla universal (UBM-*universal background model*) que representará la distribución independiente de locutor de los vectores de características, es decir, modela las características comunes a todos los locutores; por otro lado, el modelo del locutor cuya

identidad se quiere verificar, el cual se obtiene adaptando el UBM a los parámetros extraídos de las locuciones de entrenamiento de dicho locutor.

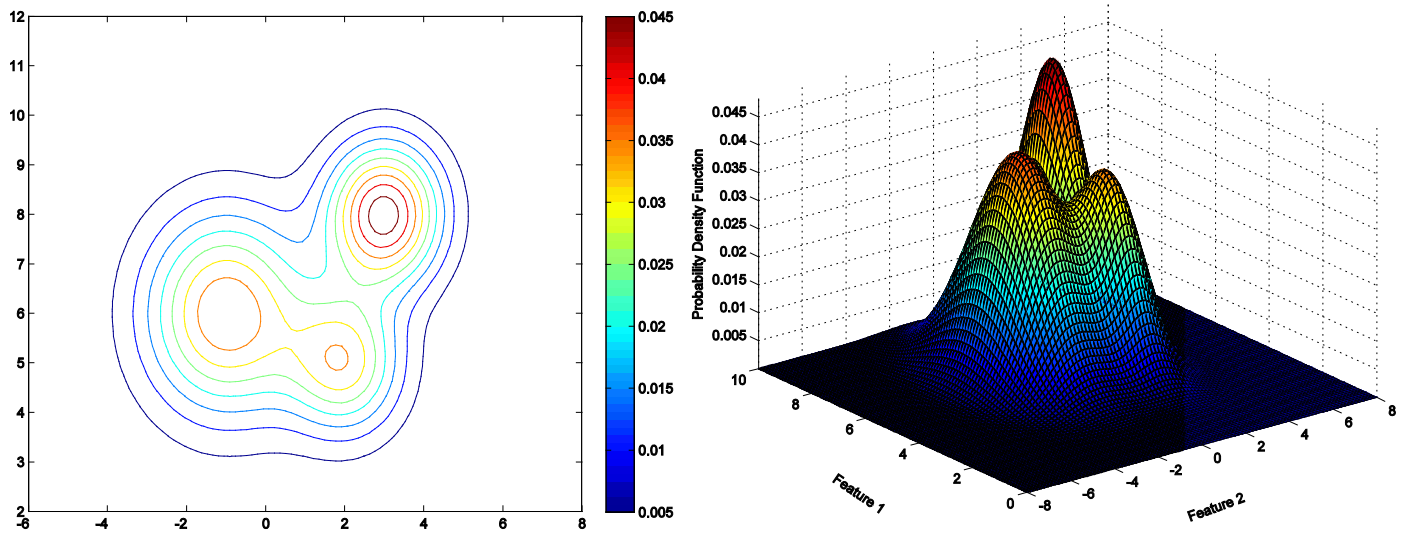


Figura 3.8: *Función densidad de probabilidad de un GMM de 3 gaussianas sobre un espacio de características bidimensional.*

Para ello, en primer lugar, se entrena por medio del algoritmo EM el modelo UBM, a partir de una gran cantidad de audio (decenas o cientos de horas) procedente de un gran número de locutores y de muy diversas condiciones acústicas. Cuando se registra a un locutor en el sistema, los parámetros del UBM se adaptan a la distribución de características del locutor, de forma que el modelo adaptado constituye el modelo de ese locutor. Por lo tanto, los parámetros del modelo no se estiman de cero, sino que se utiliza el conocimiento *a priori* dado por el UBM ("datos de voz en general").

En la adaptación de un modelo de locutor es posible adaptar todos los parámetros del UBM o sólo alguno de ellos. En [20] se demuestra que adaptando sólo los vectores de medias, se consiguen buenos resultados. Dado el conjunto de vectores del locutor a registrar, $X = \{x_1, \dots, x_T\}$, y el UBM, $\lambda\{p_i(x), \mu_i, \Sigma_i\}_{i=1}^{M}$, los vectores de medias adaptados (μ_i') por el método *maximum a posteriori* (MAP) se obtienen como sumas ponderadas de los datos de entrenamiento del locutor y de las medias del UBM:

$$\mu_i' = \alpha_i E_i(x) + (1 - \alpha_i)\mu_i \tag{3.3}$$

donde

$$\alpha_i = \frac{n_i}{n_i + r} \tag{3.4}$$

$$E_i(x) = \frac{1}{n_i}\sum_{t=1}^{T} P(i|x_t)\, x_t \tag{3.5}$$

$$n_i = \sum_{t=1}^{T} P(i|x_t) \tag{3.6}$$

$$P(i|x_t) = \frac{w_i\, p_i(x_t)}{\sum_{j=1}^{M} w_j\, p_j(x_t)} \tag{3.7}$$

El parámetro de relevancia r, y por tanto α_i, controla la influencia de los datos de entrenamiento sobre el modelo de locutor adaptado con respecto al UBM. Cuanto más grande sea r, más pequeño será α_i, y por lo tanto la influencia de los datos de entrenamiento será mayor. Esto será deseable cuando se disponga de una gran cantidad de datos de entrenamiento para un locutor concreto. En cambio, cuando apenas se disponga de datos de entrenamiento, interesará que su influencia sea menor en el modelo adaptado para que éste no termine modelando únicamente las características particulares de esos datos. Por otra parte, el proceso de adaptación puede constar de varias iteraciones MAP, sin embargo una única iteración suele ser suficiente.

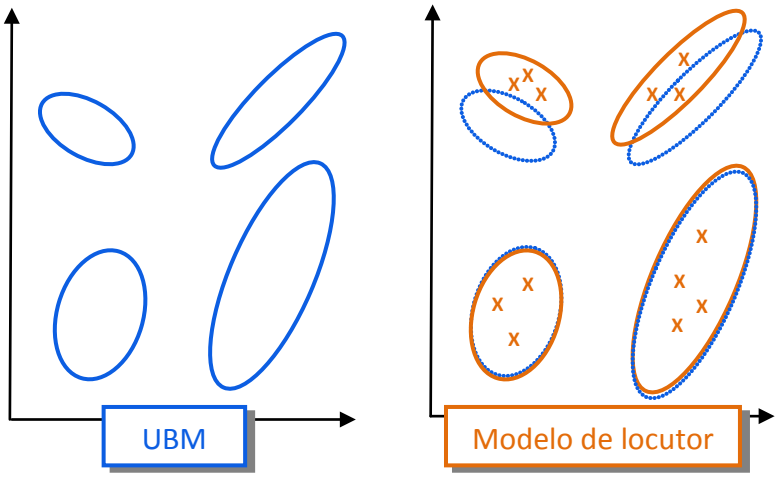


Figura 3.9: *Proceso de adaptación MAP (únicamente medias) del UBM a los datos del locutor.*

El modelo adaptado mediante esta técnica se conoce como GMM-MAP. La figura 3.9 ilustra el resultado de la adaptación de los vectores de medias de un UBM a los datos de entrenamiento de un locutor, por lo que la "forma" de las gaussianas no cambia, sólo su posición.

En el proceso de verificación se comparan las características de *test* tanto con el modelo de locutor como con el UBM, de forma que la probabilidad de que pertenezcan al locutor es proporcional a la diferencia entre las probabilidades frente a su modelo y frente al UBM (es decir, los datos de *test* deben no sólo parecerse suficientemente al modelo de locutor, sino también parecerse suficientemente poco al UBM, o lo que es lo mismo, ser más distintivos) [19]. La puntuación se obtiene mediante la relación:

$$S(X|\lambda_t) = \log p(x|\lambda_t) - \log p(x|\lambda_{UBM}) \quad (3.8)$$

donde $p(x|\lambda_t)$ y $p(x|\lambda_{UBM})$ son las funciones densidad de probabilidad para el modelo *target* y el UBM respectivamente.

3.4.2. SVM (Support Vector Machine)

Un SVM es un clasificador de patrones discriminativo en el que se modelan fronteras entre 2 regiones mediante un hiperplano. Este hiperplano representará el modelo de cada locutor calculado en la etapa de entrenamiento.

Los datos de entrenamiento serán una serie de vectores etiquetados de la forma $\{\vec{x}_i, y_i\}$, donde:

- $\vec{x}_i \in R^d$ es el vector de características de dimensión d.
- $y_i \in \{-1, 1\}$ representa las etiquetas de la clase a la que pertenece cada vector. En una tarea de verificación una de las clases consiste en los vectores de características para el entrenamiento del locutor a verificar (clase *target*, etiquetados como +1) y la otra se compone de los vectores de entrenamiento de la población de impostores (clase *non-target*, etiquetados como -1).

El problema consistirá en asignar cada vector de características a su clase correspondiente, 1 ó -1. Para ello en la etapa de entrenamiento se calculará un hiperplano de separación denotado por $\{w, b\}$ que divida el espacio R^den dos regiones. Entre todos los posibles planos de separación de las dos clases, existe sólo un hiperplano de separación óptimo, de forma que la distancia entre éste y el valor de entrada más cercano sea máxima (maximización del margen). Aquellos puntos sobre los cuales se apoya el margen máximo son los denominados vectores de soporte.

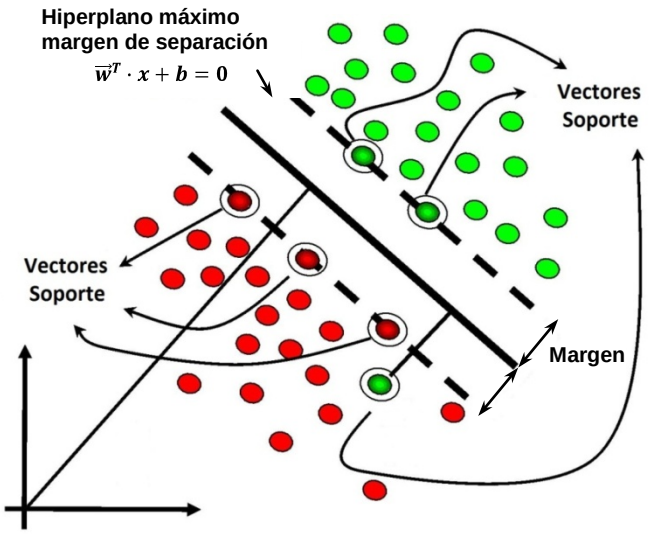


Figura 3.10: *Principio de funcionamiento de un SVM.*

El cálculo de este hiperplano supone que los vectores de ambas clases son linealmente separables, pero en la realidad no suele ser así. En el caso de que los datos no puedan ser separables por una frontera lineal en el espacio de características original R^d, se realiza una transformación a un espacio de características de mayor dimensión $R^{d'}$, donde $d' > d$, en el cual sí puedan ser separados linealmente. Esta transformación se realiza por medio de funciones de mapeo o *kernels*. En la figura 3.8. se observa el mapeo de los vectores en un espacio de características de dimensión mayor.

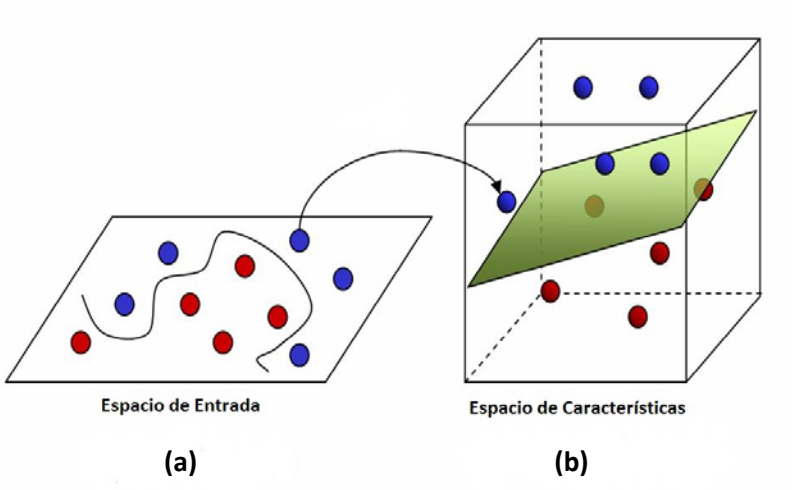


Figura 3.11: *Datos no separables linealmente (a) y datos separables linealmente en un espacio de características de mayor dimensión (b).*

Una vez se tiene definido el hiperplano de separación entre las 2 clases, el *score* de similitud será la distancia desde el vector correspondiente al fichero de *test* x_t, al hiperplano entrenado como modelo de locutor. La función que mide esta distancia se define en la ecuación 3.9. y será positiva para muestras pertenecientes a la clase +1 y negativa para las de la clase -1.

$$f(x) = \vec{w}^T \cdot x + b \tag{3.9}$$

donde w es un vector de n coeficientes que determina la orientación del plano y b es el término independiente de la ecuación paramétrica del plano.

3.5. Variabilidad

En reconocimiento de locutor, muchos de los errores cometidos por el sistema, no se deben a la diferencia entre locutores sino a la variabilidad de diferentes locuciones del mismo individuo. Este tipo de variabilidad se llama variabilidad de sesión [21]. Estos desajustes entre las condiciones del habla entre diferentes tomas de voz son causados por factores intrínsecos y extrínsecos al individuo.

- **Intrínsecos (variabilidad intra-locutor):** son aquellos que dependen sólo de características propias del sujeto como son la manera de hablar, la edad, el género, variabilidad inter-sesión, el estado emocional, etc. [22].

- **Extrínsecos (variabilidad de canal):** son los no achacables al locutor como los dispositivos de adquisición y transmisión (tipo de micrófono, distorsión de canal, etc.) y los ruidos de fondo.

Existen técnicas orientadas a minimizar el efecto de la variabilidad en diferentes dominios:

Dominio de los **parámetros**: consiste en eliminar los efectos de canal antes de modelar. Algunos ejemplos son CMN (Cepstral Mean Normalization), filtrado RASTA o FW (Feature Warping).

Dominio de los **modelos**: consiste en modificar los modelos de características para minimizar los efectos de la variabilidad. Como ejemplos: FM (Feature Mapping) y JFA (Joint Factor Analysis).

Dominio de los ***scores***: aquí la compensación se refiere a eliminar los desajustes, como escalado y desplazamiento, producidos en los *scores* debido a los efectos de la variabilidad (T-norm, Z-norm, ZT-norm, H-norm, C-norm).

A continuación se presentan las diferentes técnicas de compensación de variabilidad utilizadas en este proyecto:

- **T-Norm:** en esta técnica se compara la locución de *test* con una cohorte de modelos de impostores dando lugar a una distribución de *scores* de impostores cuya media y varianza se utiliza para la normalización. Así se consigue un alineamiento de la distribución de probabilidad de impostores dependiente del fichero de *test* a identificar.

$$S_{Tnorm} = \frac{S_{raw} - \mu_{Tnorm}}{\sigma_{Tnorm}} \tag{3.10}$$

  Está demostrado [23] que la normalización T-norm "gaussianiza" la distribución de *scores* provocando que la curva DET (sección 4.5.1) no sólo se mueva hacia valores de EER más bajas sino que también se produzca una rotación hacia tasas de falsos rechazos menores.

- **Z-Norm:** es una técnica similar a T-norm, pero en este caso la distribución de *scores* de impostores se obtiene comparando el modelo utilizado para generar S_{raw} con un conjunto de locuciones de *test*. Así se consigue un alineamiento de la distribución de probabilidad de impostores dependiente del modelo.

$$S_{Znorm} = \frac{S_{raw} - \mu_{Znorm}}{\sigma_{Znorm}} \tag{3.11}$$

- **ZT-Norm:** se refiere a la combinación de una normalización Z-norm seguida de una normalización T-Norm.

- **JFA (Joint Factor Analysis):** Técnica desarrollada recientemente que modela las direcciones de máxima variabilidad intra-locutor y de canal de las características extraídas de la señal de habla. Se usa con la técnica de modelado GMM anteriormente explicada en la que se utiliza MAP para adaptar los vectores de medias del modelo UBM

mientras que los pesos y las covarianzas se mantienen entre locutores. De esta manera el modelo de locutor se representa únicamente concatenando los vectores de medias, que puede ser interpretado como un supervector. JFA considera la variabilidad de un supervector Gaussiano como la combinación lineal de la variabilidad intra-locutor y de canal. Dada una muestra de entrenamiento, el supervector de medias se puede descomponer en dos componentes estadísticamente independientes:

$$\mu_{sh} = \underbrace{\mu + Vy_s + Dz_s}_{\text{Locutor}} + \underbrace{Ux_h}_{\text{Canal}} \tag{3.12}$$

Donde Ux_h representa el supervetor que modela la variabilidad de canal y el resto de la ecuación es la componente que modela la variabilidad de locutor.

3.6. Reconocimiento de locutor en entornos forenses

La ciencia forense se define como la aplicación de prácticas científicas en la investigación de actividades criminales para demostrar la existencia de un delito y determinar la identidad de sus autores y sus *modus operandi* [24].

La evidencia forense en reconocimiento de locutor es la relación entre una grabación de voz de identidad desconocida (material recuperado) y una grabación de voz del sospechoso (material de control), involucrados en un caso.

Por lo tanto se habrá de examinar el material recuperado y el material de control para evaluar la contribución de estos hallazgos a la decisión entre dos hipótesis contrarias. Cuando las hipótesis tratan sobre si la fuente de ambas muestras analizadas es la misma, estamos frente a un problema de atribución de fuentes [25].

En la última década se ha planteado un intenso debate con el objetivo de alcanzar tanto un marco común para evaluar la evidencia y para enviar los resultados de la misma al tribunal, como un criterio unificado para evaluar el rendimiento de los sistemas forenses.

Actualmente la metodología Bayesiana de valoración de la evidencia está firmemente establecida como marco teórico aplicable a cualquier disciplina forense. De acuerdo con este enfoque Bayesiano, los sistemas forenses ofrecen sus resultados en forma de Relaciones de verosimilitud (en adelante LRs) [3]. Esta interpretación de la evidencia mediante LRs, implica que el científico forense sólo proveerá al Ttribunal este valor como la mejor manera de expresar el peso de la evidencia, que representa el grado de apoyo hacia una de las hipótesis frente a la otra, para ayudar al tribunal en sus deliberaciones y en la toma de decisión (figura 3.11).

Para aportar resultados en forma de LRs es posible adaptar cualquier sistema biométrico basado en puntuaciones, convirtiéndose por tanto en un sistema de identificación forense. La figura 3.12. muestra un esquema representativo de este procedimiento.

Como se ha expuesto anteriormente, también se ha de evaluar el rendimiento de los sistemas forenses. El análisis de la voz en condiciones en este entorno, implica una serie de

inconvenientes que obstaculizan considerablemente el rendimiento óptimo de los sistemas forenses de reconocimiento de locutor:

Las grabaciones dubitadas, provienen de interceptaciones de telefonía móvil y línea terrestre, o de micrófonos escondidos, que a su vez son registradas en diferentes equipos y soportes de grabación. También se debe considerar la limitación frecuencial en el canal telefónico, la posible codificación de canal de la voz (por ejemplo GSM) o cualquier fuente de ruido en el ambiente de la grabación.

En definitiva, nos hallamos ante una señal degradada por causas de índole cualitativo (ruido, distorsiones, bajo nivel, etc.) e insuficiencias de tipo cuantitativo (locuciones cortas). A todo esto, hay que añadir fuertes fluctuaciones de los planos expresivos entre las muestras dubitadas e indubitadas debido al carácter habitualmente no cooperativo por parte del sospechoso y la variabilidad intra-locutor [26].

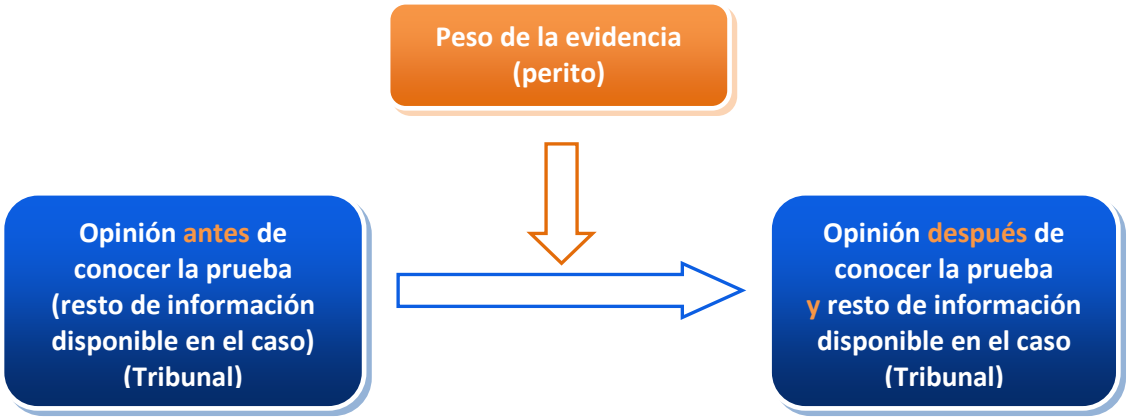


Figura 3.12: *Separación de roles*

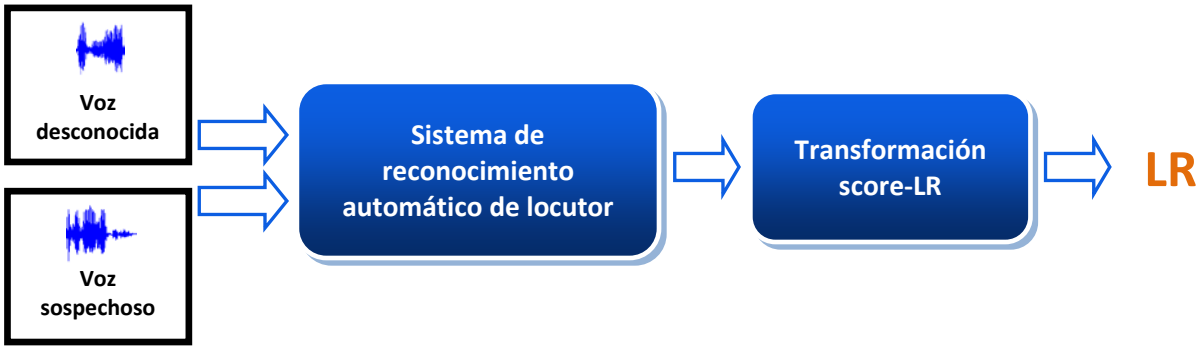


Figura 3.13: *Computación de LRs a partir de scores.*

4

Evaluación de evidencias forenses

4.1. Paradigma de identificación

Las formas clásicas de evaluación de la evidencia forense, que han sido fuertemente criticadas en las últimas décadas [27] [28], podrían resumirse en dos:

- Decisiones binarias: consiste en tomar una decisión de "Identificación" o "exclusión".
- Decisiones basadas en escalas verbales ligadas a valoraciones de probabilidad de identificación: "identificación", "muy probable", "probable", "no concluyente", "exclusión".

Ambas formas tienen dos inconvenientes importantes: la utilización de umbrales subjetivos ignorando las probabilidades a priori relacionadas con las circunstancias del caso, usurpando así el papel correspondiente al Tribunal en la toma de decisiones; en segundo lugar producen un elevado número de casos en el que el perito o laboratorio forense no se pronuncian [29].

Además, la identificación está basada o fuertemente influida por la experiencia del perito, lo que da lugar a una falta de reproducibilidad, un rendimiento difícilmente comprobable de forma repetible o incluso a una influencia de contexto los expertos [30].

4.2. Necesidad de un cambio de paradigma:

Según Saks y Koehler [31], la identificación forense debe sufrir lo que ellos llaman un "cambio de paradigma" (Paradigm Shift), es decir, una transición de los procedimientos clásicos utilizados en identificación para adecuarlos a un método con una sólida base científica y protocolos de trabajo que superen las afirmaciones no comprobables. Sustituir el concepto tradicional de unicidad y perfección por otro basado en pruebas empíricas y en probabilidades.

Algunos de los motivos de este cambio son [32]:

- **Condenas erróneas:**

  Uno de los casos más conocidos es el caso Mayfield, que fue erróneamente implicado en los atentados del 11 de Marzo en Madrid por los expertos en huellas digitales del FBI. Cuestionando así el análisis de huella dactilar que se ha asumido durante décadas como "libre de error".

- **Análisis de ADN como modelo científico de disciplina forense:**

  Desde sus orígenes en la década de 1980, la identificación genética es un método científico que evita en sus conclusiones la existencia de opiniones de expertos basadas en la experiencia y no en objetividad. Se trata de una metodología basada en procedimientos claros y transparentes que elimina los métodos no científicos. Además, los informes de identificación de ADN no son deterministas, es decir, eliminan las afirmaciones rotundas de identificación o exclusión que existen en otras disciplinas. En su lugar, se presentan informes probabilísticos objetivos basados en la información disponible y apoyados en un marco experimental y repetible. Las opiniones probabilísticas de ADN se expresan en relaciones de verosimilitud dentro de un contexto Bayesiano, tal y como muchos expertos recomiendan [32] [33] [34].

- **Cambios en las leyes:**

  Las reglas Daubert [35] han establecido el primer paso para este cambio en EEUU. Estas reglas establecen que para que la evidencia sea admitida en un juicio, las técnicas utilizadas tienen que cumplir los siguientes requisitos:

  1. Prueba empírica en condiciones reales: refutabilidad, repetibilidad.
  2. Rendimiento conocido o potencial (ej.: tasas de error).
  3. Técnica revisada y publicada en foros científicos.
  4. Existencia de estándares que definen el uso de la técnica.
  5. Aceptación general por parte de la comunidad científica.

Además, se derivan otras necesidades para que la identificación forense sea una técnica plenamente basada en métodos científicos:

- **Transparencia de los procedimientos:** La trasparencia es esencial para que en los juicios se puedan evaluar los métodos e identificar y eliminar las posibles prácticas no científicas. La claridad en la presentación de los resultados forenses es esencial a la hora de evaluar el peso de la evidencia forense y la precisión de la disciplina forense en cada caso.
- **Testabilidad de las técnicas utilizadas:** La medida de la precisión de una disciplina científica forense debería estar basada en resultados experimentales representando condiciones reales en la medida de lo posible. La existencia y disponibilidad de los datos es fundamental a la hora de poder realizar experimentos repetibles. Son necesarias técnicas de evaluación comunes y compartición de recursos para llegar a estándares que faciliten la comparación y la mejora del rendimiento de las diferentes técnicas de identificación.
- **Precisión:** La precisión es el grado de conformidad de una cantidad medida o calculada con respecto a su valor verdadero. Es importante la selección de técnicas comunes de medida de precisión para evitar confusiones y malentendidos a la hora de presentar a un tribunal los resultados sobre la precisión de las distintas técnicas forenses.
- **Procedimientos comunes:** Es importante que los forenses adopten metodologías comunes a la hora de presentar los resultados a un tribunal. Este requerimiento es necesario para evitar confusiones debidas a la incongruencia de los resultados entre disciplinas o incluso dentro de una misma disciplina. Esta convergencia debe motivarse en todos los pasos del proceso de identificación forense.

Para la satisfacción de estas necesidades, pueden adaptarse los procedimientos de identificación forense tomando como modelo la metodología seguida en identificación por

ADN. En esta disciplina, se sigue una corriente probabilística: el cálculo de LR. La metodología LR cumple los requisitos descritos en las reglas Daubert para la admisibilidad de la evidencia forense, aportando un apoyo probabilístico acerca del peso de la evidencia forense y evitando así las opiniones deterministas basadas en la experiencia [36].

4.3. Metodología de Relaciones de Verosimilitud en reconocimiento automático de locutor.

En este apartado se introduce la teoría bayesiana de relaciones de verosimilitud o LR. En concreto, se computarán LR a partir de los *scores* de salida de un sistema biométrico de identificación. El objetivo del cálculo de LR es apoyar estadísticamente una de las hipótesis del problema de atribución de fuentes, y hacerlo de manera transparente y a partir de toda la información disponible.

En un caso de atribución de fuentes normalmente se consideran dos hipótesis:

- $\boldsymbol{H_p}$**:** Hipótesis de la fiscalía. Implica que el sospechoso es el autor de las muestras incriminatorias. Según esta hipótesis, ambas muestras comparadas pertenecen a la misma fuente.
- $\boldsymbol{H_d}$**:** Hipótesis de la defensa. Conlleva que el autor de las muestras incriminatorias es otro individuo distinto al sospechoso, es decir, que las dos muestras comparadas no pertenecen a la misma fuente.

$\boldsymbol{E}$**:** Resultado de la comparación entre ambas muestras conocido como Evidencia. En el caso de este proyecto, será la salida del reconocedor automático de locutor, es decir, un *score*.

$\boldsymbol{I}$**:** Información relacionada con el caso que no se incluye en la evidencia, por ejemplo información sobre la investigación policial, información procedente de la declaración de testigos (como el sexo, la raza, la profesión, etc.) o también información sobre otro tipo de evidencia forense. En general, *I* se utiliza para definir un conjunto de posibles fuentes de la marca, conocido como población potencial [37].

A partir de esta información I, que en general no es suministrada al experto forense, se obtienen las probabilidades *a priori* de cada una de las hipótesis:

$$P(H_p|E,I) = 1 - P(H_d|E,I) \tag{4.1}$$

La decisión del tribunal estará basada en las probabilidades de las dos hipótesis teniendo en cuenta la evidencia (E) y la información sobre el caso (I). Utilizando el teorema de Bayes que relaciona las probabilidades de antes y después del análisis de la evidencia:

$$P(H_p|E,I) = \frac{p(E|H_p,I)\cdot P(H_p|I)}{p(E|I)} \tag{4.2}$$

A partir de las ecuaciones anteriores se llegaría la expresión que relaciona ambas hipótesis antes y después de la evaluación de la evidencia mediante el LR:

$$\frac{P(H_p|E,I)}{P(H_d|E,I)} = LR \cdot \frac{P(H_p|I)}{P(H_d|I)} \tag{4.3}$$

Opinión *a posteriori* (Juez) | Peso de la evidencia (científico forense) | Opinión *a priori* (Juez)

De esta manera, las probabilidades a posteriori requeridas por el Tribunal, se pueden separar en 2 valores: las probabilidades a priori, que como ya se ha mencionado están basadas en la información relacionada con el caso y es competencia exclusiva del Tribunal o Jurado y el LR que representa el peso de la evidencia y es calculado por el científico forense.

$$LR = \left.\frac{p(e|H_p,I)}{p(e|H_d,I)}\right|_{e=E} \tag{4.4}$$

El valor de LR es el cociente entre dos probabilidades, y se puede demostrar que es igual al cociente de dos densidades de probabilidad en el caso de variables continuas. En el numerador del LR se tiene una medida de similitud, que muestra la variabilidad de evidencias correspondientes a muestras pertenecientes a la misma fuente. Y el denominador es una medida de tipicidad o rareza de la muestra incriminatoria con respecto a una población relevante.

Antes de la evaluación de la evidencia, la información disponible es la que proporciona I. Después de la evaluación de la evidencia, el valor de ésta, E, se incluye a la información conocida, por lo que se modifica el valor de la apuesta a priori. Lo que sucede es un cambio de opinión como respuesta a disponer de nueva información [4]. El experto forense no usurpa así el papel que le corresponde al Tribunal, sino que proporciona una valoración, el LR, que modifica la apuesta a priori. Por tanto, el LR se puede interpretar como un grado de apoyo en favor a una de las dos hipótesis:

- Si LR > 1 La evidencia apoyará la hipótesis del Fiscal.
- Si LR < 1 La evidencia apoyará la hipótesis de la defensa.
- Si LR = 1 Sin apoyo, prueba sin valor.

Cuanto mayor sea el valor del LR, más apoyo a la hipótesis del fiscal, y cuanto menor, más apoyo a la hipótesis de la defensa.

Este marco de cálculo de LR presenta numerosas ventajas en el campo de las ciencias forenses:

- Permite a los expertos forenses evaluar y determinar el LR, un valor lleno de significado, ya que por sí mismo aporta el peso de la evidencia forense al caso.

- Define claramente el papel del científico forense, siendo éste el de evaluar cuál es el peso de la evidencia forense y dejando el papel de decidir sobre la culpabilidad o inocencia del acusado al tribunal encargado del caso.

- Las probabilidades pueden ser interpretadas como grados de creencia acerca del problema de atribución de fuentes, dejando a un lado las decisiones categóricas y subjetivas.

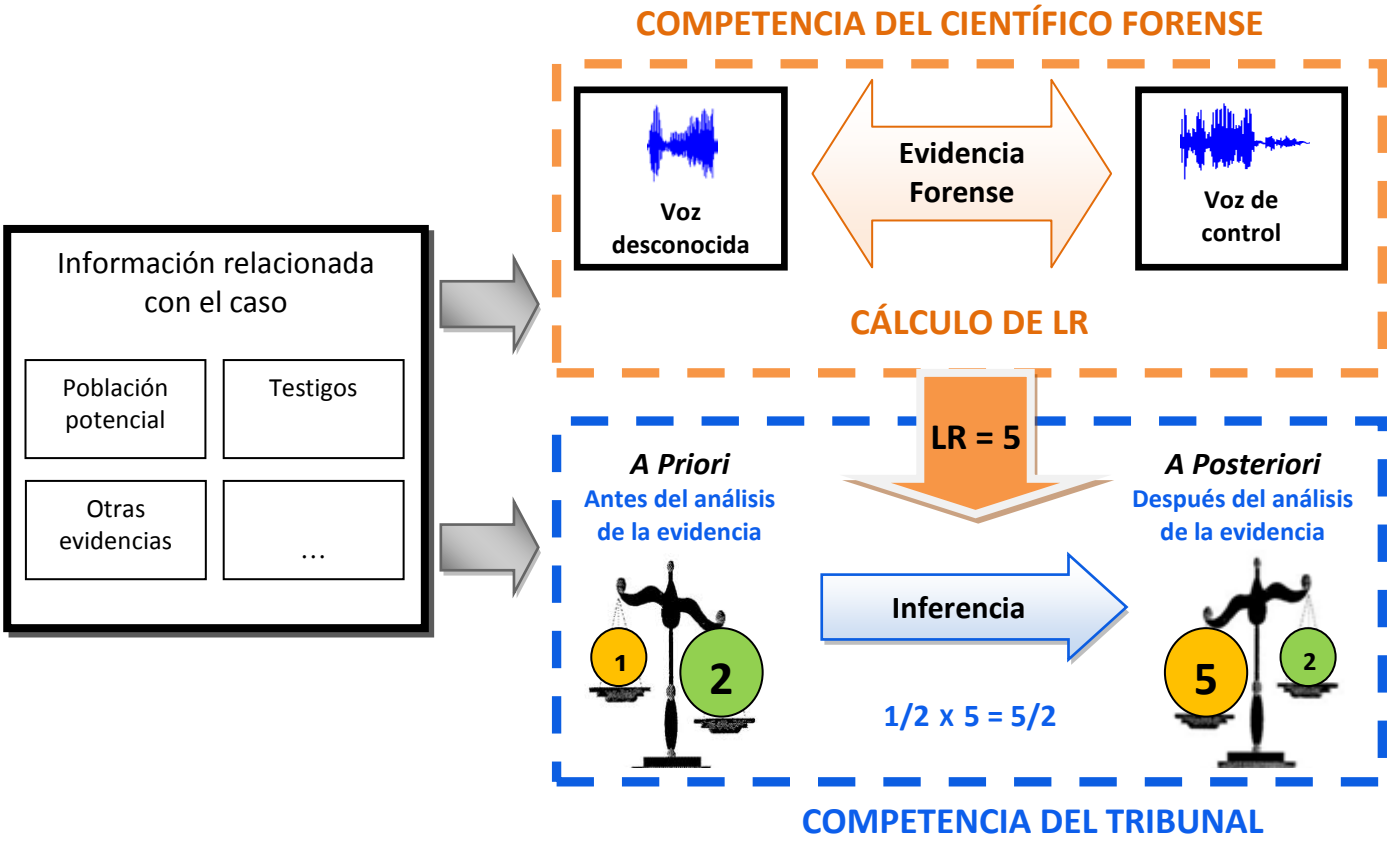


Figura 4.1: *Inferencia Bayesiana en el análisis de la evidencia forense mediante LRs [32].*

4.4. Métodos de cálculo de LR

Como se comentó en la sección 3.7, cualquier sistema biométrico puede ser adaptado para presentar los resultados en forma de LRs de acuerdo con el enfoque Bayesiano de análisis forense.

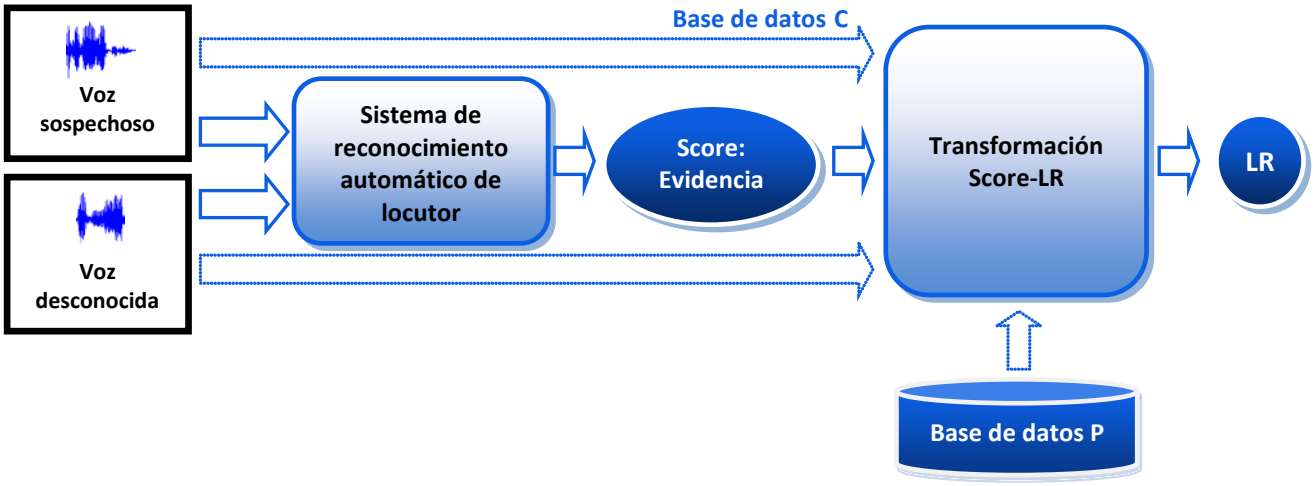


Figura 4.2: *Cálculo de LRs a partir de scores. Válido para técnicas generativas y discriminativas[32].*

El numerador del LR, representa una estimación de la intra-variabilidad (*within-source*) del sistema, mientras que el denominador representa la tipicidad o rareza de la muestra incriminatoria con respecto a una población relevante (*between-source).*

Para modelar la intra-variabilidad se utilizan *scores target* resultado de la comparación de diferentes locuciones del material de control contenidas en la base de datos (C). Por otro lado, la variabilidad entre fuentes se estima mediante los *scores non-target* resultantes de comparaciones entre las muestras recuperadas de identidad desconocida y un conjunto de modelos de la población relevante contenidos en la base de datos (P). Para construir esta base de datos se tiene en cuenta la información proporcionada por las circunstancias del caso.

En esta sección, se presentarán algunas técnicas tanto generativas (Modelado Gaussiano, GMM) como discriminativas (Regresión Logística) de cálculo de LRs en reconocimiento de locutor.

4.4.1. Modelado Gaussiano

La transformación en LR se realiza mediante una distribución gaussiana del conjunto de *scores*. Las distribuciones *target* (H_p cierta) y *non-target* (H_d cierta) quedarán definidas por la media y la desviación típica de sus respectivos conjuntos de *scores*. Debido a que la normalización T-norm "gaussianiza" las distribuciones [23], es previsible que este tipo de modelado genere mejores resultados para el sistema normalizado con esta técnica.

$$p(E|\mu_{Target}, \sigma_{Target}) \equiv \frac{1}{2\pi^{1/2}\sigma_{target}} e^{-\frac{1}{2}\left(\frac{x-\mu_{target}}{\sigma_{target}}\right)} \tag{4.5}$$

$$\boldsymbol{p(E|\mu_{Non-Target}, \sigma_{Non-Target}) \equiv \frac{1}{2\pi^{1/2}\sigma_{non-target}} e^{-\frac{1}{2}\left(\frac{x-\mu_{non-target}}{\sigma_{non-target}}\right)}} \tag{4.6}$$

Por tanto el LR queda como:

$$LR = \frac{p(E|\mu_{Target}, \sigma_{Target})}{p(E|\mu_{Non-Target}, \sigma_{Non-Target})} \tag{4.7}$$

4.4.2. GMM (*Gaussian Mixture Models*)

En este caso el LR se calcula como la división entre un GMM correspondiente al conjunto de *scores target* y un GMM correspondiente al conjunto de *scores non-target*. Como ya se vio en la sección 3.4.1, la función densidad de probabilidad puede expresarse mediante la ecuación 3.1, que adaptado al problema de atribución de fuentes puede expresarse como:

$$p(E|H_p) = \sum_{i=1}^{M} w_{target\,i} \cdot p(E|\mu_{target\,i}, \sigma_{target\,i}) \tag{4.8}$$

$$p(E|H_d) = \sum_{i=1}^{M} w_{non-target\,i} \cdot p(E|\mu_{non-Target\,i}, \sigma_{non-Target\,i}) \tag{4.9}$$

donde M es el número de gaussianas componentes, x es el conjunto de *scores* a modelar, w_i son los pesos de cada una de las mezclas donde debe cumplirse $\sum_{i=1}^{M} w_i = 1$ y $p\left(E|\mu_{target\,i}, \sigma_{target\,i}\right)$ y $p\left(E|\mu_{non-Target\,i}, \sigma_{non-Target\,i}\right)$ son las M densidades componentes para cada hipótesis H, definidas en las ecuaciones 4.5 y 4.6.

Por lo tanto, es posible definir la expresión del LR como:

$$LR = \frac{\sum_{i=1}^{M} w_i\, p(E|\mu_{target\,i}, \sigma_{target\,i})}{\sum_{j=1}^{M} w_j\, p(E|\mu_{non-Target\,i}, \sigma_{non-Target\,i})} \tag{4.10}$$

4.4.3. Regresión logística

La regresión logística es un método muy utilizado en calibración y otros ámbitos como fusión de sistemas biométricos [38]. El objetivo de la regresión logística es el de obtener una transformación lineal (desplazamiento y escalado) de un conjunto de puntuaciones de entrada para optimizar una función objetivo [32].

La transformación realizada por el modelo de regresión logística puede definirse como:

$$f_{lr} = \log\left(O\left(H_p|E\right)\right) = a_0 + a_1 \cdot E_1 + a_{02} \cdot E_{21} + \cdots + a_K \cdot E_K \tag{4.11}$$

despejando de la ecuación del teorema de Bayes (3.9) se obtiene:

$$\begin{aligned} \log(LR) &= a_0 + a_1 \cdot E_1 + a_2 \cdot E_2 + \cdots + a_K \cdot E_K - \log\left(O\left(H_p\right)\right) \\ &= a_0' + a_1 \cdot E_1 + a_2 \cdot E_2 + \cdots + a_K \cdot E_K \end{aligned} \tag{4.12}$$

Si se deshace la transformación logarítmica, se obtiene el modelo de regresión logística denotado por la siguiente expresión:

$$P\left(H_p|E\right) = \frac{1}{1+e^{-f_{lr}}} = \frac{1}{1+e^{-\log(LR)-\log\left(O(H_p)\right)}} \tag{4.13}$$

Los valores de los pesos $\{a_0,\ a_1,\ a_2, \cdots, a_K\}$ pueden obtenerse a partir de un conjunto de *scores* de entrenamiento, haciendo que sea los más cercano a 1 para los *scores target* y lo más cercano posible a 0 para *scores non-target* [32].

4.4.4. PAV (*Pool Adjacent Violators*)

Propuesta en [39], el algoritmo transforma un conjunto de *scores* en un conjunto de LR calibrados. Con este algoritmo sólo es posible calcular una transformación óptima cuando la verdadera hipótesis para cada *score* es conocida. Sin embargo, puede aplicarse una transformación óptima si se entrena con un conjunto de *scores* cuya hipótesis verdadera es conocida y después se aplica la transformación entrenada a conjunto de *test* de hipótesis verdadera desconocida [32]. El algoritmo se puede resumir en:

- Se ordenan los *scores* de manera ascendente
- Se asigna una probabilidad de 1 a cada *score* de H_p y 0 a cada *score* de H_d. El algoritmo sólo usa esta secuencia de 0 y 1 como entrada, no los *scores* originales. Este método proporciona un coste 0 pero hace que la transformación no sea monótona creciente, condición necesaria para conservar el poder de discriminación de los *scores* [32].
- El algoritmo toma iterativamente los *scores* adyacentes que violan la monotocidad y sustituye el valor de su probabilidad por la media sobre esa región.
- Se consigue por tanto una calibración óptima, conservando el poder de discriminación del conjunto de *scores*.

4.5 Evaluación del rendimiento de métodos de reconocimiento forense de locutor

Como ya se introdujo en la sección 3.7, hay varios factores que influyen en el rendimiento final del sistema de reconocimiento forense de locutor. Algunos de esos factores son: la variabilidad de sesión que provoca desajustes entre las condiciones del habla dubitada e indubitada, la falta de material de sospechoso y el desajuste en bases de datos debido a que el sistema se entrena con datos en condiciones muy diferentes a las del funcionamiento real.

Para evaluar el rendimiento de un sistema de reconocimiento de manera empírica se hacen con él una serie de comparaciones sobre muestras o grupos de muestras de origen conocido. Estas comparaciones son de dos tipos:

- **Comparaciones *target*:** donde la muestra indubitada y la muestra dubitada son producidas por la misma persona. Son casos en que la hipótesis del fiscal H_p es cierta y la comparación debería generar un LR > 1.
- **Comparaciones *non-target*:** donde la muestra indubitada y la muestra dubitada no son producidas por la misma persona. Son casos en que la hipótesis de la defensa H_d es cierta y la comparación debería generar un LR < 1.

Si se usa el LR para tomar decisiones, se pueden producir dos tipos de errores:

- **Falsas aceptaciones (FA):** donde se da como *target* una comparación *non-target*
- **Falsos rechazos (FR):** donde se da como non-*target* una comparación *target*

Dichos valores dependerán no sólo del sistema, sino del umbral establecido que separa resultados que se clasificarán como *target* de los que se clasificarán como *non-target*, por lo que una pareja de FA y FR sólo determinarán el funcionamiento del sistema en un punto particular, determinado por el umbral escogido.

El punto en el que ambos errores toman el mismo valor se denomina EER (*equal error rate*) y caracteriza el funcionamiento del sistema de forma resumida en un único valor, aunque sólo para un punto de funcionamiento. Sin embargo, puede que los objetivos del sistema (como por ejemplo reducir de forma específica la probabilidad de FA o FR) o un funcionamiento mejor del mismo en alguna región específica, lleven a trabajar con un valor diferente del umbral.

Según [32], la evaluación del rendimiento se puede dividir en dos componentes, llamadas pérdidas de discriminación y pérdidas de calibración cuya interpretación es la siguiente:

- **Pérdidas de discriminación**: es una medida de la separación, en términos de solapamiento, entre la distribución generada a partir de comparaciones *target* y la distribución a partir de las comparaciones *non-target*. Un sistema con alto poder de discriminación dará lugar a valores de LR más altos para comparaciones *target* que para comparaciones *non-target*.

- **Pérdidas de calibración**: se refiere a la interpretación estadística de un conjunto de LR. Mide la similitud entre las probabilidades a posteriori de H_p (predicciones) y la frecuencia de ocurrencia real de la hipótesis H_p. Está relacionada con la manera en que

la información se presenta ante el Tribunal. Así, el objetivo de una buena calibración, es presentar la información contenida en los LR para permitir que el Tribunal tome buenas decisiones en el contexto de la teoría de decisión bayesiana vista anteriormente. Si la calibración mejora, (las pérdidas de calibración disminuyen) la información contenida en los LRs mejorará y consecuentemente la decisión que tome el Tribunal también será mejor en promedio [32].

A continuación se presentan algunos métodos utilizados para medir el rendimiento del sistema.

4.5.1. DET (Detection Error Trade-off)

La curva DET resume la discriminación del conjunto experimental de valores de LR en una única curva [40]. En ella, se representa la probabilidad de FA frente a la probabilidad de FR para todos los puntos de funcionamiento del sistema. Se puede obtener el valor de EER, que se corresponde con el la intersección de la curva DET y la recta P_{fa}= P_{fr}. Además, permite una comparación entre sistemas clara y fácil de realizar ya que cuanto más cercana está la curva al origen de coordenadas, mejor es el poder de discriminación [10]. En la figura 4.3. se puede deducir, por tanto, que el sistema 1 tiene mayor poder de discriminación al estar su curva DET más próxima al origen de coordenadas (ser el valor del EER menor al del sistema 2).

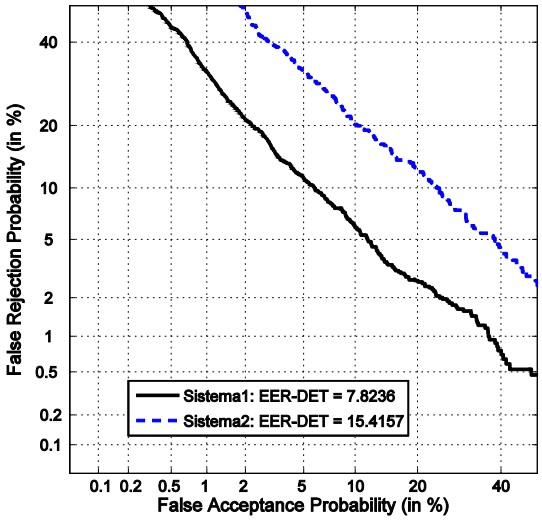


Figura 4.3: *Ejemplo de curva DET.*

Aunque la curva DET es buena para medir el poder de discriminación del sistema, no es útil para medir la calidad de la calibración, por lo tanto no sería capaz de distinguir entre el rendimiento general de sistemas con el mismo poder discriminativo. Además, no pueden utilizarse para aportar ninguna conclusión a un tribunal que implique una aceptación o rechazo de la hipótesis de partida, al no ser éste el papel del análisis forense como se comentó anteriormente. Por ello, se hace necesario el uso de otros métodos a través de los cuales se pueda evaluar el rendimiento de un sistema forense de reconocimiento de locutor. En los siguientes apartados se presentarán algunos de ellos como C_{llr}, y curvas Tippett, APE y ECE.

4.5.2. Tippett

La curva Tippett representa la proporción de casos en los que se cumple que "LR es mayor que...". Por lo tanto, se dibujan dos curvas en un mismo diagrama Tippett simultáneamente, una de ellas para la hipótesis H_p, definida en la sección 4.3, para la cual el sistema debe proporcionar valores de LR altos (LR>>1), y otra para la hipótesis contraria, para la cual el sistema debe proporcionar valores de LR bajos (LR<<1). De este modo, para cualquier valor X del eje de abscisas cada curva muestra la proporción de casos con LR mayor que X. Por tanto, cuanto más grande sea la separación de las curvas mayor será la capacidad de discriminación y mejor será el sistema (en un sistema ideal las curvas deberían ajustarse respectivamente a los márgenes superior derecho e inferior izquierdo del diagrama). Además, es muy deseable un buen rendimiento en los valores de LR cercanos a uno, esto es, LRs de la curva *target* mayores que uno y LR de la curva *non-target* menores que uno [41].
La figura 4.4 muestra una curva Tippett, el área sombreada representa la proporción de "evidencia errónea", es decir, la proporción de LRs apoyando a la hipótesis equivocada.

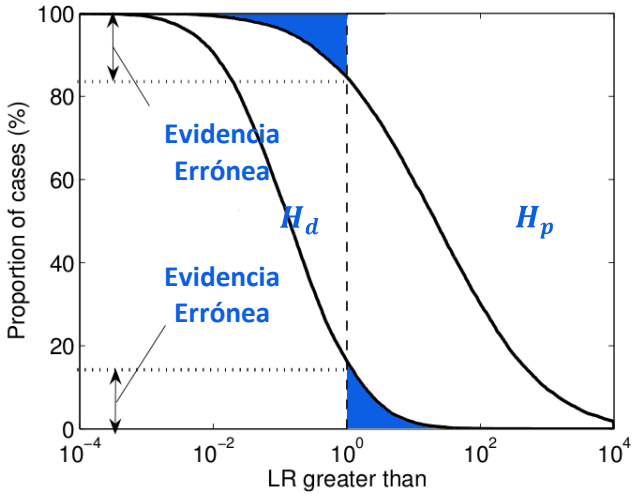


Figura 4.4: *Ejemplo de curva Tippett. Adaptada de [32].*

4.5.3. C_{llr} y C_{llr}^{min}

Para definir el rendimiento de los LR finales generados por el sistema se pueden utilizar valores C_{llr} propuestos en [39]:

$$C_{llr} = \frac{1}{2 \cdot N_p} \sum_{i_p} \log_2 \left(1 + \frac{1}{LR_i}\right) + \frac{1}{2 \cdot N_d} \sum_{j_d} \log_2 \left(1 + LR_j\right) \tag{4.14}$$

siendo N_p y N_d el número de comparaciones (valores LR) para las que H_p y H_d son ciertas respectivamente en el conjunto experimental, y que representan por tanto comparaciones target (entre muestras del mismo individuo) y non-target (entre muestras correspondientes a individuos diferentes).
Se puede observar en la ecuación 4.11 que el C_{llr} es una medida de rendimiento sobre un conjunto de LRs. Cuanto mayor sea C_{llr}, peor es el rendimiento del sistema que generará

dichos valores. Mediante el algoritmo PAV (sección 4.4.4), se realiza una transformación en la que se consigue el valor óptimo del C_{llr} conservando el poder discriminativo del conjunto de *scores* original. La pérdida general de rendimiento reflejada por C_{llr} se puede descomponer en:

1. **Pérdida de discriminación:** este valor se denomina C_{llr}^{min} y es capaz de resumir una curva DET en un único valor, cuanto menor sea éste, mayor poder discriminativo del conjunto experimental.

2. **Pérdida de calibración** del sistema bajo evaluación comparado con el sistema óptimo calculado con PAV. La pérdida por calibración puede calcularse según la relación $C_{llr}^{cal} = C_{llr} - C_{llr}^{min}$.

4.5.4. APE (*Applied Probability of Error*)

Las curvas APE propuestas en [39], representan gráficamente el rendimiento de un sistema. Son una manera de representar la probabilidad de error total si se escogen umbrales óptimos (sección 4.4.1). La probabilidad de error total se define como

$$P_e = P(H_p)P_{fr}(-\lambda) + P(H_d)P_{fa}(-\lambda) \tag{4.15}$$

siendo P_{fr} y P_{fa} las probabilidades de FA y FR en el valor negativo del Umbral de Bayes λ expresado como:

$$\lambda = log\frac{P(H_d)}{P(H_p)} \tag{4.16}$$

En una gráfica APE (figura 4.5) se observan claramente los siguientes elementos:

- **La curva roja**: es la tasa de error real del sistema tal y como está calibrado. Esta es la medida de rendimiento válida. El área de esta curva a lo largo de todo su dominio se corresponde con el valor de C_{llr}, y está representado en la altura total de la barra en el diagrama de barras bajo las APEs.

- **La curva punteada azul:** representa la probabilidad de error optimizada con el algoritmo PAV, que representa el sistema de referencia, supuesta una calibración perfecta. El área de esta curva a lo largo de todo su dominio se corresponde con el valor de C_{llr}^{min} , que está representado en la altura de la sección azul de la barra en el diagrama de barras bajo las APEs. Cuanto más cercanas estén las curvas roja y azul, mejor calibrado estará el sistema y cuanto más baja se encuentre la curva azul, mayor poder de discriminación. Por tanto el sistema 2 de la figura 4.5. presenta un mejor rendimiento que el sistema 1.

- **La curva punteada negra:** representa la tasa de error un sistema que no procesa la entrada sino que asigna a la muestra el valor más probable de entrada, y que por lo tanto tiene una probabilidad de error $P_e = \min(P(H_d), P(H_p))$. El área de esta curva es la unidad.

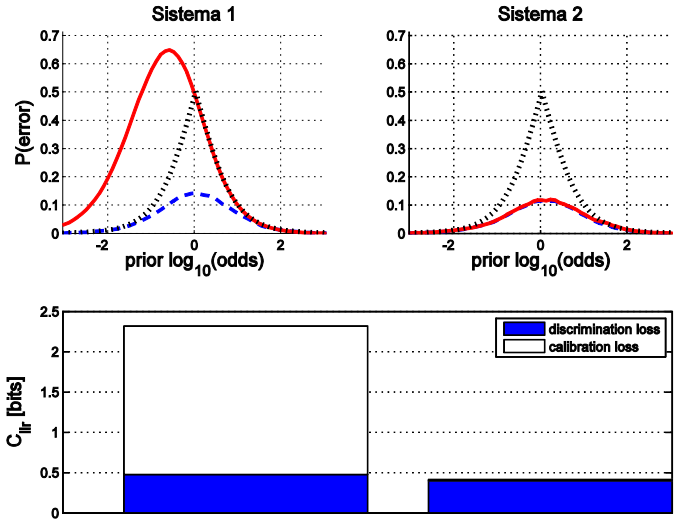


Figura 4.5: *Ejemplo de curva APE.*

4.5.5. ECE

La entropía es un concepto creado en teoría de la información, y representa el grado de incertidumbre acerca de una variable determinada con respecto a la información conocida [42]. En el marco forense, se utiliza la entropía para representar la incertidumbre que existe en cada caso acerca del valor verdadero de las hipótesis [32].

La incertidumbre sobre una variable desconocida se cuantifica mediante la entropía. El conocimiento adicional sobre las variables bajo estudio contribuyen a una reducción de la entropía y por tanto, la información sobre la variable desconocida aumentará.

- **La curva sólida** es la entropía cruzada, es decir, la pérdida media de información de los valores de LR calculados. Cuanto más alta es esta curva, más información se necesita para saber cuál de las dos hipótesis enfrentadas es la verdadera, y por lo tanto, peor es el poder de discriminación del sistema. El valor de la curva para una probabilidad prior de 0.5, representa el valor C_{llr}

- **La curva punteada azul** al igual que en la curva APE, representa el sistema de referencia, que optimiza la ECE conservando la discriminación y es obtenida por el algoritmo PAV (sección 4.4.4.). Cuanto más cercanas estén la curva azul y la roja, mejor calibrado estará el sistema. Además, la curva azul es una medida del poder de discriminación del sistema, y por tanto curvas DET iguales dan lugar a curvas azules iguales. Por esta razón, y al igual que en el ejemplo anterior, el sistema 2 presenta un mejor rendimiento que el sistema 1.

- **La curva punteada negra** representa el rendimiento de un sistema con LR=1 siempre, conocido como sistema neutral. Esta curva sirve como referencia para la curva sólida, que siempre debe estar por debajo para que el sistema tenga algún poder de discriminación. Si la curva sólida estuviera por encima, el sistema estaría perdiendo más información con el cálculo de LR que si la decisión se basara solamente en la información inicial del caso.

En el caso de atribución de fuentes, el Tribunal establece la probabilidad *a priori* $P(H_p)$ antes del análisis de la evidencia. Posteriormente, a través de la curva ECE junto con la probabilidad *a priori* se obtiene una medida de la información media (sobre el caso forense) que se necesitaría para saber el verdadero valor de la hipótesis.

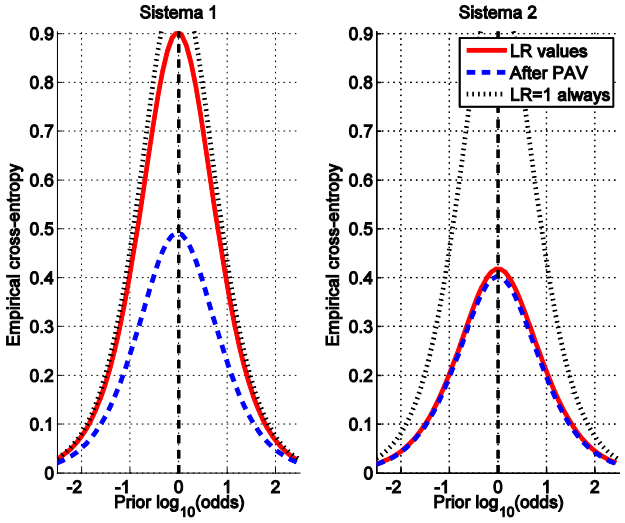


Figura 4.6: *Ejemplo de curva ECE.*

5

Redes Bayesianas

5.1. Introducción a las redes Bayesianas

Las redes bayesianas son un formalismo con multitud de aplicaciones para una representación compacta de las relaciones inciertas entre parámetros en un dominio.
Estos modelos probabilísticos gráficos combinan la teoría de la probabilidad con la de grafos. Proporcionan una herramienta natural para abordar dos de los problemas que afrontan las matemáticas y la ingeniería aplicadas [43] como son la incertidumbre y la complejidad [33].

Se componen de dos partes, una cualitativa y otra cuantitativa:

- **La parte cualitativa** es un grafo dirigido acíclico donde los nodos representan variables aleatorias del dominio $X_1, X_2, \ldots, X_n$, mientras que los arcos representan valores de dependencias entre las variables. Si hay un arco desde el nodo X hasta el nodo Y, se dice que X es padre de Y e Y es hijo de X. Las redes bayesianas asumen que un nodo depende únicamente de sus padres.
- **La parte cuantitativa** representa la incertidumbre del problema por medio de probabilidades condicionadas: posibles relaciones causa efecto entre los nodos. Cada nodo posee una tabla de probabilidades condicionales asociada, que define la probabilidad de cada uno de los estados en los que puede estar una variable, dados los posibles estados de sus padres.

La propiedad clave de una red bayesiana, es que a través del teorema de Bayes, cada distribución de probabilidad conjunta puede descomponerse como un producto de probabilidades condicionales [33] y esto facilita la investigación de relaciones entre variables en el contexto de un caso particular.

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P\big(x_i | padres(X_i)\big) \tag{5.1}$$

donde x_i representa el valor que toma la variable X y $padres(X_i)$ denota los valores que tienen el conjunto de los padres en la red Bayesiana del nodo X_i.

Otra virtud de las redes bayesianas es que proporciona un sistema de inferencia, donde una vez encontradas nuevas evidencias sobre el estado de ciertos nodos, se modifican sus tablas

de probabilidad y, a su vez, las nuevas probabilidades son propagadas al resto de nodos. Esta propagación de probabilidades se conoce como inferencia probabilística. Las evidencias pueden ser de dos tipos:

- **Evidencia dura (hard),** conocida también como instanciación, se da cuando se asigna un valor concreto a una variable, es decir, se tiene certeza del estado de dicha variable.
- **Evidencia parcial (soft),** permite analizar las probabilidades a priori de los estados que puede tomar la variable.

Existen varios algoritmos de propagación de probabilidad. Las diferencias entre ellos se basan principalmente en la precisión de los resultados y en el consumo de recursos durante el tiempo de ejecución.

Los algoritmos de propagación se dividen inicialmente en “exactos” o “aproximados”, según cómo calculen los valores de las probabilidades. Los métodos exactos calculan los valores por medio del teorema de Bayes, mientras que los métodos aproximados utilizan técnicas iterativas de muestreo, en las que los valores se aproximarán más o menos a los exactos dependiendo del punto en que se detenga el proceso.

5.2. Aplicaciones

El campo de aplicación de las redes Bayesianas es muy grande y variado. Algunas áreas de aplicación son: bioinformática [44]; psicología [45]; diagnóstico médico [46], datos georeferenciados [47][48], robótica o problemas de inferencia en ciencia forense [33][49]. Esta última aplicación es la que se utiliza en este proyecto y que será desarrollada en el siguiente punto.

5.3. Redes bayesianas en entornos forenses

Como ya se ha comentado en este capítulo, las redes bayesianas se han manifestado como una valiosa ayuda para la representación de relaciones entre características de interés en situaciones en las que existe incertidumbre. En la ciencia forense también se han propuesto las redes Bayesianas como un método formal de razonamiento que ayuda a los expertos forenses a entender las dependencias que pueden existir entre diferentes aspectos de la evidencia, y a abordar el análisis formal de una toma de decisión.

La utilización de modelos gráficos para representar asuntos jurídicos no es nueva. Los métodos gráficos de Wigmore [50] pueden considerarse como los predecesores de los modernos modelos gráficos como son las redes bayesianas. En [51],[52] y [53] se pueden encontrar ejemplos de uso de gráficos que fueron desarrollados para proporcionar soporte formal para llegar a obtener conclusiones basadas en numerosas evidencias.

El uso de redes probabilísticas ha alcanzado notoriedad con el análisis de complejos y famosos casos como el caso Collins y Sacco-vanzetti o más recientemente el juicio de O.J Simpson, todos ellos analizados mediante modelos gráficos.

Como principales **ventajas** del uso de las redes bayesianas especialmente en entornos forenses se tiene:

- La difícil tarea intelectual de organizar y combinar conjuntos complejos de evidencias para resaltar las dependencias e independencias puede realizarse de modo visual e intuitivo. Esto hace que sea más fácil de entender por parte del Tribunal y facilitará así el proceso de la toma de decisión.
- Se permite estudiar la sensibilidad de cambios ante estados de verdad de otras variables de interés.
- La combinación de nodos y flechas constituyen caminos a través de red. Por consiguiente, una red puede considerarse como una representación gráfica compacta de una evolución de todas las posibles historias relacionadas con un escenario.
- Cuando se dispone de más conocimiento, las especificaciones cualitativas y/o cuantitativas se pueden adaptar para alcanzar una nueva comprensión de las propiedades del dominio.
- La tarea de especificar las ecuaciones relevantes puede hacerse invisible al usuario, es decir, no es necesario conocer la teoría matemática subyacente. Además el cálculo aritmético puede automatizarse casi completamente.

En este proyecto, se estudiará la utilización de redes bayesianas para la combinación de evidencias forenses. Para una ampliación sobre la utilización de modelos gráficos en la valoración de la evidencia forense se pueden consultar libros como [33] o [49].

5.3.1 Combinación de evidencias mediante redes bayesianas

Históricamente, la dificultad en la combinación de pruebas ha sido abordada a través de una discusión sobre el problema llamado dificultad de conjunción: dos evidencias, cuando son combinadas, parecen producir una probabilidad menor que cuando son consideradas por separado. Este problema fue tema de debate entre Cohen [54][55] y Dawid [56]. Un resumen sobre esta cuestión y la solución propuesta por Dawid puede también encontrarse en Aitken y Taroni [33].

El siguiente ejemplo ilustra el problema de combinación de evidencias [49]:

Supóngase que E_1 y E_2 representan dos evidencias utilizas para extraer una conclusión sobre una proposición determinada, H, con dos posibles salidas: H_p, la proposición de la acusación y H_d, la proposición de la defensa. Supóngase que la probabilidad de H_p dada E_1 o dada E_2 es 0.7, es decir,

$$P(H_p|E_1) = P(H_p|E_2) = 0.7$$

La probabilidad de interés es:

$$P(H_p|E_1, E_2)$$

Si E_1 y E_2 se consideran independientes, dado H_p o H_d, su probabilidad conjunta puede escribirse como el producto de las probabilidades individuales:

$$P(E_1, E_2|H_p) = P(E_1|H_p) \cdot P(E_2|H_p)$$

Es tentador pensar que $P(H_p|E_1, E_2)$ se obtiene de manera análoga, es decir,

$$P(H_p|E_1, E_2) = P(H_p|E_1) \cdot P(H_p|E_2)$$

El aparente resultado contradictorio de este procedimiento (incorrecto) resulta $0.7 \cdot 0.7 = 0.49$, que es una probabilidad menor de H_p que considerando sólo E_1 o E_2.

Considerando el teorema de Bayes para dos evidencias E_1 y E_2 y las proposiciones H_p y H_d

$$\frac{P(H_p|E_1, E_2)}{P(H_d|E_1, E_2)} = \frac{P(E_1, E_2|H_p)}{P(E_1, E_2|H_d)} \cdot \frac{P(H_p)}{P(H_d)} \tag{5.2}$$

Asumiendo que las probabilidades priores son iguales, $P(H_p) = P(H_d)$, la probabilidad de interés, viene dada por [49]:

$$P(H_p|E_1, E_2) = \frac{LR}{1 + LR} \tag{5.3}$$

donde el LR puede calcularse como:

$$LR = \frac{P(E_1|H_p)P(H_p)}{P(E_1|H_d)P(H_d)} \cdot \frac{P(E_2|H_p)P(H_p)}{P(E_2|H_d)P(H_d)} \tag{5.4}$$

$$= \frac{P(H_p|E_1)}{P(H_d|E_1)} \cdot \frac{P(H_p|E_2)}{P(H_d|E_2)} = \frac{0.7}{0.3} \cdot \frac{0.7}{0.3} = \frac{0.49}{0.09}$$

Aplicando la ecuación (5.3) se obtiene finalmente:

$$P(H_p|E_1, E_2) = 0.84$$

Así, la combinación de evidencias proporciona una probabilidad posterior mayor para H_p que cuando son consideradas de manera individual. Por lo tanto, en casos en que se tienen dos evidencias E_1 y E_2 hay que evaluar el efecto combinado para revisar la creencia en una proposición de interés H, y además no deben tenerse en cuenta sólo las probabilidades posteriores de las respectivas proposiciones [49], como se ha propuesto por ejemplo, en campos como huellas de zapatos [57] o escritura [58] siendo medios inadecuados para la evaluación de la evidencia científica [59].

Por lo tanto, las dificultades de interpretación cuando hay que combinar dos o más evidencias, se pueden abordar utilizando métodos de evaluación basados en LRs. Usando los LRs, se pueden ir añadiendo evidencias y examinar la probabilidad posterior de la proposición de interés, H. Las probabilidades *a posteriori* cuando se considera una sola evidencia, por ejemplo E_1, se convierten en las probabilidades *a priori* para la siguiente evidencia, E_2:

$$\frac{P(H_p|E_1)}{P(H_d|E_1)} = \frac{P(E_1|H_p)}{P(E_1|H_d)} \cdot \frac{P(H_p)}{P(H_d)} \tag{5.5}$$

El término de la izquierda de (5.5) representa las probabilidades en favor de H_p dado E_1. Cuando se tiene en cuenta el segundo valor de evidencia E_2, se obtiene por la ecuación 5.1:

$$\frac{P(H_p|E_1,,E_2)}{P(H_d|E_1,E_2)} = \frac{P(E_2|H_p,E_1)}{P(E_2|H_d,E_1)} \cdot \frac{P(H_p|E_1)}{P(H_d|E_1)} = \frac{P(E_2|H_p,E_1)}{P(E_2|H_d,E_1)} \cdot \frac{P(E_1|H_p)}{P(E_1|H_d)} \cdot \frac{P(H_p)}{P(H_d)} \tag{5.6}$$

Así, las probabilidades *a posteriori* a favor de H_p incorporan la nueva información E_1 y E_2.

Uno de las partes del presente proyecto, se centrará en el estudio de la combinación de evidencias mediante el uso de redes Bayesianas. Concretamente se examinarán dos posibilidades, cuyas representaciones gráficas se pueden ver en la figura 5.

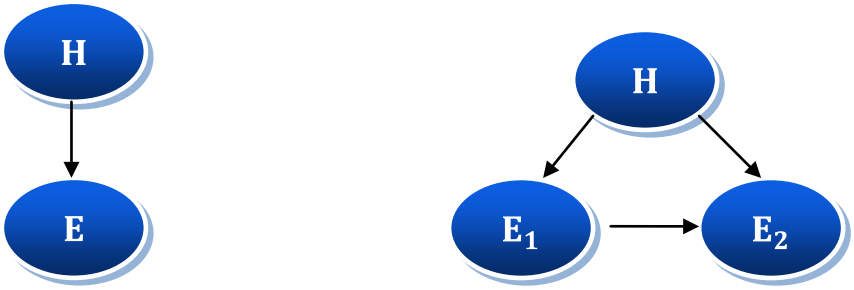


Figura 5.1: *Representación del problema de atribución de fuentes mediante redes Bayesianas.*

En ambas redes, el nodo H, representa las dos posibles hipótesis del problema de atribución de fuentes H_p y H_d; y el (los) nodo(s) E, la evidencia, que en este caso será la puntuación de salida del sistema de reconocimiento de locutor. En la primera red Bayesiana se tiene un único nodo de evidencia resultado de la combinación previa de dos evidencias individuales. Y en la segunda, se utilizarán dos nodos con evidencias sin combinar. El procedimiento tanto de construcción como de instanciación será descrito con detalle en el capítulo 7.

5.4. Hugin Expert

Para la parte de combinación de evidencias mediante redes Bayesianas se ha hecho uso de la herramienta Hugin Expert [60]. Este programa permite la construcción, aprendizaje y análisis de redes bayesianas y diagramas de influencia. Está compuesto por un motor de decisión y una interfaz gráfica, aunque también hay disponibles APIs para diferentes lenguajes de programación como C, C++ o Java.

En este proyecto se ha utilizado la API para C++ para la creación de las diferentes redes Bayesianas e instanciación de nodos de manera automática, debido a la gran cantidad de datos manejados.

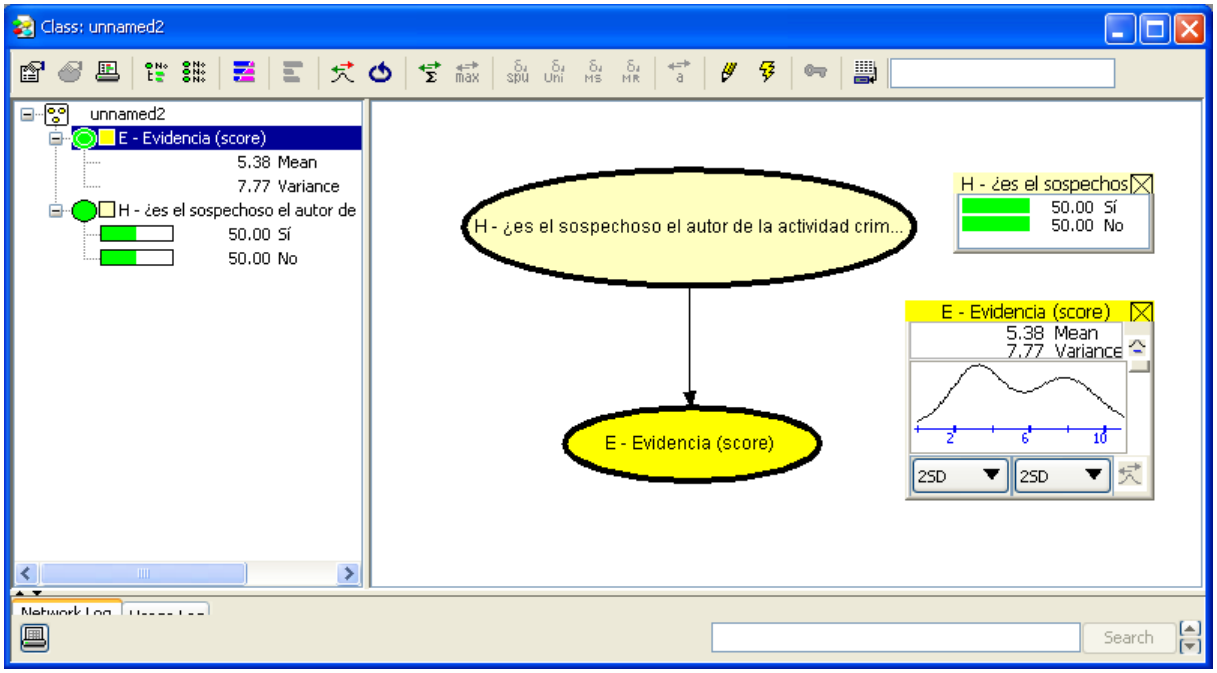


Figura 5.2: *Interfaz gráfica de Hugin Expert.*

En el anexo A se encuentra un tutorial en el que se construye e instancia una red de ejemplo utilizando tanto la interfaz gráfica como la API de Hugin para C++.

6

Marco experimental

Este capítulo introduce el marco experimental usado en este proyecto, incluyendo las bases de datos de voz y el sistema utilizado para la presentación de resultados.

6.1 Sistema Utilizado

Para la realización de los experimentos desarrollados en este proyecto, se ha utilizado un sistema biométrico de reconocimiento de locutor en el estado del arte.

El sistema de reconocimiento de locutor ha sido desarrollado por el grupo de reconocimiento biométrico ATVS-UAM. Este sistema genera un *score* por cada comparación de muestras, no genera LRs. Los LRs serán calculados, posteriormente mediante diferentes técnicas y de acuerdo a la teoría Bayesiana explicada en el capítulo 4.

El proceso de parametrización se ha realizado segmentando los archivos de audio mediante ventanas tipo Hamming de 20 ms de duración con solapamiento del 50% (tasa de 10 ms), extrayendo los vectores de características utilizando 19 coeficientes MFCC.

El sistema empleado para la construcción de modelos es el ATVS-UAM GMM que modela la distribución de probabilidad mediante Modelos de Mezclas de Gaussianas, cuyos fundamentos han sido vistos con detalle en la sección 3.4.1. El *score* de similitud viene dado por la ecuación 3.8, que mide la diferencia entre la probabilidad de que los datos de *test* provengan del modelo de locutor adaptado y la probabilidad de que provengan del UBM.

El UBM consta de 1024 gaussianas en un espacio de características de 38 dimensiones. Se realiza una iteración K-means y 5 iteraciones EM, a partir del cual se obtienen los modelos de locutor por adaptación MAP (con una sola iteración) de únicamente los vectores de medias, empleando un valor para el parámetro de relevancia r = 16. Tanto para la adaptación de los modelos de locutor como para el *scoring* se emplean únicamente las 5 gaussianas más pesadas de cada vector de entrenamiento y *test*, respectivamente.

Para compensar los efectos de la variabilidad de sesión vista en la sección 3.6 se utilizan diferentes técnicas: CMN RASTA *warping* a nivel de parámetros, JFA a nivel de modelos y T-norm, Z-norm y ZT-norm a nivel de *scores*.

6.2 Bases de Datos

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

El "*Speech Group*" del Instituto Nacional de Estándares y Tecnologías de los Estados Unidos (NIST) [61] realiza evaluaciones anuales desde el año 1996. A partir a partir del año 2006 han pasado a ser bianuales intercalándose con evaluaciones de reconocimiento de idioma. El objetivo general de las evaluaciones es impulsar el desarrollo tecnológico, medir el estado del arte y encontrar las aproximaciones algorítmicas más prometedoras. Para ello, diseña una serie de *tests* que tratan de verificar el rendimiento de dichos sistemas, tomando como punto de partida cuatro ejes de referencia: el tipo de entrenamiento, la duración de los segmentos-muestra, edad/sexo de los locutores y la influencia del "factor canal".

La mayor parte de las bases de datos de este proyecto derivan de las bases de datos utilizadas en los experimentos realizados en estas evaluaciones:

- **Switchboard 1:** contiene habla conversacional en inglés americano grabada sobre línea telefónica convencional. No recoge variabilidad dialectal, pero sí la derivada de las diversas líneas de teléfono y los distintos tipos de terminales telefónicos (de micrófono tipo electrect, de carbón, etc.). Esta base de datos fue empleada en la evaluación NIST del año 2001.

- **Switchboard 2:** contiene habla conversacional en inglés americano grabada sobre línea telefónica convencional. Como en Switchboard 1, recoge la variabilidad de diferentes tipos de líneas y terminales telefónicos pero en un mayor grado. Además recoge variabilidad dialectal grabada en diferentes fases: Fase 1 (inglés americano de la mitad Atlántica), Fase 2 (inglés americano de la mitad oeste) y Fase 3 (inglés americano del sur). La Fase 3 fue empleada en las evaluaciones NIST de los años 1996 a 1999, así como en las de 2002 y 2003 junto con la Fase 2; en la del año 2000 se empleó la base completa.

- **Switchboard 3** (o Switchboard Cellular): contiene habla conversacional en inglés americano, recogiendo distintos dialectos, y está grabada sobre redes móviles. Fue grabada en 2 fases en las que se obtienen distintos tipos de canal: Fase 1 (canal de transmisión GSM) y Fase 2 (canal de transmisión CDMA). La Fase 1 fue empleada en la evaluación NIST de 2001, mientras que la 2 fue empleada en las de 2002 y 2003.

- **MIXER** y datos adicionales **multi-lenguaje**: presenta tres diferencias fundamentales respecto las versiones de Switchboard. En primer lugar, la variabilidad de canal y terminales es significativamente mayor, incluyendo habla grabada sobre teléfonos inalámbricos de líneas telefónicas convencionales y redes móviles, con terminales tipo tele-operador, manos libres, etc. En segundo lugar, es multi-lenguaje, conteniendo habla en inglés americano, español, árabe, chino mandarín y ruso. En tercer lugar, se empleó un protocolo para aleatorizar las conversaciones entre locutores en la base. Por todo ello, MIXER contiene mucha más variabilidad que las bases de datos previas. Esta base de datos (sin la ampliación multi-lenguaje) se empleó en la evaluación NIST del año 2004. En las evaluaciones de 2005 y 2006 se añadieron nuevas grabaciones incluyendo

los idiomas anteriores y siguiendo el mismo protocolo, añadiendo variaciones de dialecto y locutores no nativos.

- Datos **multi-micrófono**: durante las evaluaciones NIST de los años 2005 y 2006 se hizo un esfuerzo considerable para recoger bases de datos multi-micrófono. En éstas se registra la conversación de un locutor de forma simultánea a través de la línea telefónica y de una variedad de micrófonos: micrófono de solapa, micrófonos a distintas distancias, micrófono de PC, etc. Esto hace que la variabilidad sea mucho mayor que para habla sólo telefónica, suponiendo un punto de referencia realista y desafiante. Es por ello que este tipo de datos han sido empleados también en la evaluación del año 2008.

- **Ahumada III** [62] es una base de datos forense que fue adquirida por el departamento de Procesamiento de Imagen y Acústica de la Guardia Civil. Incluye datos de locuciones de conversaciones reales que fueron registradas entre los años 1995 y 2004.
  Las locuciones pertenecen a 69 locutores masculinos. Por cada uno de los individuos hay un fichero de entrenamiento de 120 s de duración y entre 4 y 10 ficheros de test con una duración media de 13 s. Las grabaciones provienen de terminales GSM y presentan una gran variabilidad en cuanto a entornos, estados emocionales, ruido ambiental, etc. Esta base de datos presenta la peculiaridad de que en el formateo de la misma se suprimieron los silencios, quedando sólo las zonas de voz.

Los sistemas empleados en la parte experimental de este proyecto utilizan UBMs entrenados con datos procedentes de las bases Swichboard 1 y 2, y de las evaluaciones NIST SRE 2004 y 2005 (MIXER, datos multi-lenguaje y multi- micrófono); estos datos también se utilizaron en el entrenamiento de las matrices de *eigenvoices* y *eigenvectors* para la aplicación de Joint Factor Analysis. Así mismo, las cohortes de normalización para Z-Norm, T-Norm y ZT-norm, también se obtuvieron de la base de datos MIXER.

La base de datos Ahumada III se ha utilizado por un lado, para la realización de las pruebas y por otro lado en la parte de entrenamiento de las distribuciones de probabilidad *target* y *non-target* (H_p y H_d verdadera respectivamente) para posteriormente realizar el cálculo de los LRs (ecuación 4.4).

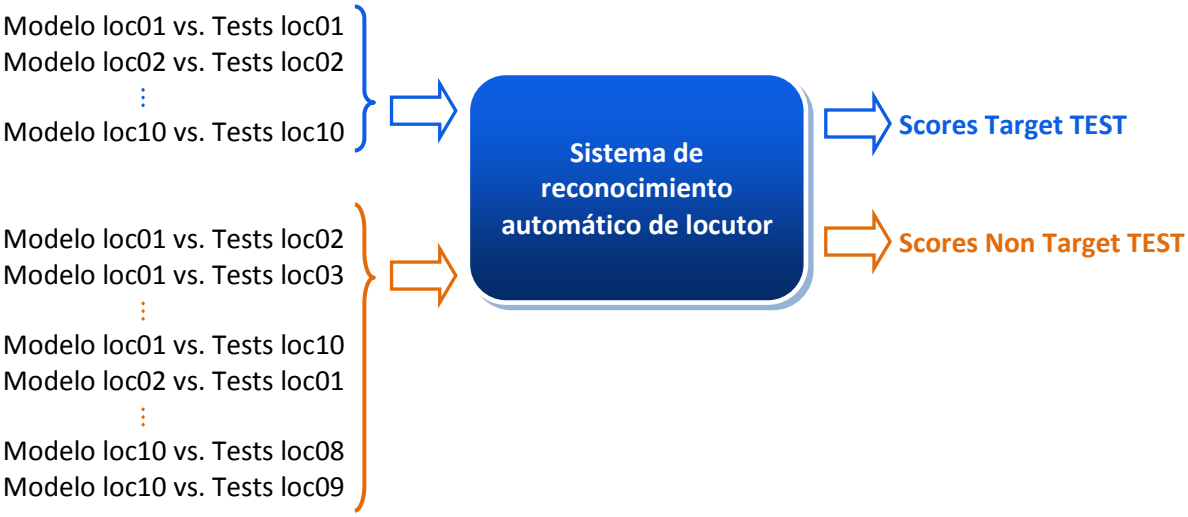
6.3. Validación cruzada

En este proyecto se hace uso de la función validación cruzada para que la evaluación del sistema sea honesta ya que se evalúa y se entrena con la base de datos Ahumada III. Para que esto ocurra, es necesario que en el conjunto de datos de *test* con los que se va a evaluar el sistema, no existan *scores* provenientes de comparaciones con modelos utilizados para el conjunto de entrenamiento.

Para ello, se divide el conjunto de los 69 locutores disponibles en 2 grupos, uno de ellos se utilizará para formar el conjunto de datos de *test* y el otro en el conjunto de datos de entrenamiento. El procedimiento sería el siguiente:

1. Se seleccionan N locutores que formarán una división de la validación cruzada. Con estos locutores se constituye el conjunto de *scores* de *test* con los que se instanciará la red Bayesiana para el posterior cálculo de los LRs. Estos *scores* se obtendrán al enfrentar los modelos con locuciones dentro de ese conjunto de datos. Por ejemplo, si se selecciona N=10, el conjunto de datos de *test* estará formado por los locutores del 1-10 para la

primera división, por lo tanto los *scores* para instanciar la red serán los que surjan de las siguientes comparaciones:



Pero no se calcularán, por ejemplo, scores que surgen de comparar Modelo loc09 vs. Tests loc11, porque loc11 no está en la división 1.

2. En el conjunto de datos de entrenamiento (*train*) correspondiente a la primera división, no podrán formar parte los locutores que constituyen el conjunto de datos de t*est*, es decir, los locutores 1-10. Por lo tanto, los *scores* para entrenar la red Bayesiana serán los que surjan de las siguientes comparaciones:

Pero no formará parte de este conjunto de entrenamiento el *score* resultante de Modelo loc11 vs. Tests loc09, porque loc09 se encuentra en el conjunto de datos de *test*.

3. Se repite el procedimiento para el resto de divisiones de N locutores tantas veces como ceil (69/N).

7

Experimentos y Resultados

Uno de los mayores problemas a la hora de evaluar el peso de la evidencia forense ocurre cuando se disponen de diversos fragmentos de voz cuya procedencia es desconocida. La aportación total de cada uno de esos fragmentos al peso total de la evidencia de voz constituye un importante tema de investigación.

Por lo tanto, se parte de un caso con una toma indubitada y múltiples tomas dubitadas. A partir de este escenario, el objetivo será evaluar el peso de dicha evidencia utilizando relaciones de verosimilitud, es decir, se generará un único LR por cada comparación de toma indubitada frente a una combinación de tomas dubitadas. El problema es determinar la mejor manera de combinar la información que aportan dichas tomas dubitadas de cara a generar el LR final. Para ello, a lo largo del capítulo se estudiarán diferentes métodos de combinación de evidencias y se realizará una posterior evaluación de los resultados.

Los resultados mostrados en las gráficas a lo largo de todo el capítulo, corresponden al rendimiento proporcionado por el sistema con normalización T-norm, compensado en locutor y canal ya que es el que mejor resultado ofrece para la totalidad de los resultados. No obstante en las gráficas resumen se podrá apreciar el rendimiento comparado con los demás sistemas.
El resto de resultados tanto de manera gráfica como de forma numérica se recogen en el anexo B, el cual no resulta necesario de cara a entender las principales conclusiones obtenidas en el proyecto.

La parte experimental se divide en dos bloques principales: El primero de ellos consiste en la comparación de diferentes métodos más inmediatos de combinación de evidencias y el segundo consiste en la utilización de redes bayesianas para calibración y combinación de *scores* de salida de un sistema de reconocimiento automático de locutor.

1. **Comparación de diferentes estrategias de combinación de evidencias y calibración de los resultados.**

   En esta primera parte se estudiarán diferentes esquemas de combinación de 2 y 3 evidencias procedentes de un sistema de reconocimiento automático de locutor. Se hace necesario un proceso de calibración, que ajuste los valores de LR y establezca el grado correcto de apoyo a una hipótesis de acuerdo con el funcionamiento del sistema, compensando el grado de desviación de los resultados generados por el mismo. La calibración se realiza apoyándose en comparaciones entre muestras de control conociendo *a priori* que H_p y H_d son ciertas, y compensa las desviaciones entre muestras que no son interpretables probabilísticamente. El resultado final perseguido es un único LR por cada

comparación de toma indubitada frente a una combinación de tomas dubitadas, que exprese el peso de la evidencia, arrojado por el sistema. Para ello, se han seguido tres estrategias de combinación:

- **Calibración y suma de Log-LR calibrados.**

El conjunto de scores de salida del sistema de reconocimiento de locutor se calibra utilizando regresión logística, de tal manera que para cada comparación en el sistema, se obtiene un Log-LR calibrado que se sumará dando lugar a un Log-LR por cada combinación realizada. En las figuras 7.1 y 7.2 se muestra un esquema del procedimiento seguido para esta estrategia de combinación.

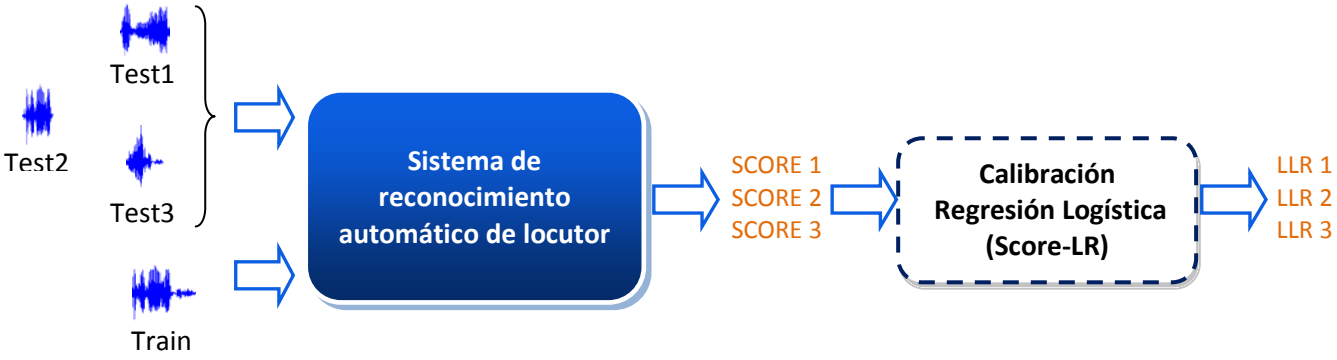


Figura 7.1: *Cálculo de los log-LR resultantes de cada comparación de muestra de test (dubitadas) con la muestra train (indubitada).*

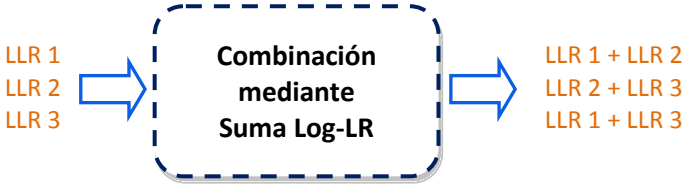


Figura 7.2: *Combinación de evidencias mediante suma de log-LRs. Para el caso de combinación de 3 evidencias, el cálculo sería similar, calculando las combinaciones de 3 de los LLRs de entrada.*

- **Suma Log-LR y calibración posterior.**

  En primer lugar, se genera un único Log-LR por cada toma dubitada y se suman de 2 en 2 y de 3 en 3 suponiendo independencia (Bayes ingenuo) obteniendo un único valor para cada combinación de evidencias. A continuación de aplica una etapa posterior de calibración, para compensar la falsa presunción de independencia.

- **Concatenación de los archivos de audio correspondientes a tomas dubitadas.**

  En primer lugar, se combinan las tomas dubitadas concatenando los archivos de audio de 2 en 2 y de 3 en 3, creando dos nuevas bases de datos con archivos combinados. Posteriormente, se utilizan los sistemas de reconocimiento de locutor, generando un *score* por cada comparación de la toma indubitada con cada combinación generada. Se calibra posteriormente mediante regresión logística. En las figuras 7.3 y 7.4 se muestra un esquema del procedimiento seguido para esta estrategia de combinación.

Figura 7.3: *Combinación de evidencias mediante concatenación de ficheros de audio. Para el caso de combinación de 3 evidencias, sería similar, calculando las combinaciones de 3 de los archivos de entrada.*

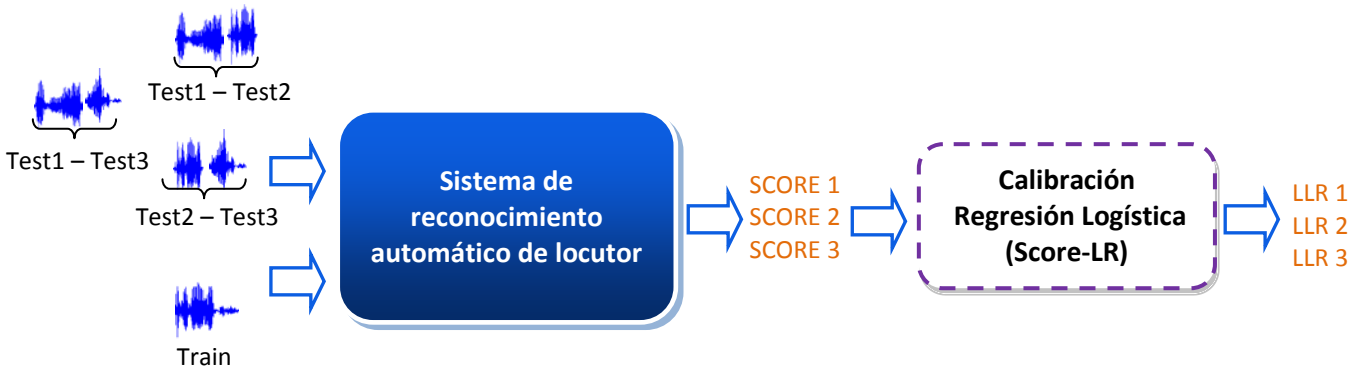


Figura 7.4: *Cálculo de los log-LR resultantes de cada comparación de muestra de test (dubitadas) ya combinada, con la muestra train (indubitada).*

2. Uso de redes Bayesianas.

Esta parte del proyecto tiene como objetivo la adaptación redes Bayesianas como posible método de calibración y combinación de evidencias. Para ello se propone la utilización de dos redes diferentes, una más sencilla, de dos nodos para implementar los métodos de calibración y otra de tres nodos para la combinación de dos evidencias. En ambas redes, el nodo H, representa la dos posibles hipótesis del problema de atribución de fuentes H_p y H_d; y el (los) nodo(s) E, la evidencia, que será la puntuación de salida del sistema de reconocimiento de locutor.

- **Contribución: Utilización de redes Bayesianas como método de calibración.**

  El procedimiento seguido para calibrar mediante redes Bayesianas está representado en la figura 7.5. Como entrada al sistema de reconocimiento de locutor se tienen los archivos combinados mediante concatenación, por ser éste el que mejor resultado ofrece. Posteriormente los *scores* resultantes son calibrados con la red Bayesiana de dos nodos entrenándola mediante las técnicas Gauss, GMM y PAV explicadas en la sección 4.4. En la sección 7.2.1 se describirá en detalle esta metodología.

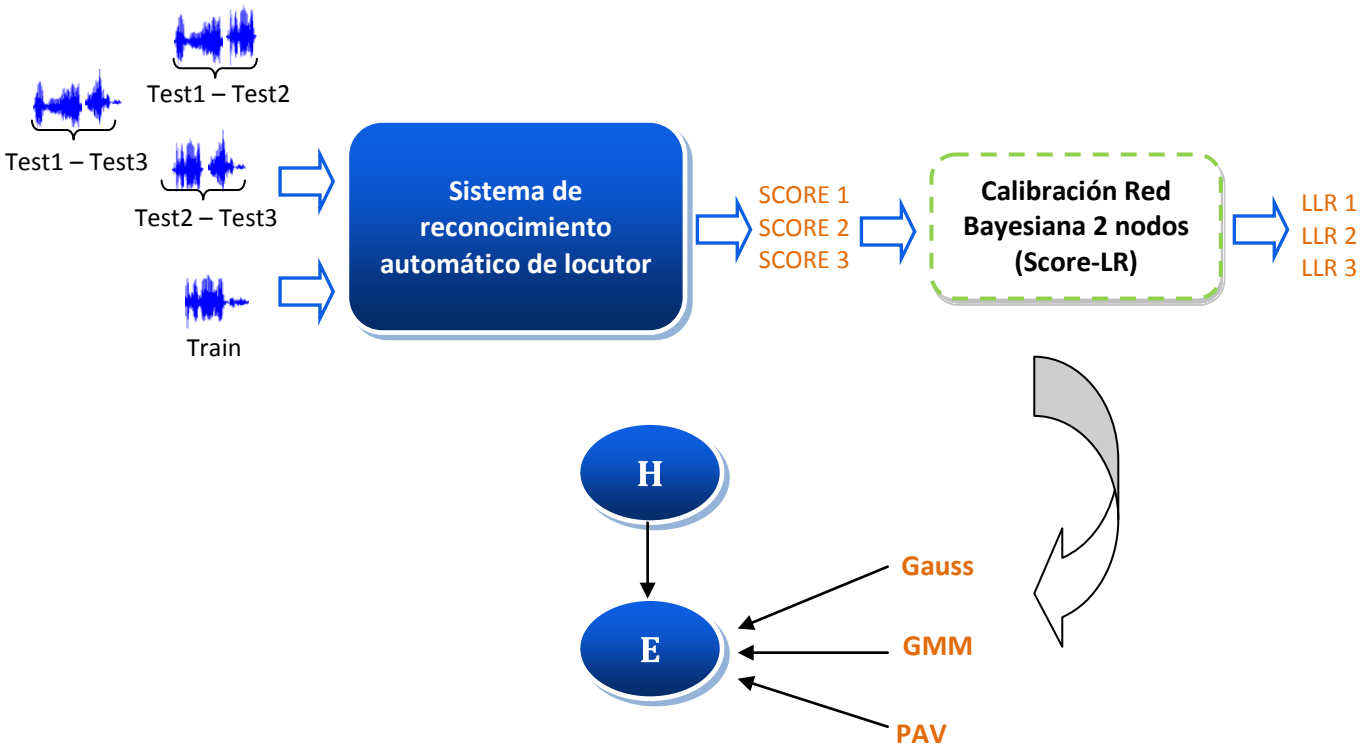


Figura 7.5: *Esquema de la red Bayesiana utilizada en el estudio de diferentes técnicas de calibración.*

- **Contribución: Implementación de una red Bayesiana para combinación de dos evidencias forenses.**

  Una vez estudiadas las diferentes estrategias de combinación y calibración, en esta parte del proyecto se pretende realizar la combinación y calibración en un solo paso mediante el uso de redes Bayesianas. En la figura 7.6. se muestra la metodología seguida para su implementación. Como entrada al sistema se tiene los diferentes archivos de audio individuales con los que se calculará la evidencia a combinar. Posteriormente, los *scores* de salida son combinados y calibrados mediante una red Bayesiana compuesta de tres nodos, el nodo hipótesis y los dos nodos de evidencias que se combinarán. El procedimiento será descrito en detalle en la sección 7.2.2.

  Finalmente, se hará un estudio comparativo de todas las estrategias de combinación: suma, concatenación de audio y la red Bayesiana de tres nodos.

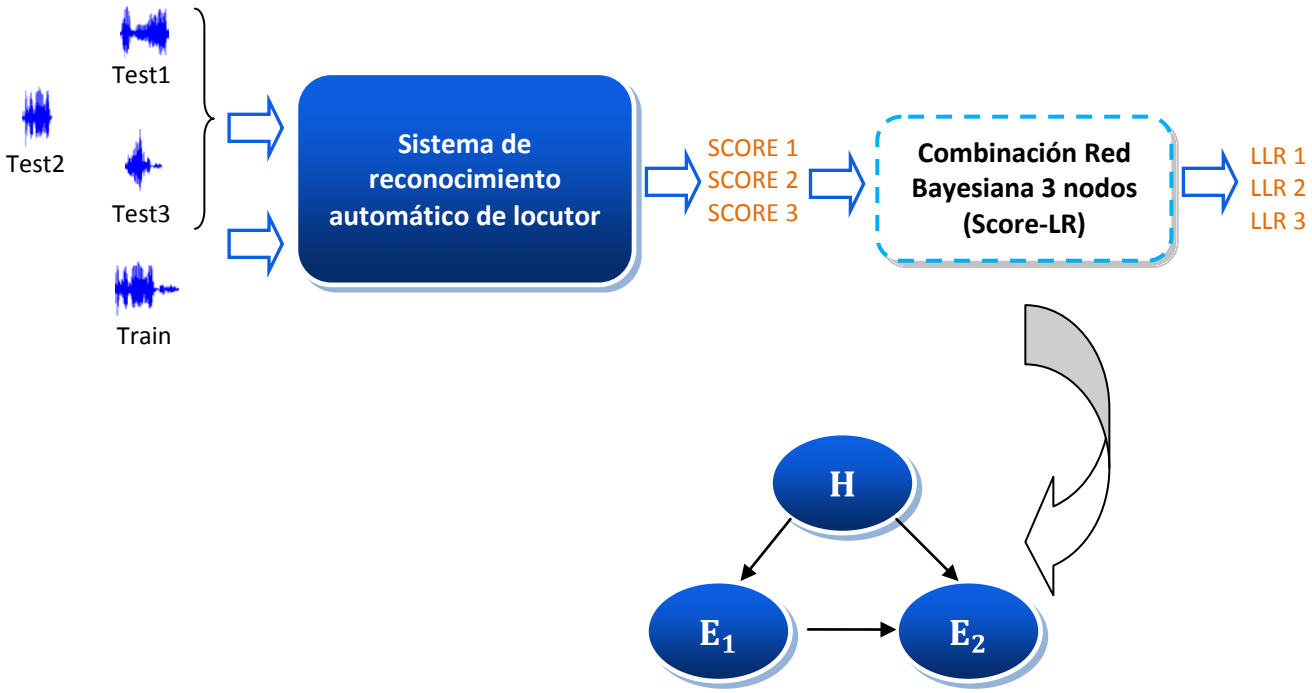


Figura 7.6: *Esquema de la red Bayesiana utilizada en el estudio de combinación de evidencias mediante modelado Gaussiano.*

La figura 7.7 muestra un esquema de relación entre los experimentos que forman parte de este proyecto: evaluación de estrategias de combinación de evidencias, evaluación de estrategias de calibración mediante redes Bayesianas, y una vez evaluados ambos métodos, evaluación de la posibilidad de utilizar redes Bayesianas para realizar ambos procedimientos: combinación y calibración.

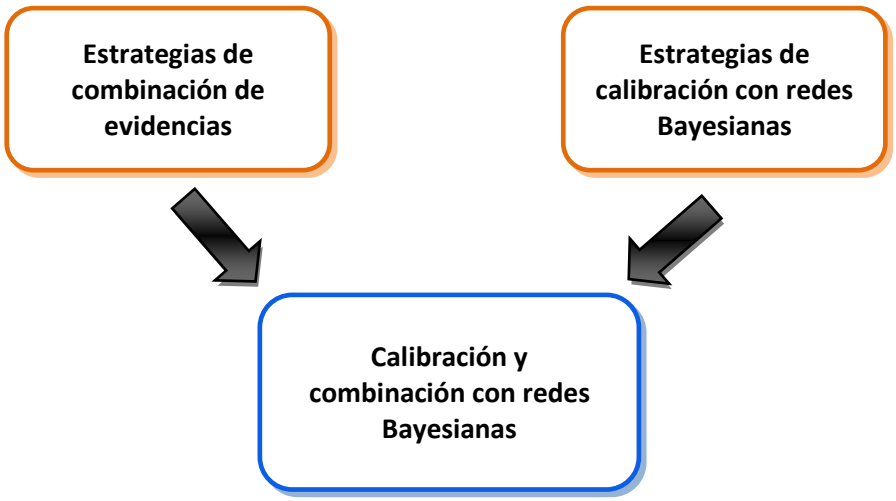


Figura 7.7: *Relación entre experimentos.*

A continuación, se presenta una explicación detallada de los diferentes experimentos realizados para combinar evidencias forenses y los resultados generados a partir de ellos.

7.1. Comparación de diferentes estrategias de combinación de evidencias

En esta sección se va a analizar el rendimiento de los sistemas de combinación propuestos en el primer punto de la parte experimental: suma pre-calibrada, suma post-calibrada y concatenación de archivos con combinaciones de 2 y 3 evidencias.

7.1.1 Combinación de evidencias mediante suma pre-calibrada y suma post-calibrada

Como se describió en la introducción del capítulo, existen diferentes posibilidades para combinar varios LR. La más inmediata es la suma directa de los LR en escala logarítmica, pero es necesario aplicar un proceso de normalización o calibración en alguna fase del proceso para ofrecer resultados lo más acordes posible a todas las fuentes de información disponibles. Se puede optar entonces por aplicar el proceso de calibración a los resultados individuales y sumar resultados ya calibrados, lo que genera un resultado global que puede no precisar ser calibrado de nuevo, si las fuentes de información son suficientemente independientes. Para el sistema T-norm compensado en locutor y canal este proceso genera los resultados mostrados en la figura 7.8.

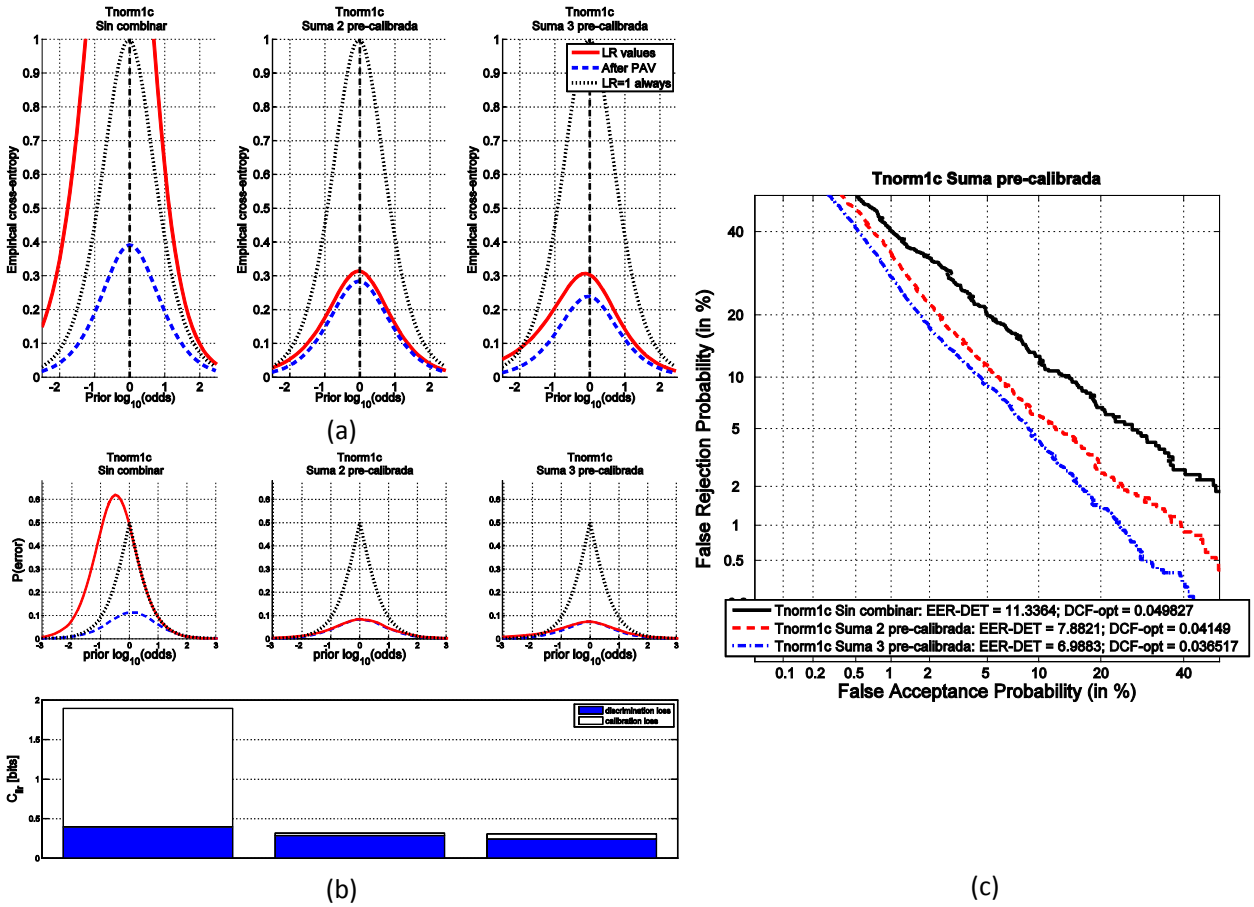


Figura 7.8: Curvas ECE (a), APE (b) y DET (c) comparativas del rendimiento ofrecido por el sistema antes de combinar y después de aplicar una combinación de evidencias mediante suma pre-calibrada de los log-LR de 2 en 2 y de 3 en 3.

Por el contrario, se puede optar por sumar en primera instancia los *scores* y aplicar a continuación un único proceso de calibración sobre el resultado global. De este procedimiento cabe esperar arrastrar el posible error de calibración de cada uno de los *scores* a sumar, así como la falta de normalización entre diferentes scores, que se soluciona en parte con el uso de técnicas como T-norm o Z-norm, sin embargo el tiempo de cálculo empleado será menor puesto que bastará con una sola calibración final. Este proceso genera los resultados que se muestran a continuación para el sistema T-norm compensado en locutor y canal para combinaciones de 2 y 3 log-LR.

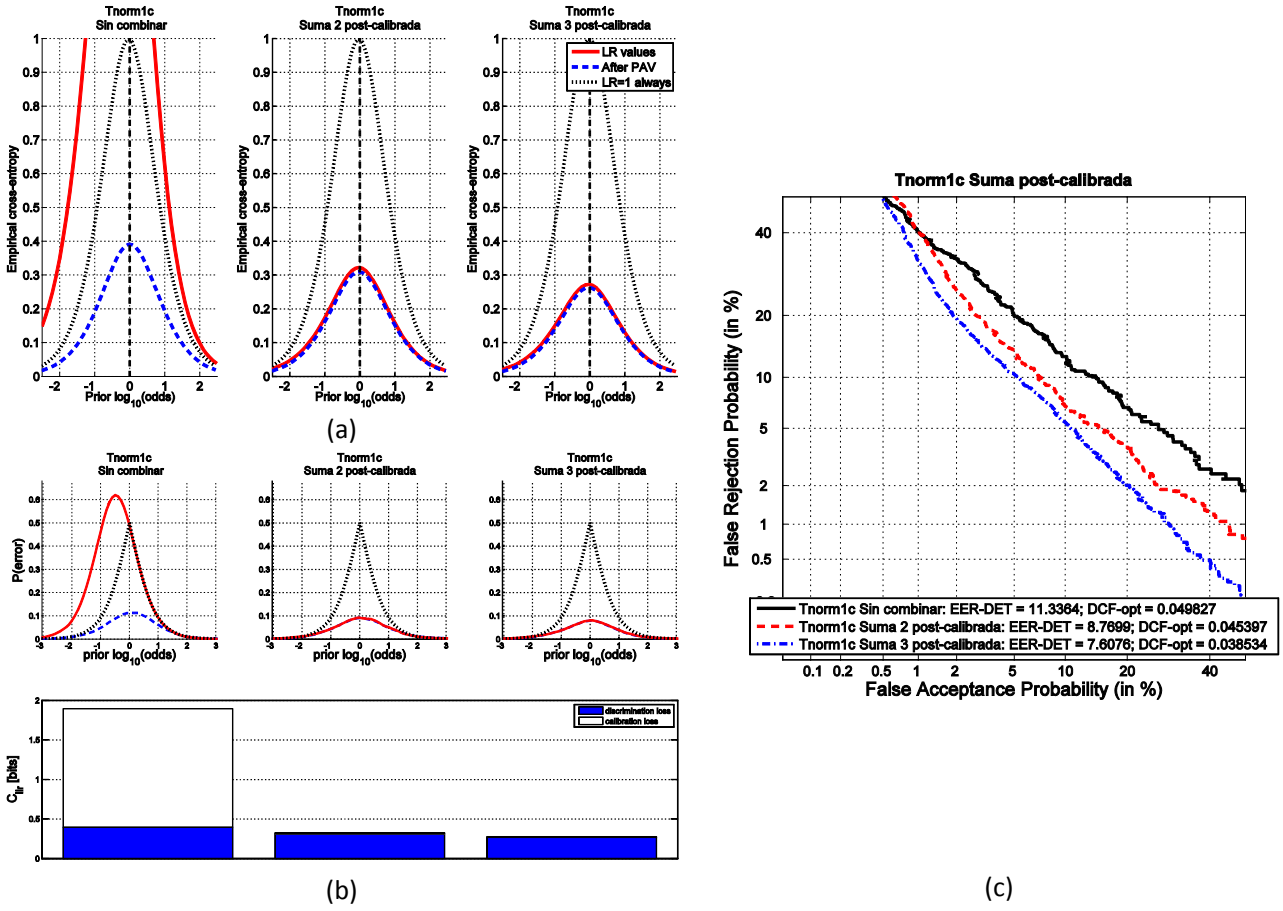


Figura 7.9: Curvas ECE (a), APE (b) y DET (c) comparativas del rendimiento ofrecido por el sistema antes de combinar y después de aplicar una combinación de evidencias mediante suma post-calibrada de los log-LR de 2 en 2 y de 3 en 3.

Se observa que empleando estos procedimientos de combinación, los resultados mejoran con respecto al sistema sin combinar debido a que se conoce más información sobre la identidad del locutor. Esto se representa por una curva DET más cercana al origen que el sistema sin combinar, o de manera análoga, una curva azul más baja en las curvas ECE o APE.

Por otra parte, se observa que en cuanto a discriminación, la calibración previa a la suma de LR ofrece mejor resultado que la suma post-calibrada pero empeora en cuanto a pérdidas de calibración a medida que aumenta el número de combinaciones, lo que implicaría que cuando se interpreten los log-LR de forma probabilística se van a cometer muchos y/o graves errores. Esto es debido a que se está suponiendo independencia estadística entre muestras y es una

presunción errónea, además, cuanto mayor es el número de combinaciones, mayor error se comente. Esto se representa en las curvas APE y ECE de las figuras 7.8 (a) y 7.8 (b) con una curva roja más lejana a la azul que en las figuras 7.9 (a) y 7.9 (b). Para solventar este error, sería necesaria una etapa adicional de post-calibrado implicando un tiempo de procesamiento adicional. Por lo tanto, es preferible utilizar suma post-calibrada porque la diferencia en cuanto a discriminación no es significativa y además, los resultados de calibración están más controlados.

	PRECALIBRADO		POSTCALIBRADO	
SISTEMA	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}
TNORM COMPENSADO (2)	0.287	0.317	0.313	0.326
TNORM COMPENSADO (3)	0.242	0.308	0.266	0.275

Tabla 7.1: *Resultados para los métodos de combinación suma pre-calibrada y suma post-calibrada.*

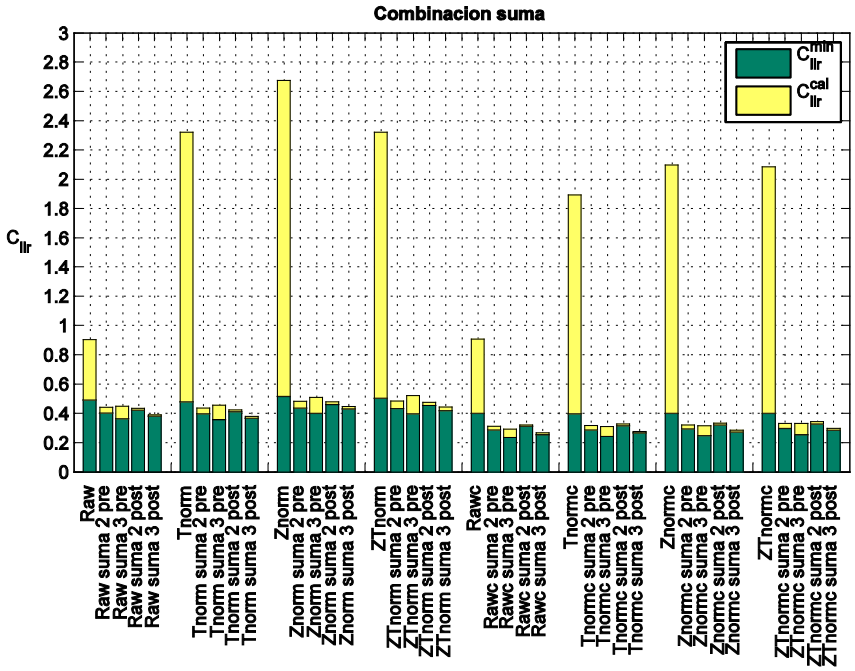


Figura 7.10: *Diagrama de barras representativo del rendimiento del sistema sin combinar (y sin calibrar) y combinado mediante las estrategias suma pre-calibrada y suma post-calibrada para combinaciones de 2 y 3 LRs. Se muestran los sistemas sin normalización, RAW, y con normalización T-norm, Z-norm, ZT-norm cada uno de ellos sin compensar y compensado en locutor y canal. La parte verde representa las pérdidas de discriminación y la amarilla las pérdidas de calibración.*

La figura 7.10. representa un resumen del comportamiento de los diferentes sistemas estudiados sin combinación de evidencias y combinado mediante las técnicas de suma vistas anteriormente, donde se puede observar que el sistema T-norm compensado para el esquema

de combinación suma post-calibrada es el que mejores resultados obtiene. Se muestra claramente como en la estrategia suma pre-calibrada, las pérdidas de calibración, representadas en amarillo, aumentan con el número de combinaciones, efecto que no ocurre en la suma post-calibrada.

7.1.2. Combinación de evidencias mediante concatenación de archivos de audio

Otro método de combinación propuesto en la introducción consiste en combinar los ficheros de audio correspondientes a las tomas dubitadas de la base de datos. En este esquema, se combinan las muestras de audio concatenando los archivos de 2 en 2 y de 3 en 3 y posteriormente, se utilizan los sistemas de reconocimiento de locutor, generando un *score* por cada comparación de la toma indubitada con cada combinación generada. Por último se calibran los resultados mediante regresión logística. Este proceso genera los resultados que se muestran a continuación para el sistema T-norm compensado en locutor y canal.

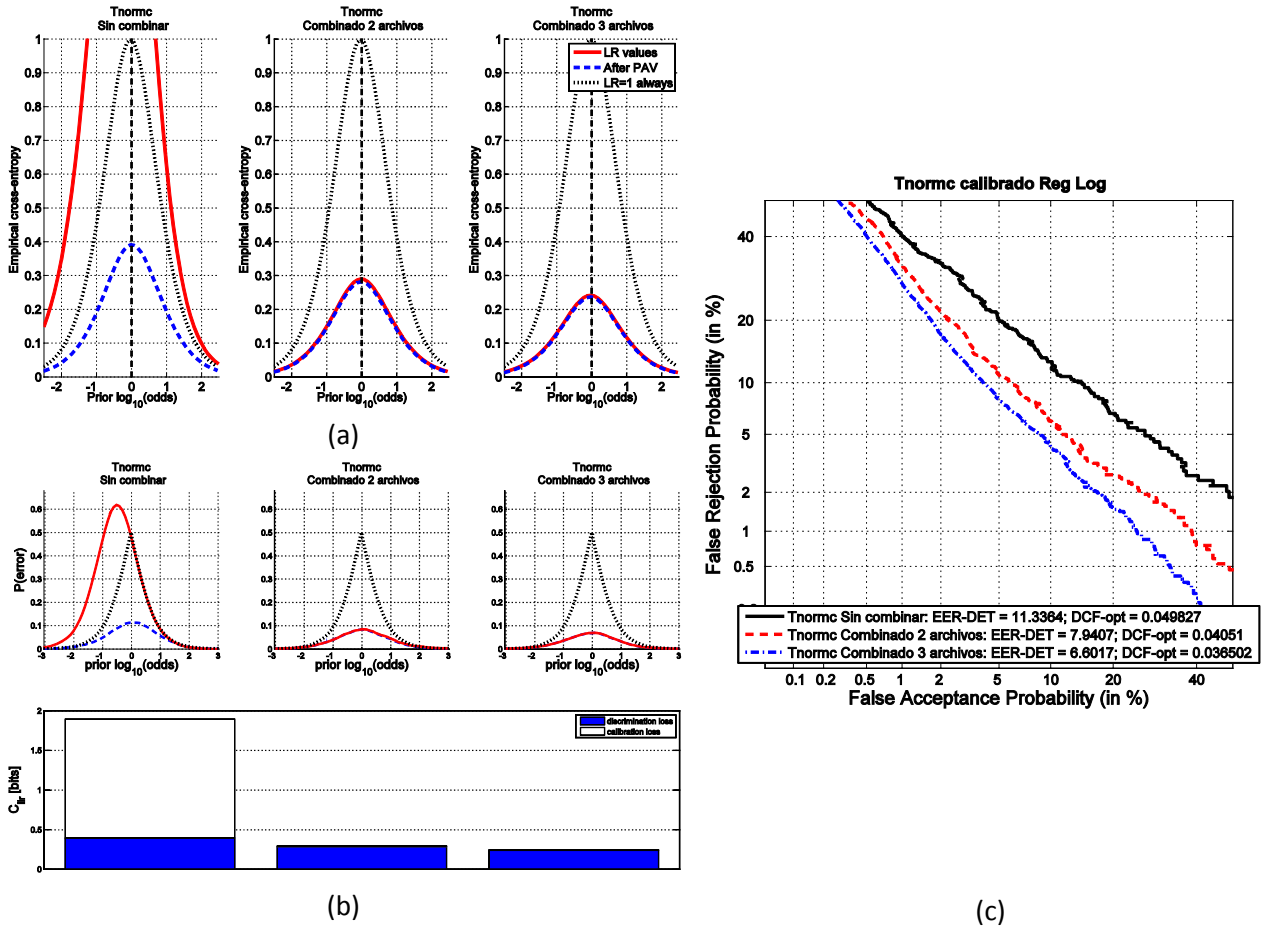


Figura 7.11: *Curvas ECE (a), APE (b) y DET (c) comparativas del rendimiento ofrecido por el sistema antes de combinar y después de aplicar una combinación de evidencias mediante concatenación de 2 y 3 archivos.*

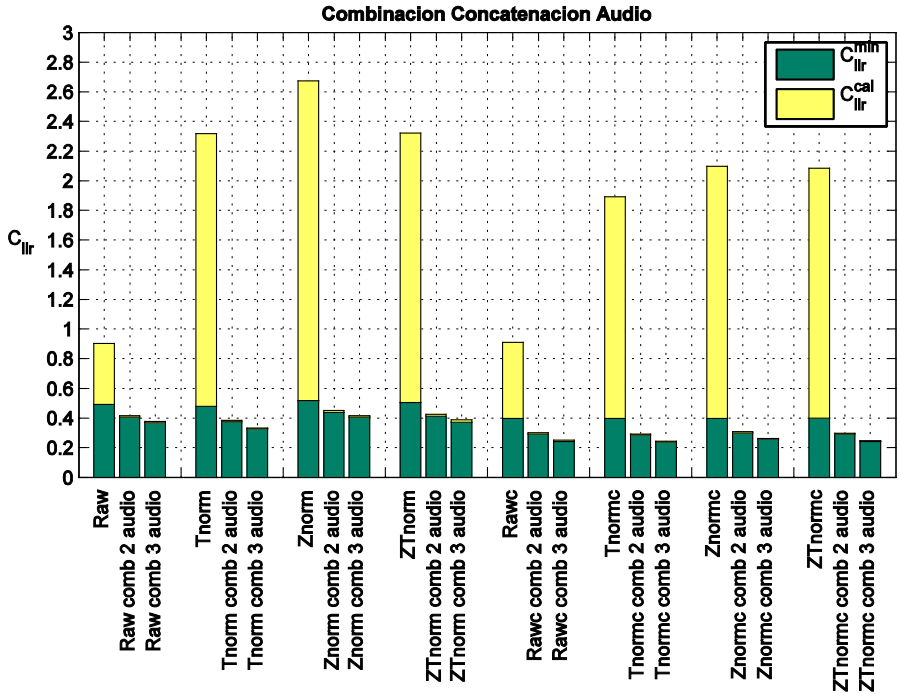


Figura 7.12: *Diagrama de barras representativo del rendimiento del sistema sin combinar (y sin calibrar) y combinado mediante la estrategia de concatenación de audio para combinaciones de 2 y 3 archivos. Se muestran los sistemas sin normalización, RAW, y con normalización T-norm, Z-norm, ZT-norm cada uno de ellos sin compensar y compensado en locutor y canal.*

Se observa que empleando este método de combinación, los resultados mejoran con respecto al sistema sin combinar, produciéndose una mejora mayor con mayor número de combinaciones. Al combinar los archivos, aumenta el poder de discriminación, representado por una curva DET más cercana al origen, pero las pérdidas de calibración aumentan y esto hace que no sean interpretables probabilísticamente, como se comentó anteriormente. Para solucionarlo, se ha realizado una calibración posterior mediante regresión logística y como se puede observar en la gráfica de barras de la figura 7.12 las pérdidas de calibración son prácticamente nulas.

Comparación de los métodos de combinación suma y concatenación

A continuación se realiza una comparación entre los resultados obtenidos a través de las diferentes técnicas de combinación vistos hasta ahora: suma pre-calibrada, suma-post-calibrada y concatenación de audio para combinaciones de 2 y 3 muestras. Para ello nos basaremos en las figuras 7.13 y 7.14 Y los resultados mostrados por la tabla 7.2 Donde C_{llr}^{min} representa el poder discriminativo y $C_{llr} - C_{llr}^{min}$ representan las pérdidas de calibración de cada sistema.

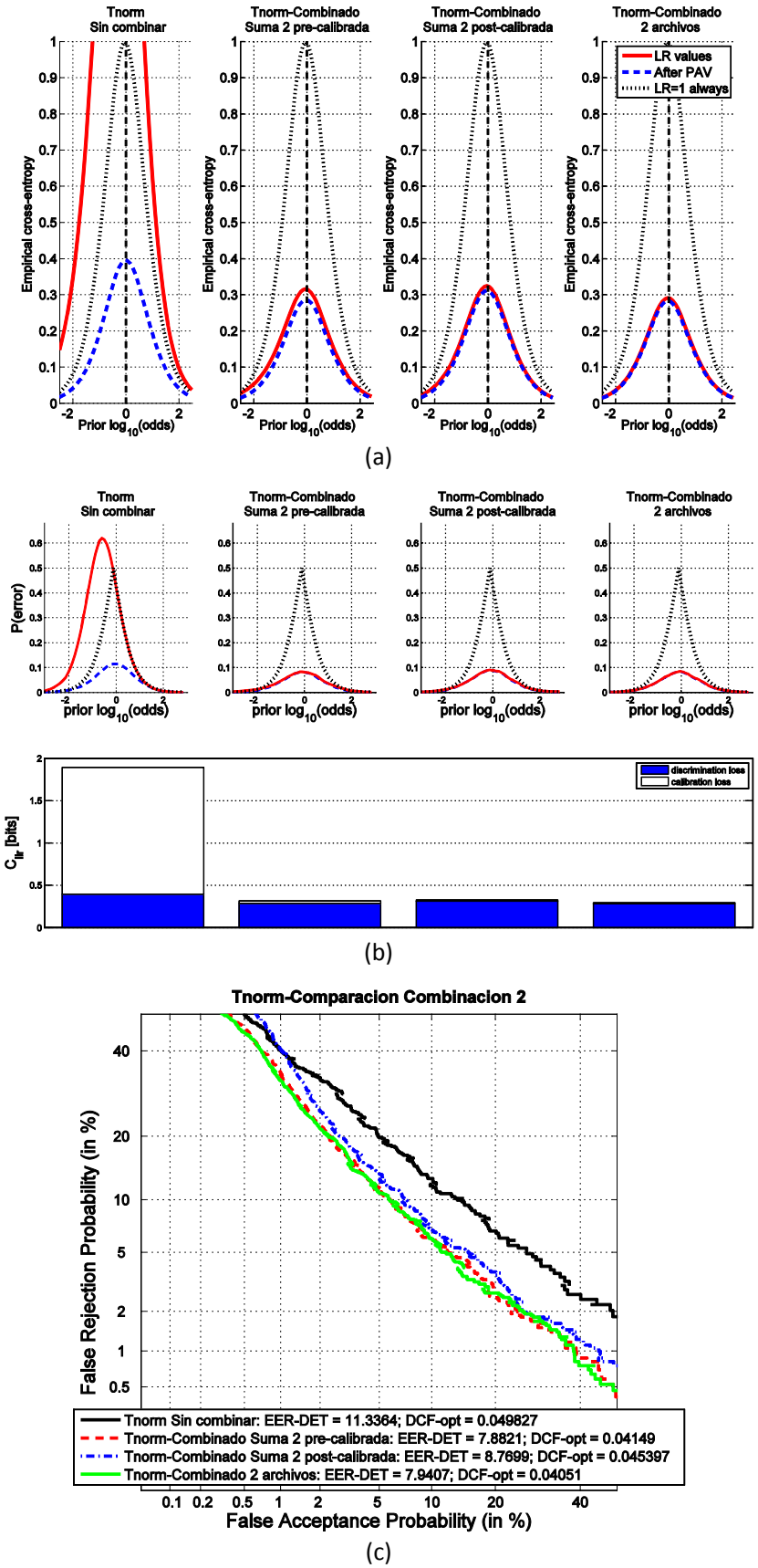


Figura 7.13: *Curvas ECE (a), APE (b) y DET (c) comparativas del rendimiento ofrecido por el sistema antes de combinar y después de aplicar una combinación de evidencias mediante suma pre y post-calibrada de 2 LRs y concatenación de 2 archivos.*

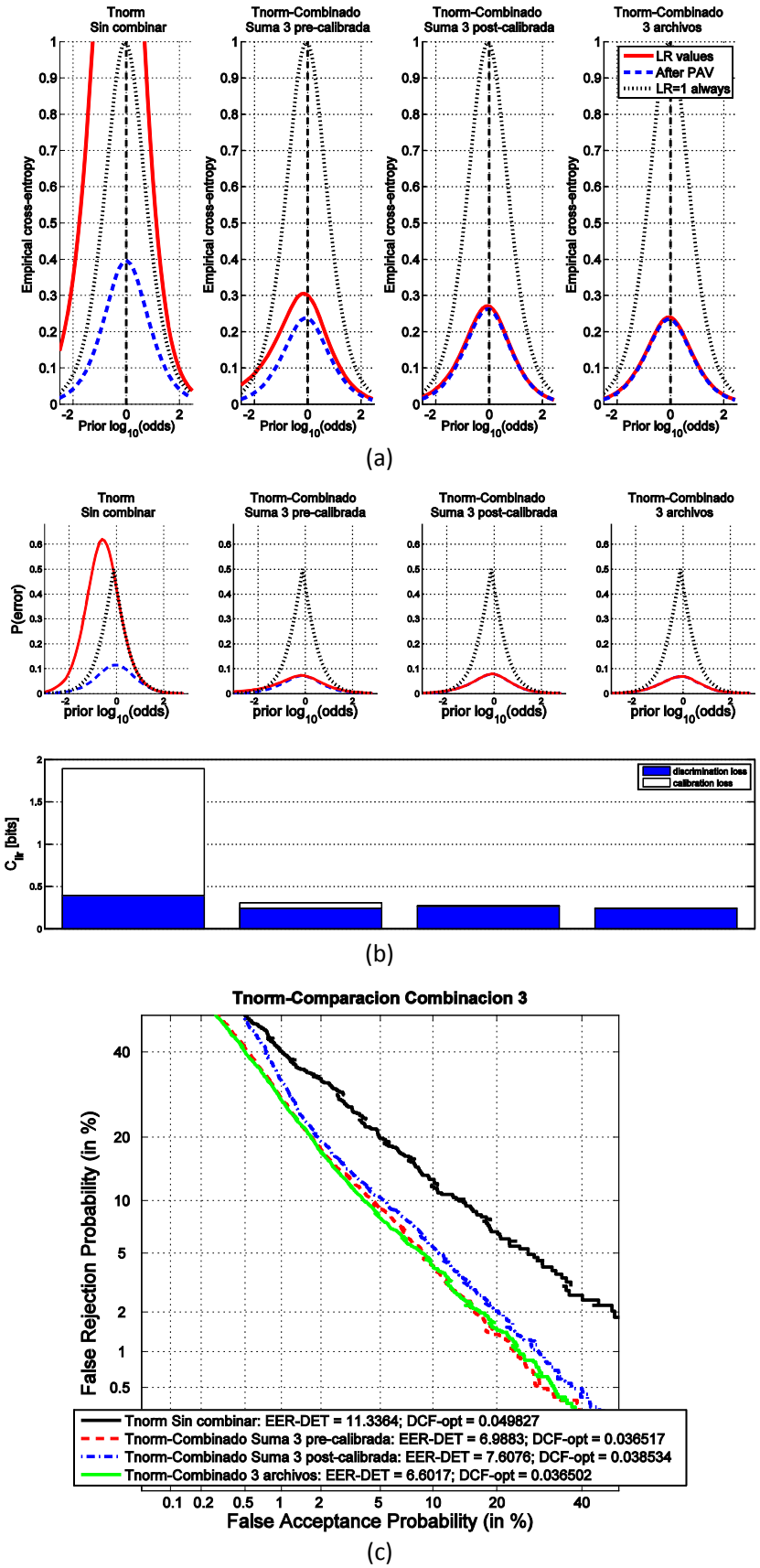


Figura 7.14: *Curvas ECE (a), APE (b) y DET (c) comparativas del rendimiento ofrecido por el sistema antes de combinar y después de aplicar una combinación de evidencias mediante suma pre y post-calibrada de 3 LRs y concatenación de 3 archivos.*

	PRECALIBRADO		POSTCALIBRADO		COMBINACIÓN AUDIO	
SISTEMA	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}
SIN COMBINAR	0.395	1.893	0.395	1.893	0.395	1.893
TNORM COMPENSADO (2)	0.287	0.317	0.313	0.326	**0.285**	**0.293**
TNORM COMPENSADO (3)	0.242	0.308	0.266	0.275	**0.240**	**0.244**

Tabla 7.2: *Comparación del rendimiento ofrecido por las diferentes técnicas de combinación de 2 y 3 evidencias.*

Se observa que empleando cualquiera de los procedimientos de combinación vistos, los resultados mejoran con respecto al sistema sin combinar, dando lugar a valores de EER y C_{llr}^{min} menores. Esto era de esperar debido a que cuantos más ficheros de dubitada se utilicen, más información de la identidad del locutor se puede extraer. Se representa por una curva DET más cercana al origen que el sistema sin combinar, y una curva azul más baja en las curvas ECE o APE.

Como ya se mencionó anteriormente, el mejor sistema de combinación mediante suma resulta ser la suma post-calibrada ya que a pesar de obtener resultados de discriminación ligeramente peores, las pérdidas de calibración son prácticamente nulas. Las pérdidas de calibración en la suma pre-calibrada son producidas por la falsa presunción de independencia de las muestras, y éstas son mayores a medida que aumenta el número de combinaciones ya que se estaría cometiendo más error. Para que la suma pre-calibrada obtuviera un rendimiento similar a la suma post-calibrada, sería necesario compensar las pérdidas de calibración, mediante una etapa de post-calibrado adicional dando lugar a un tiempo de procesamiento mayor.

Si se compara el procedimiento de combinación de suma pre y post-calibrada con la combinación mediante archivos de audio, se observa que este último procedimiento obtiene mejores resultados que los anteriores, esto puede deberse a que la combinación de la información de la identidad se lleva a cabo en el nivel de las características MFCC, que se modela de forma más compleja que el modelado que pueda hacerse de los *scores*.

Sin embargo esta técnica puede no ser del todo adecuada ya que se pueden estar combinando archivos de diferentes sesiones haciendo que el modelado de variabilidad posterior sea poco representativo y la aplicación de la técnica de compensación de variabilidad mediante JFA no sea óptima. Se ha de destacar que este importante problema no se manifiesta en los experimentos realizados, debido a que la base de datos utilizada no dispone de una suficiente cantidad de datos como para representar variabilidad fuera del espacio que se modela con JFA a partir de las grandes bases de datos de desarrollo utilizadas para entrenarlo (basadas en evaluaciones NIST).

Por ello, se propone una tercera técnica de combinación de evidencias mediante el uso de redes Bayesianas que no presente los inconvenientes de las técnicas de combinación vistas hasta ahora.

7.2. Combinación de evidencias con Redes Bayesianas

En esta sección se evaluará el rendimiento de los sistemas propuestos en el segundo punto de la parte experimental. En primer lugar, el uso de redes Bayesianas como método de calibración de *scores* a partir de una red de un nodo de evidencias previamente combinadas mediante la concatenación de archivos de audio y entrenada con diferentes técnicas: Modelado Gaussiano, GMM y PAV; y en segundo lugar, se implementará una red Bayesiana de 3 nodos para combinación de evidencias individuales mediante la técnica más sencilla: modelado Gaussiano.

Para ello, se recordará el significado del LR y su cálculo a partir de sistemas basados en *scores*. Posteriormente se explicarán con detalle las diferentes técnicas utilizadas y el proceso de implementación de redes Bayesianas a partir de ellas.

Tal y como puede observarse en la figura 4.2, los *scores* del sistema de reconocimiento de locutor son utilizados para calcular los LRs mediante la ecuación 4.4:

$$LR = \left. \frac{p(e|H_p, I)}{p(e|H_d, I)} \right|_{e=E}$$

El valor del LR, es el cociente entre dos probabilidades, y se puede demostrar que es igual al cociente de dos densidades de probabilidad en el caso de variables continuas. El numerador, $p(e|H_p, I)$, representa una estimación de la intra-variabilidad (*within-source*) del sistema y muestra la variabilidad de evidencias correspondientes a muestras pertenecientes a la misma fuente. Esta función, se modela a partir de *scores target* asumiendo H_p cierta. Utilizando validación cruzada, los *scores* utilizados son los correspondientes al conjunto de datos de entrenamiento *scores target train.*

Por otro lado, el denominador $p(e|H_d, I)$, representa una estimación de la inter-variabilidad (*between-source*) del sistema, y muestra la tipicidad o rareza de la muestra incriminatoria con respecto a una población relevante. Este modelo de población se especifica a partir de la información que proporciona las circunstancias del caso. Se obtiene a partir de *scores non-target* asumiendo H_d cierta. Utilizando validación cruzada, los *scores* utilizados son los correspondientes al conjunto de datos de entrenamiento *scores non-target train.*

En la figura 7.15 se muestra un esquema representativo del procedimiento seguido para la realización de los experimentos de esta sección. A continuación se describen cada una de las fases:

1. En primer lugar, se hace uso de la función validación cruzada para separar la base de datos en un conjunto de datos de entrenamiento y un conjunto de datos de *test.* Habrá tantos conjuntos de datos, y por lo tanto, estimación de distribuciones diferentes, como 69/N siendo N el número de divisiones especificado para la función validación cruzada y 69 el número total de locutores de la base de datos.

2. Para modelar la fdp correspondiente a la inter-variabilidad $p(e|H_p, I)$, se hace uso del conjunto de datos *scores target train* y para modelar la fdp $p(e|H_p, I)$, que se corresponde con la intra-variabilidad, se utiliza el conjunto de datos *scores non-target train.* La información relativa a estas distribuciones de probabilidad, se exporta a Hugin a través de un archivo de texto.

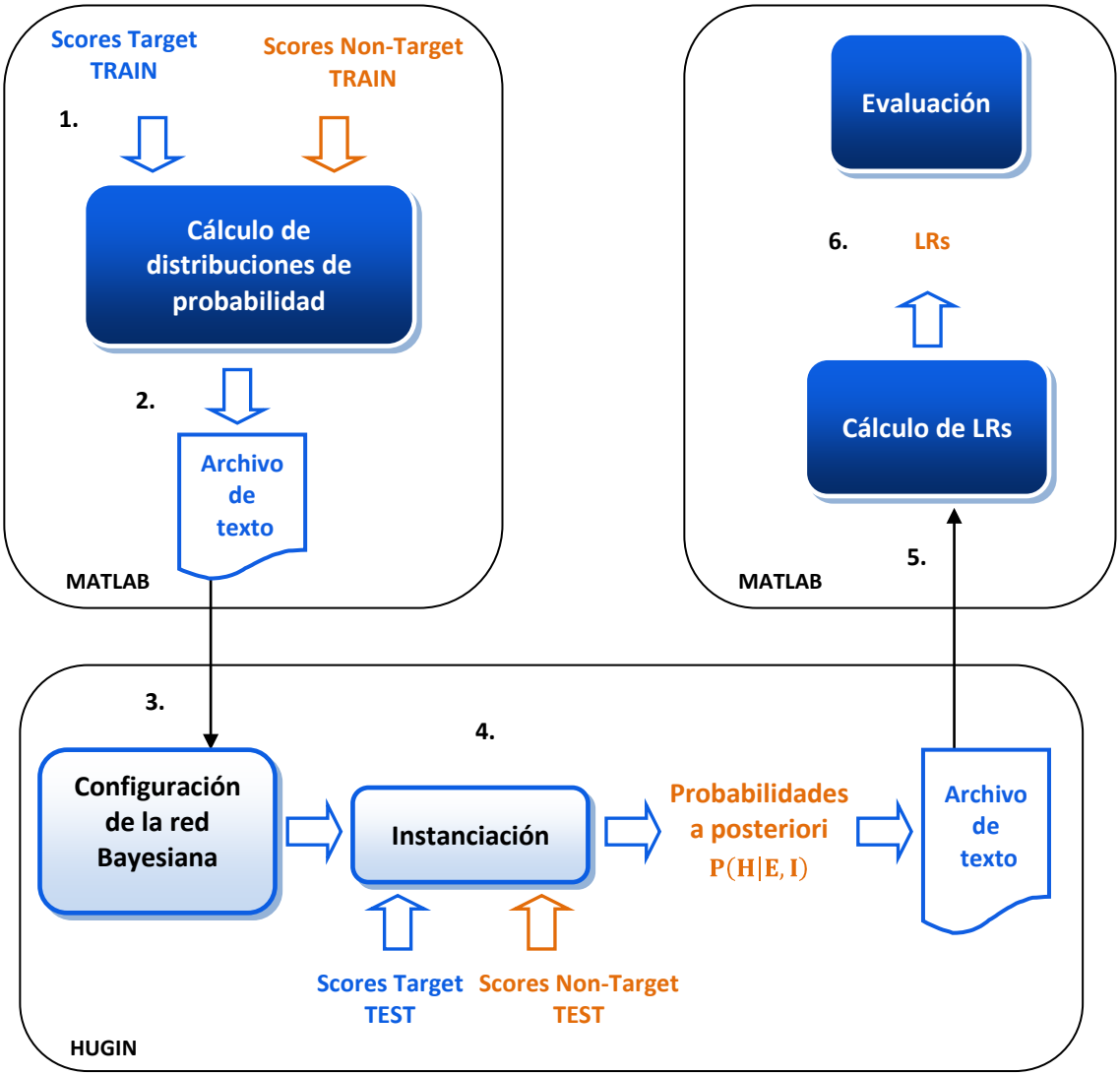


Figura 7.15: *Esquema del procedimiento seguido para la realización de los experimentos utilizando redes Bayesianas.*

3. A partir de las fdps calculadas anteriormente, se generan las tablas de probabilidad para cada nodo con las que se configurará la red Bayesiana.

4. Instanciación del nodo E a partir del conjunto de datos de *test.* Cuando se instancia el nodo con las evidencias observadas, se modifican sus tablas de probabilidad y, a su vez, las nuevas probabilidades son propagadas al nodo H. Esta propagación de probabilidades se conoce como inferencia probabilística y da lugar a las probabilidades *a posteriori* con las que se calculará el LR.

5. Las probabilidades *a posteriori* obtenidas en el paso anterior, se exportan a Matlab a través de un archivo de texto. Con esta información se calcularán los valores de LR posteriormente. Sabiendo que las probabilidades *a priori* $P(H_p)$ = $P(H_d)$ = 0.5, bastaría con calcular el cociente de probabilidades *a posteriori* para una evidencia dada.

$$\frac{P(H_p|E,I)}{P(H_d|E,I)} = LR \cdot \frac{P(H_p|I)}{P(H_d|I)}, \qquad \frac{P(H_p|I)}{P(H_d|I)} = 1 \Rightarrow \boxed{LR = \frac{P(H_p|E,I)}{P(H_d|E,I)}} \tag{7.1}$$

6. Por último, se medirá el rendimiento de cada uno de los sistemas mediante los métodos explicados en la sección 4.5.

Esta metodología, es común para cada uno de los métodos de calibración, difiriendo únicamente en los pasos 2 y 3. En ellos, la distribución de probabilidad será calculada por los diferentes métodos a estudiar: Modelado Gaussiano, GMM y PAV, dando lugar a sendas tablas de probabilidad utilizadas en la configuración de la red Bayesiana. En las próximas secciones se explicará el procedimiento para implementar cada uno de estos 3 métodos de calibración.

7.2.1. Evaluación del uso de diferentes métodos de calibración con redes Bayesianas

A continuación se describirá en detalle el procedimiento seguido para hacer posible la utilización de redes Bayesianas como método de calibración de *scores.* La red Bayesiana utilizada se muestra en la figura 7.16 y se entrenará mediante diferentes técnicas: modelado Gaussiano, GMM y PAV. La evidencia, será la puntuación de salida del sistema de reconocimiento de locutor, cuya entrada son los archivos concatenados correspondientes al esquema de combinación mediante concatenación (figura 7.5).

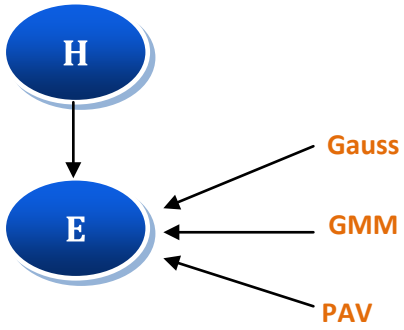


Figura: 7.16: *Red genérica utilizada para calibración.*

- **Modelado Gaussiano**

El modelado Gaussiano de calibración es el más sencillo de implementar. Consiste en modelar las funciones densidad de probabilidad involucradas en el cálculo de LRs a partir de *scores*, mediante funciones Gaussianas o normales. En los siguientes apartados se describirán métodos más complejos en los que el cálculo de probabilidades para configurar la red, no es tan inmediato.

La red bayesiana que representa este escenario viene dada como:

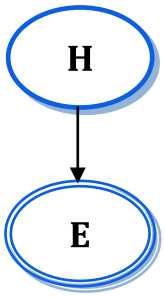


Figura: 7.17: *Red utilizada para calibración mediante modelado Gaussiano*

Está formada por las dos variables H y E que representan las dos variables aleatorias presentes en la ecuación del LR y que se utilizará en el proceso de calibración. Por lo tanto se deben especificar las siguientes probabilidades:

- $P(H_p) = P(H_d) = 0.5$
- $p(E|H_p)$ y $p(E|H_d)$

  En Hugin, se asume que los nodos continuos tienen una distribución Gaussiana, por tanto, en este caso bastará con especificar una media (μ) y una varianza (σ^2) para cada configuración del nodo padre (H_p o H_d). Por lo tanto, para cada conjunto de datos *train* (*target y non-target*) se debe calcular un dato de media y varianza que posteriormente serán utilizados para configurar la red Bayesiana correspondiente. Así las probabilidades $p(E|H_p)$ y $p(E|H_d)$ quedan representadas por sendas funciones densidad de probabilidad gaussianas:

$$p(E|H_p) = p\left(E|\mu_{Target}, \sigma_{Target}\right) \equiv \frac{1}{2\pi^{1/2}\sigma_{target}} e^{-\frac{1}{2}\left(\frac{x-\mu_{target}}{\sigma_{target}}\right)} \quad (7.2)$$

$$p(E|H_d) = p\left(E|\mu_{Non-Target}, \sigma_{Non-Target}\right) \equiv \frac{1}{2\pi^{1/2}\sigma_{non-target}} e^{-\frac{1}{2}\left(\frac{x-\mu_{non-target}}{\sigma_{non-target}}\right)} \quad (7.3)$$

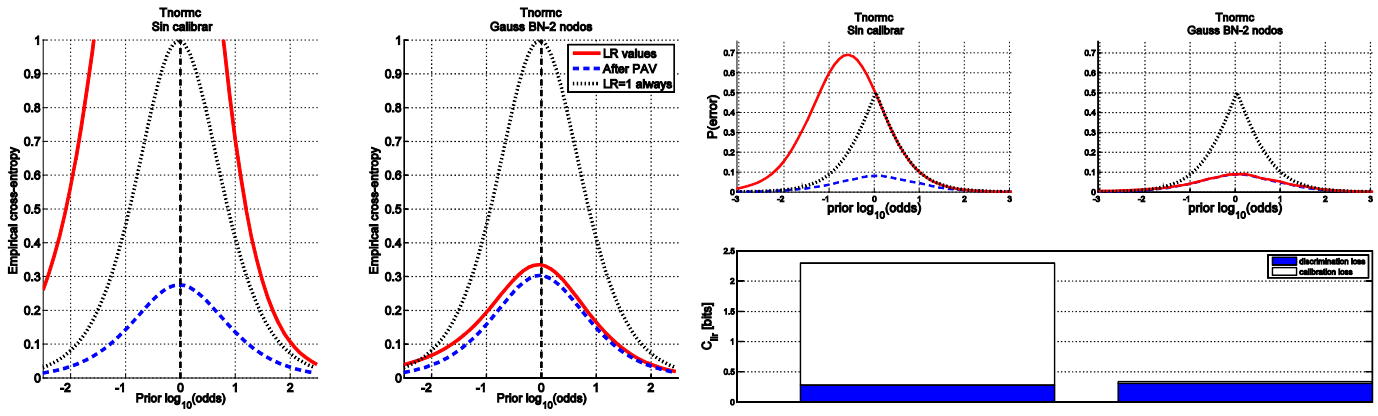
H	
H_p	0.5
H_d	0.5

Tabla 7.3: *Tabla de probabilidad del nodo H para modelado Gaussiano.*

	E	
H	H_p	H_d
Media (μ)	μ_{target}	$\mu_{non-target}$
Varianza (σ^2)	σ^2_{target}	$\sigma^2_{non-target}$

Tabla 7.4: *Tabla de probabilidad del nodo E condicionado al nodo H, para modelado Gaussiano.*

Para el sistema T-norm compensado en locutor y canal calibrado mediante modelado Gaussiano con redes Bayesianas se generan los resultados mostrados en la figura 7.18.



(a) (b)

Figura 7.18: *Curvas ECE (a) y APE (b) comparativas del rendimiento ofrecido por el sistema antes de calibrar y después de calibrar mediante modelado Gaussiano.*

En la figura 7.18, se puede observar que aunque los resultados de calibración son muy buenos, todavía existen algunas pérdidas de calibración, representado gráficamente como una separación entre las curvas roja y azul.

• GMM

La motivación fundamental del uso de GMMs se basa en que una combinación lineal de funciones base gaussianas es capaz de representar un gran abanico de posibles distribuciones muestrales, tales como los diferentes *scores* de salida del sistema de reconocimiento de locutor. Como ya se vio en la sección 3.4.1, la función densidad de probabilidad puede expresarse mediante la ecuación 3.1, que adaptado al problema de atribución de fuentes puede expresarse como:

$$p(E|H_p) = \sum_{i=1}^{M} w_{target\,i} \cdot p(E|\mu_{target\,i}, \sigma_{target\,i}) \tag{7.4}$$

$$p(E|H_d) = \sum_{i=1}^{M} w_{non-target\,i} \cdot p(E|\mu_{non-Target\,i}, \sigma_{non-Target\,i}) \tag{7.5}$$

donde M es el número de gaussianas componentes, w_i son los pesos de cada una de las mezclas donde debe cumplirse $\sum_{i=1}^{M} w_i$ = 1 y $P(E|\mu_i, \sigma_i)$ son las M densidades componentes para cada hipótesis H, definidas en las ecuaciones 7.2 y 7.3. En este proyecto, el valor usado como número de mezclas es M = 16.

En este caso, se plantea el problema de cómo representar un GMM con Hugin, ya que para variables continuas, sólo admite funciones densidad de probabilidad Gaussianas. Aquí, como indican las ecuaciones 7.4 y 7.5, la función densidad de probabilidad del nodo E viene dada por una mezcla de funciones para cada estado de H, por ello, se necesita introducir un nodo adicional llamado "selector", que indica los pesos de cada una de las funciones densidad componentes del GMM. Por tanto, la red Bayesiana utilizada para implementar el método GMM en Hugin es:

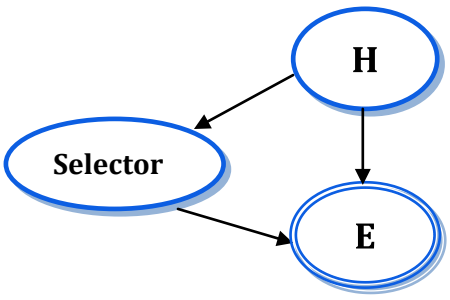


Figura: 7.19: *Red utilizada para calibración mediante GMM*

Las probabilidades a calcular en este escenario serían por tanto:

- $P(H_p)$ = $P(H_d)$ = 0.5

- $P(selector|H_p)$ y $P(selector|H_d)$ que se corresponden con los diferentes vectores de peso de dimensión 1x16 para cada una de las hipótesis.

- $p(E|H_p)$ y $p(E|H_d)$ representadas por sendas funciones densidad de probabilidad GMM, cada una con su vector de dimensión 1x16 de pesos, medias y varianzas correspondiente:

$$p(E|H_p) = \sum_{i=1}^{M} w_{target\ i} \cdot p(E|\mu_{target\ i}, \sigma_{target\ i}) \tag{7.6}$$

$$p(E|H_d) = \sum_{i=1}^{M} w_{non-target\ i} \cdot p(E|\mu_{non-Target\ i}, \sigma_{non-Target\ i}) \tag{7.7}$$

Con $i = 0 \dots 15$ para ambos casos

Así las tablas de probabilidad quedarían como:

H	
H_p	0.5
H_d	0.5

Tabla 7.5: *Tabla de probabilidad del nodo H para GMM.*

	Selector	
H	H_p	H_d
0	$w_{target\ 0}$	$w_{non-target\ 0}$
⋮	⋮	⋮
15	$w_{target\ 15}$	$w_{non-target\ 15}$

Tabla 7.6: *Tabla de probabilidad del nodo Selector condicionada al nodo H, para GMM.*

	E				
H	H_p	H_d		H_p	H_d
Media (μ)	$\mu_{target\ 0}$	$\mu_{non-target\ 0}$	…	$\mu_{target\ 15}$	$\mu_{non-target\ 15}$
Varianza (σ^2)	$\sigma^2_{target\ 0}$	$\sigma^2_{non-target\ 0}$	…	$\sigma^2_{target\ 15}$	$\sigma^2_{non-target\ 15}$

Tabla 7.7: *Tabla de probabilidad del nodo E condicionada al nodo H y el valor del nodo Selector, para GMM.*

Para el sistema T-norm compensado en locutor y canal calibrado mediante modelado Gaussiano con redes Bayesianas se generan los resultados mostrados en la figura 7.20. En ella, se puede observar que los resultados de calibración son muy buenos, ya que cuanto más cercanas estén la curva azul y la roja, mejor calibrado estará el sistema.

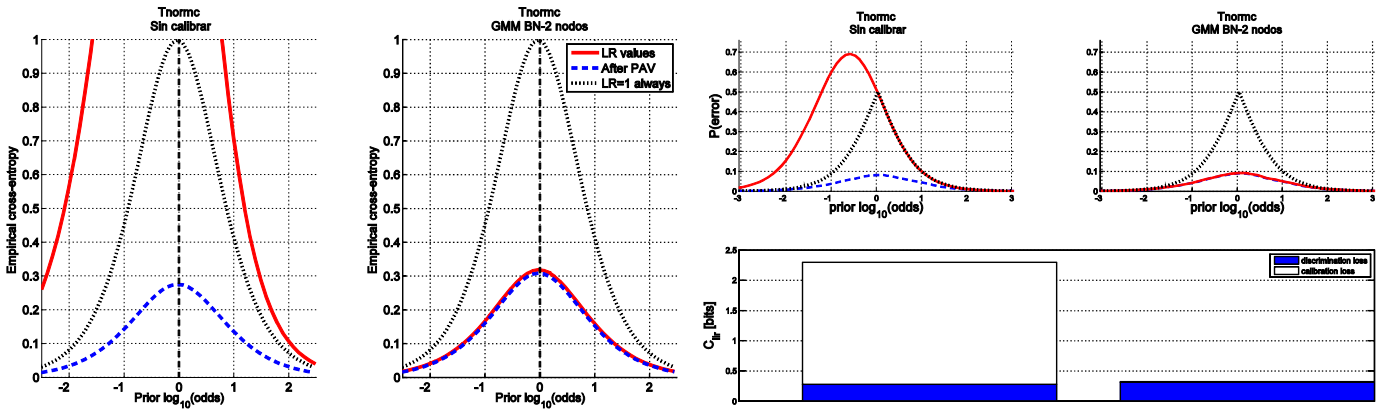


Figura 7.20: *Curvas ECE (a) y APE (b) comparativas del rendimiento ofrecido por el sistema antes de calibrar y después de calibrar mediante GMM.*

• PAV

La técnica de calibración PAV es un método basado en histogramas donde las diferentes regiones que lo componen son variables. En cada una de esas regiones, se sigue un criterio de entrenamiento en el que se tiene que mantener una determinada función de coste monótona creciente como ya se describió en la sección 4.4.4. En la figura 7.21. se puede observar un ejemplo de histograma PAV, donde el eje de abscisas representa los diferentes intervalos generados por PAV y el eje de ordenadas, representa la probabilidad de pertenencia de un *score* dado en cada una de las regiones, para la hipótesis H_p cierta.

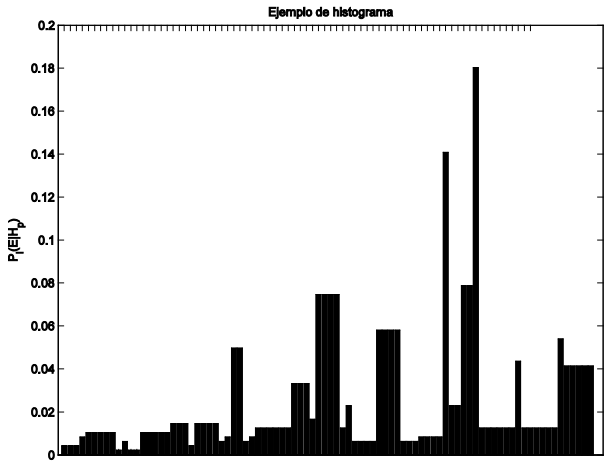


Figura 7.21: *Ejemplo de histograma PAV.*

Para especificar este tipo de distribuciones en Hugin, es necesario configurar el nodo E como discreto de tipo *interval node* que permite indicar el valor de la función de probabilidad por regiones tal y como se requiere para la implementación de este método de calibración. En este proyecto, se utiliza la versión de evaluación de Hugin y existe una limitación de 48 intervalos por lo que se ha debido adaptar el histograma de probabilidad de manera que no superara este número máximo de regiones.

La red Bayesiana en este escenario se representa como:

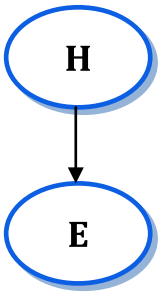


Figura: 7.22: *Red utilizada para calibración mediante PAV.*

Teniendo en cuenta esta estructura, las probabilidades a calcular son las siguientes:

- $P(H_p)$ = $P(H_d)$ = 0.5.
- $P(E|H_p)$ y $P(E|H_d)$

El cálculo de la función de probabilidad se realiza intervalo a intervalo, estableciendo un valor de probabilidad en cada uno de los existentes. Este valor se calcula como el número de *scores* perteneciente a una determinada clase en un intervalo dado, dividido entre el número total de *scores* de esa misma clase en el conjunto de *scores* total. Por tanto se tiene para cada región:

$$P_i(E|H_p) = \frac{Número\ Scores\ Target}{Número\ Total\ Scorres\ Target} \tag{7.8}$$

$$P_i(E|H_d) = \frac{Número\ Scores\ Non{-}Target}{Número\ Total\ Scores\ Non{-}Target} \tag{7.9}$$

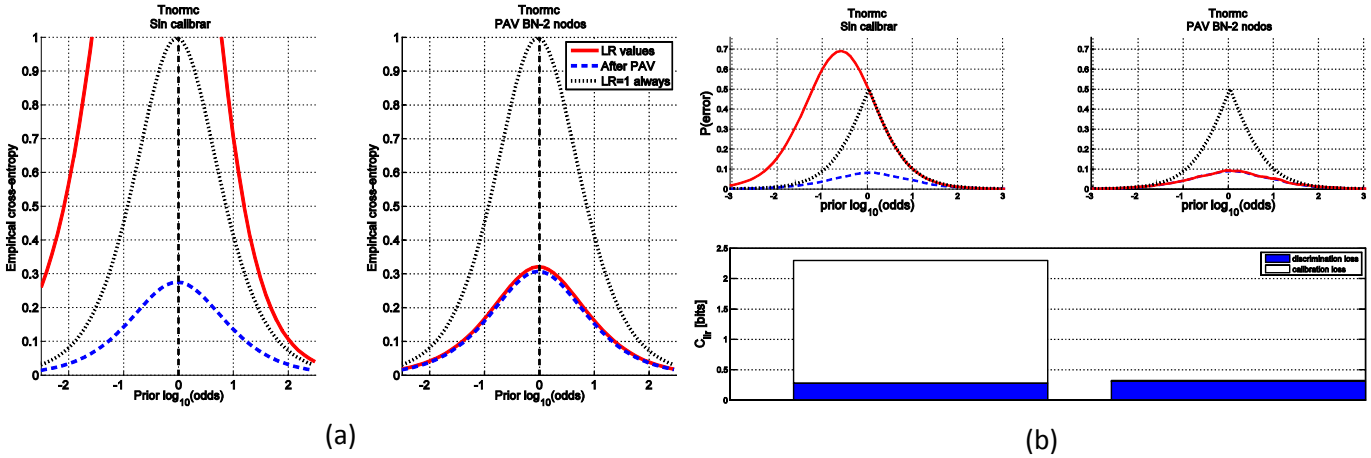
Una vez calculada la probabilidad de cada intervalo para las hipótesis H, se crea un archivo de texto para exportar a Hugin. Aquí, mediante estos datos, se configura la red Bayesiana a partir de las tablas de probabilidad para el nodo padre H y las tablas de probabilidad condicionada para el nodo hijo E. A continuación se muestran las tablas de probabilidad para la técnica de calibración PAV:

H	
H_p	0.5
H_d	0.5

Tabla 7.8: *Tabla de probabilidad del nodo H para PAV.*

E		
H	H_p	H_d
(-∞, int1]	$P_1(E\|H_p)$	$P_1(E\|H_d)$
(int1, int2]	$P_2(E\|H_p)$	$P_2(E\|H_d)$
⋮	⋮	⋮
int 47, ∞,]	$P_{48}(E\|H_p)$	$P_{48}(E\|H_d)$

Tabla 7.9: *Tabla de probabilidad del nodo E condicionado al nodo H para PAV.*



(a) (b)

Figura 7.23: *Curvas ECE (a) y APE (b) comparativas del rendimiento ofrecido por el sistema antes de calibrar y después de calibrar mediante modelado PAV.*

En la figura 7.23, se puede observar que resultados de calibración son muy buenos, ya que en la figura (a) y (b) la curva azul y la roja están muy próximas, y las pérdidas de calibración representadas en la figura (b) como una barra color blanco, no se aprecian después de la calibración.

Comparación de los diferentes métodos de calibración mediante redes Bayesianas

A continuación se realiza una comparación entre los resultados obtenidos a través de las diferentes técnicas de calibración: modelado Gaussiano, GMM y PAV. Para ello nos basaremos en la figuras 7.24 y los resultados mostrados por la tabla 7.10 donde C_{llr}^{min} representa el poder discriminativo y $C_{llr} - C_{llr}^{min}$ representan las pérdidas de calibración de cada sistema.

	CALIBRACIÓN GAUSS		CALIBRACIÓN GMM		CALIBRACIÓN PAV	
SISTEMA	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}
T-NORM Sin calibrar Compensado	0.279	2.296	0.279	2.296	0.279	2.296
T-NORM Compensado Calibrado	0.297	0.338	**0.305**	**0.322**	**0.303**	**0.324**

Tabla 7.10: *Comparación del rendimiento ofrecido por las diferentes técnicas de calibración.*

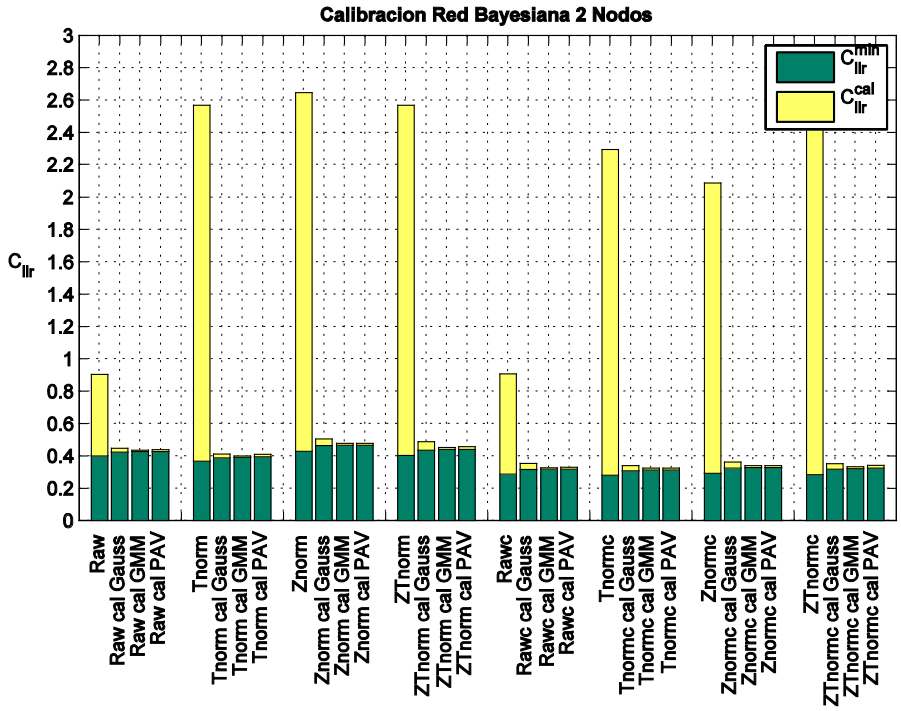


Figura 7.24: *Diagrama de barras representativo del rendimiento del sistema sin combinar (y sin calibrar) y calibrado mediante redes Bayesianas. Se muestran los sistemas sin normalización, RAW, y con normalización T-norm, Z-norm, ZT-norm cada uno de ellos sin compensar y compensado en locutor y canal.*

Tras la realización de los experimentos, se ha podido observar que la calibración de *scores* mediante redes Bayesianas ofrece el resultado esperado. De las 3 técnicas utilizadas, con la que peor resultado se obtiene es con el modelado Gaussiano con unas pérdidas de calibración de 0.041, esto puede ser debido a que la distribución de los datos se modela mediante una función de probabilidad Gaussiana y puede que esta aproximación no sea del todo cierta, es decir, el conjunto de datos no siga una distribución Gaussiana real. Sin embargo, los métodos GMM y PAV obtienen mejores resultados puesto que las pérdidas de calibración son prácticamente nulas, 0.017 y 0.021 respectivamente, debido a que son métodos que se adaptan muy bien al conjunto de datos. También se observa que las pérdidas de discriminación aumentan ligeramente con respecto al sistema original sin calibrar, lo cual no tiene sentido, ya que la transformación aplicada en la calibración no cambia el poder de discriminación de los LR generados con respecto a los *scores*. Esto puede deberse al efecto de la validación cruzada, que genera conjuntos de prueba ligeramente diferentes a los originales.

7.2.2. Implementación de una red Bayesiana de 3 nodos para combinación de evidencias

Una vez evaluados los diferentes métodos de calibración, en esta sección se implementará una combinación "real" de evidencias mediante una red Bayesiana de 3 nodos. De los 3 métodos evaluados, PAV y GMM resultan más difíciles de implementar ya que son métodos complejos sin una solución analítica por lo que habría que recurrir a soluciones aproximadas que requieren algoritmos iterativos y no es el motivo de estudio del presente proyecto. En el caso de la calibración Gaussiana, existe una solución analítica al problema, por lo que será la técnica utilizada de cara a evaluar la posibilidad de la utilización de redes Bayesianas para la combinación de evidencias en reconocimiento de locutor.

La red representativa de este escenario es la siguiente:

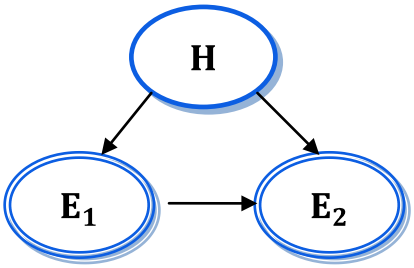


Figura: 7.25: *Red utilizada para combinación de 2 evidencias mediante modelado Gaussiano.*

En este caso, existen 3 nodos, el nodo H que como en los demás casos vistos, representa la hipótesis y los nodos E_1y E_2 que representan cada una de las dos evidencias a combinar. En el capítulo 5, se definió la ecuación del LR para esta red:

$$\frac{P(H_p|E_1,,E_2)}{P(H_d|E_1,E_2)} = \frac{P(E_2|H_p,E_1)}{P(E_2|H_d,E_1)} \cdot \frac{P(H_p|E_1)}{P(H_d|E_1)} = \frac{P(E_2|H_p,E_1)}{P(E_2|H_d,E_1)} \cdot \frac{P(E_1|H_p)}{P(E_1|H_d)} \cdot \frac{P(H_p)}{P(H_d)} \quad (7.10)$$

Teniendo en cuenta la estructura de la red, y la ecuación 7.10 habría que especificar las siguientes probabilidades:

- $P(H_p)$ = $P(H_d)$ = 0.5.
- $p(E_1|H_p)$ y $p(E_1|H_d)$ representadas por sendas funciones densidad de probabilidad Gaussianas, cada una con su media y desviación típica correspondiente:

$$p(E_1|H_p) = \frac{1}{2\pi^{1/2}\sigma_{target}} e^{-\frac{1}{2}\left(\frac{x-\mu_{target}}{\sigma_{target}}\right)} \quad (7.11)$$

$$p(E_1|H_d) = \frac{1}{2\pi^{1/2}\sigma_{non-target}} e^{-\frac{1}{2}\left(\frac{x-\mu_{non-target}}{\sigma_{non-target}}\right)} \quad (7.12)$$

- $p(E_2|E_1,H_p)$ y $p(E_2|E_1,H_d)$ cuyo procedimiento para calcularlo se describe a continuación:

Se puede demostrar que la densidad condicionada de E_2 dado E_1 se define como [63]:

$$p(E_2|E_1,H_i) = \frac{p(E_1,E_2|H_i)}{p(E_1|H_i)} \quad (7.13)$$

donde H_i representa la hipótesis del fiscal o la hipótesis de la defensa y $p(E_1,E_2|H_i)$) representa la densidad conjunta de 2 variables aleatorias normales de media 0 y, a su vez, se puede expresar como [63]:

$$p(E_1, E_2|H_i)) = \frac{1}{2\pi\sigma_{i1}\sigma_{i2}\sqrt{1-r^2}}\, e^{-\left[\frac{1}{2(1-r^2)}\left(\frac{E_1^2}{\sigma_{i1}^2} - \frac{2rE_1E_2}{\sigma_{i1}\sigma_{i2}} + \frac{E_2^2}{\sigma_{i2}^2}\right)\right]} \tag{7.14}$$

donde $\sigma_i = \sigma_{target}$ cuando $H_i = H_p$ y $\sigma_i = \sigma_{non-target}$ cuando $H_i = H_d$.

Por tanto, teniendo en cuenta las ecuaciones 7.13 y 7.14 se tiene:

$$p(E_2|E_1, H_i) = \frac{1}{\sigma_{i2}\sqrt{2\pi(1-r^2)}}\, e^{\left[-\frac{(E_2 - r\sigma_{i2}E_1/\sigma_{i1})^2}{2\sigma_{i2}^2(1-r^2)}\right]} \tag{7.15}$$

Se sigue el mismo razonamiento si E_1y E_2 son normales con media $E\{E_1\} = \mu_1$ y $E\{E_2\} = \mu_2$, entonces $p(E_2|E_1, H_i)$ está dado por la ecuación 7.14 Sustituyendo E_1 por $E_1 - \mu_1$ y E_2 por $E_2 - \mu_2$.

Por tanto, para una evidencia $E_2 = E$, $p(E_2|E_1)$ tiene una densidad de probabilidad normal, con media y varianza:

$$\mu = \mu_{e_2} + r\sigma_2(E_1 - \mu_{e_1})/\sigma_1 \tag{7.16}$$

$$\sigma = \sigma_2^2(1-r^2) \tag{7.17}$$

En Hugin, para una red Bayesiana con nodos discretos y continuos, se asume que cada nodo continuo sigue una distribución Gaussiana. Si un nodo continuo tiene solamente padres discretos, para configurar las tablas de probabilidad basta con especificar la media y la varianza del hijo para cada configuración de los padres.

Si el nodo continuo tiene padres discretos y continuos, entonces para cada configuración discreta se especifica la varianza para el hijo y un conjunto de parámetros que incluyen la media del hijo como una combinación lineal de los padres continuos. Así, para la media, se debe especificar una constante (*intercept*) y un coeficiente (regresión) para cada nodo continuo.

En el caso de estudio, se asume que el nodo E_2 tiene un valor de media representado por una combinación lineal de su nodo padre E_1 condicionado a los estados de H por lo que se tiene:

$$\mu_{E_2} = intercept + coeficiente_{E_1} \cdot E_1 \tag{7.18}$$

A continuación se muestran las tablas de probabilidad de cada uno de los nodos que forman la red:

H	
H_p	0.5
H_d	0.5

Tabla 7.11: *Tabla de probabilidad del nodo H para combinación de 2 evidencias con modelado Gaussiano.*

	E_1	
H	H_p	H_d
Media (μ)	μ_{target}	$\mu_{non-target}$
Varianza (σ^2)	$\sigma^2_{targetE_1}$	$\sigma^2_{non-targetE_1}$

Tabla 7.12: *Tabla de probabilidad del nodo E_1 condicionado al nodo H para combinación de 2 evidencias con modelado Gaussiano.*

	E_2	
H	H_p	H_d
Intercept	$\mu_{targetE_2} - r \cdot \sigma_{targetE_2} \cdot \frac{\mu_{targetE_1}}{\sigma_{targetE_1}}$	$\mu_{non-targetE_2} - r \cdot \sigma_{non-targetE_2} \cdot \frac{\mu_{non-targetE_1}}{\sigma_{non-targetE_1}}$
Coeficiente E_1	$r \cdot \frac{\sigma_{targetE_2}}{\sigma_{targetE_1}}$	$r \cdot \frac{\sigma_{non-targetE_2}}{\sigma_{non-targetE_1}}$
Varianza (σ^2)	$\sigma^2_{targetE_2}$	$\sigma^2_{non-targetE_2}$

Tabla 7.13: *Tabla de probabilidad del nodo E_2 condicionado al nodo H y al nodo E_1 para combinación de 2 evidencias con modelado Gaussiano.*

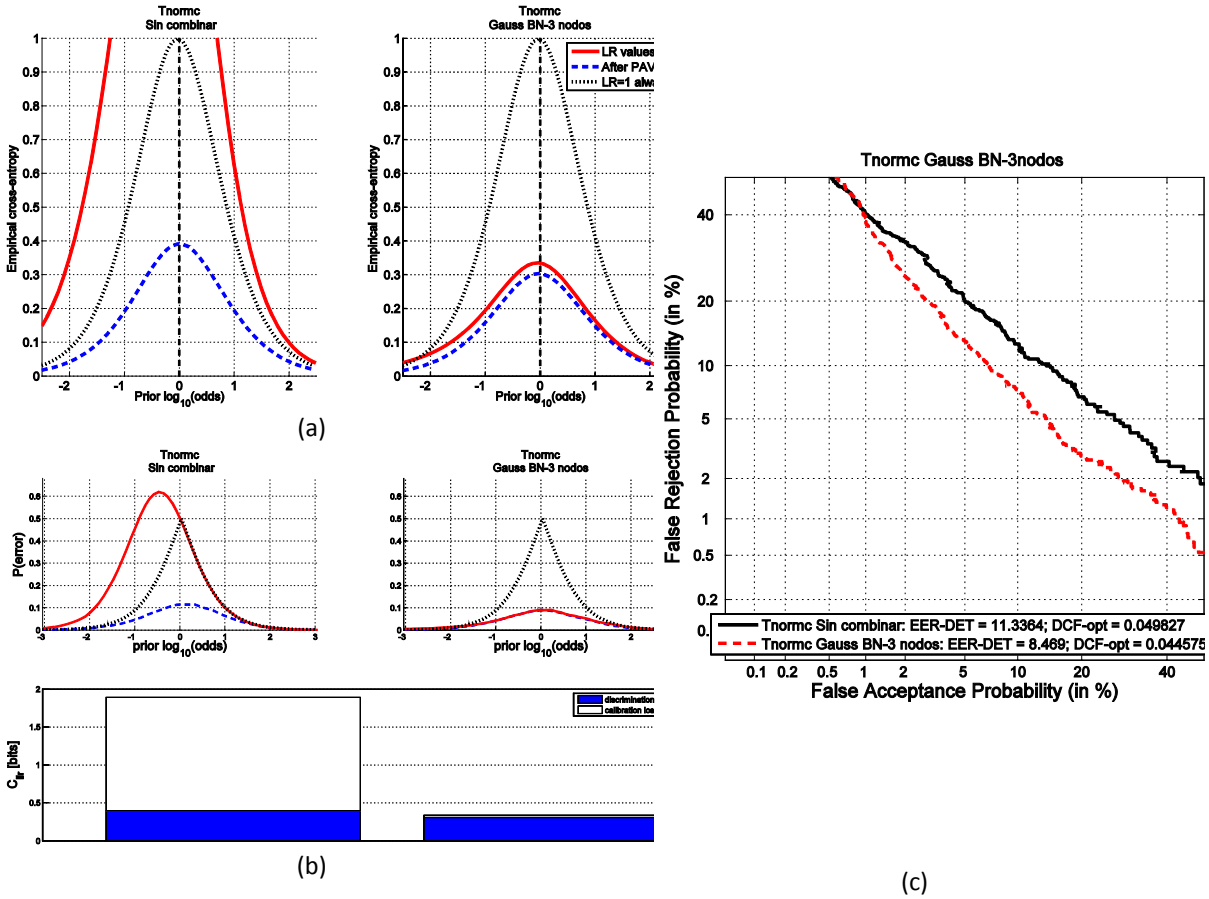


Figura 7.26: *Curvas ECE (a), APE (b) y DET (c) comparativas del rendimiento ofrecido por el sistema antes de combinar y después de aplicar una combinación de evidencias mediante la red Bayesiana de 3 nodos.*

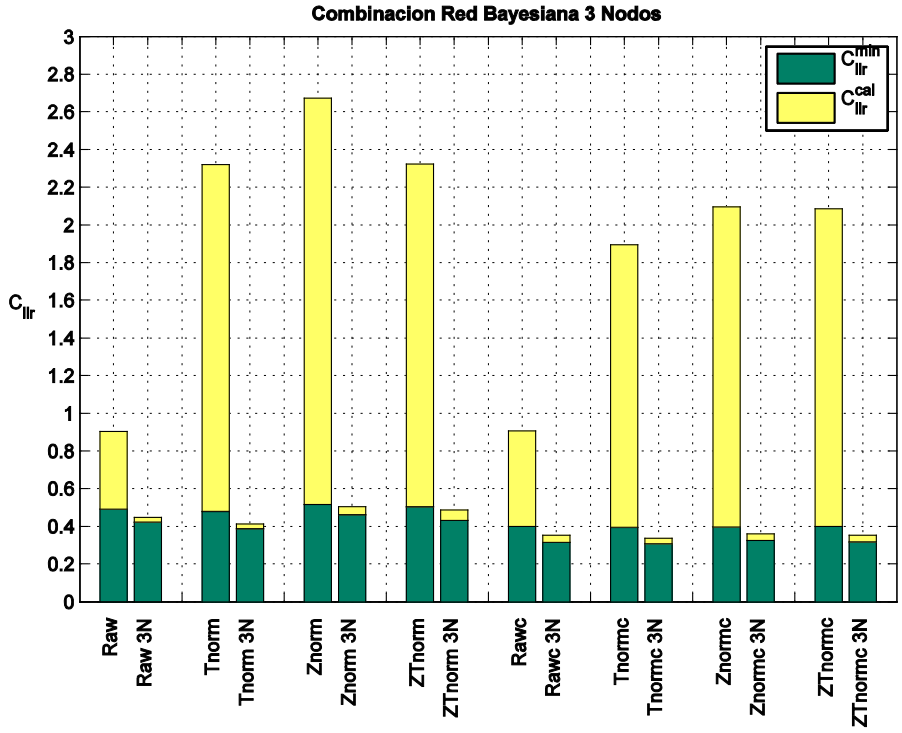


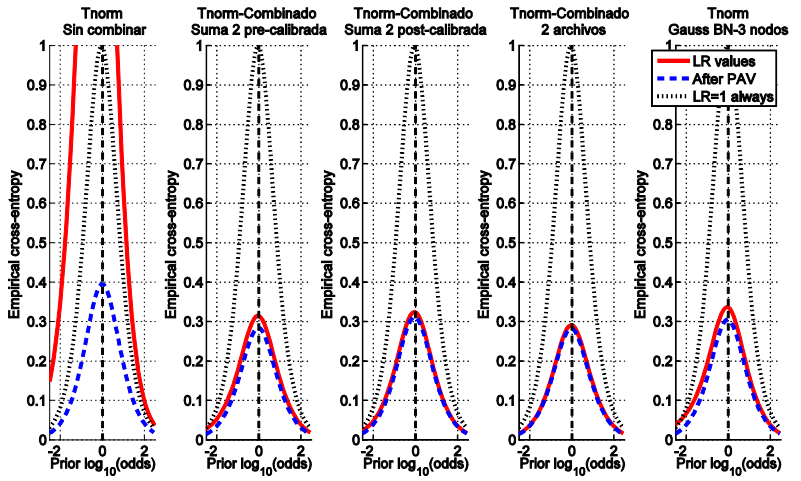
Figura 7.27: *Diagrama de barras representativo del rendimiento del sistema sin combinar (y sin calibrar) y combinado mediante redes Bayesianas. Se muestran los sistemas sin normalización, RAW, y con normalización T-norm, Z-norm, ZT-norm cada uno de ellos sin compensar y compensado en locutor y canal.*

En las figuras 7.26 y 7.27 se observa que mediante la combinación de evidencias con redes Bayesianas se obtienen mejores resultados tanto de discriminación como de calibración, no requiriendo una etapa de calibrado adicional como en el resto de las estrategias propuestas, es decir, mediante la red Bayesiana se realiza la combinación y calibración en un sólo paso por lo que el tiempo de proceso se reduce.

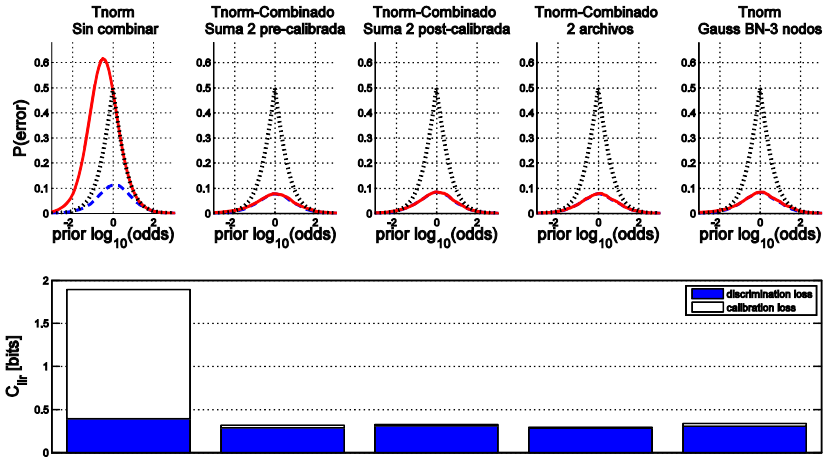
La mejora en discriminación se observa por una curva DET más cerca al origen que el sistema original, equivalente a curvas azules en las gráficas ECE y APE más bajas (figura 7.26) o una gráfica de barras de color verde más baja (figuras 7.27). La mejora en cuanto a calibración se observa de manera muy clara en la figura 7.27 donde la barra amarilla correspondiente a las pérdidas de calibración, desaparece casi por completo.

Comparación de los diferentes métodos de combinación propuestos

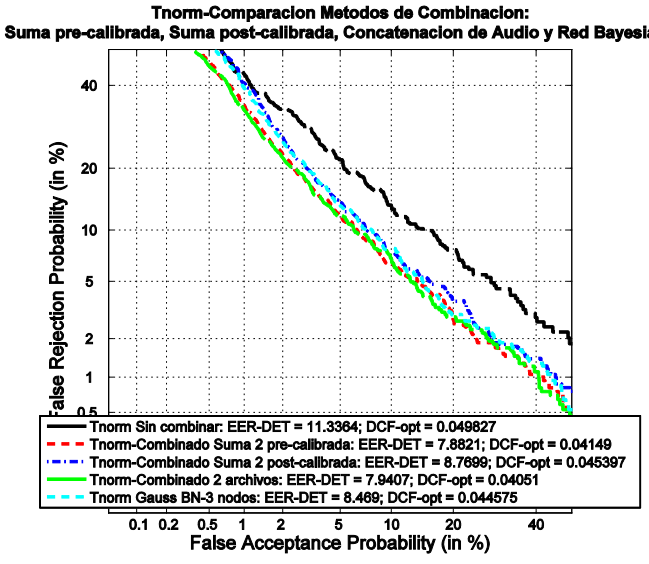
A continuación se realiza una comparación entre los resultados obtenidos a través de las diferentes técnicas de calibración: modelado Gaussiano, GMM y PAV. Para ello nos basaremos en la figuras 7.28 y 7.29 y los resultados mostrados por la tabla 7.14 donde C_{llr}^{min} representa el poder discriminativo y $C_{llr} - C_{llr}^{min}$ representan las pérdidas de calibración de cada sistema.



(a)



(b)



(c)

Figura 7.28: *Curvas ECE (a), APE (b) y DET (c) comparativas del rendimiento ofrecido por el sistema antes de combinar y después de aplicar una combinación de evidencias mediante suma pre y post-calibrada de 2 LRs y concatenación de 2 archivos y combinación con la red Bayesiana de 3 nodos.*

	PRECALIBRADO		POSTCALIBRADO		COMBINACIÓN AUDIO		COMBINACIÓN RED BAYESIANA	
SISTEMA	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}	C_{llr}^{min}	C_{llr}
SIN COMBINAR	0.395	1.893	0.395	1.893	0.395	1.893	0.395	1.893
TNORM COMPENSADO (2)	0.287	0.317	0.313	0.285	0.285	0.293	0.311	0.340

Tabla 7.14: *Comparación del rendimiento ofrecido por las diferentes técnicas de combinación propuestas.*

A partir de los resultados obtenidos se puede observar que utilizando redes Bayesianas el rendimiento del sistema en cuanto a calibración y discriminación son ligeramente peores que en las demás combinaciones (figuras 7.28 y 7.29), sin embargo, esta diferencia no es significativa si te tiene en cuenta que se han conseguido resolver los inconvenientes encontrados en ellas. Esto quiere decir que no se ha perdido prácticamente información durante el proceso de combinación, por lo tanto se puede concluir que las redes Bayesianas son un método válido para combinar evidencias forenses en reconocimiento automático de locutor.

Este resultado es muy importante debido a las grandes ventajas que ofrece el uso de las redes Bayesianas en el ámbito forense. Una de las más relevantes es, que al ser un modelo gráfico, la tarea de organizar y combinar conjuntos de evidencias para resaltar las dependencias entre ellas puede realizarse de modo visual e intuitivo y especificar las ecuaciones relevantes puede hacerse invisible al usuario, es decir, no es necesario conocer la teoría matemática subyacente. Esto hace que sea más fácil de entender por parte del Tribunal y facilitará así el proceso de la toma de decisión. Además, son escalables por lo que cuando se dispone de más conocimiento, la estructura de la red se puede adaptar para alcanzar una nueva comprensión de las propiedades del dominio.

La figura 7.29 muestra un diagrama de barras resumen del rendimiento de todas las técnicas de combinación propuestas en este proyecto. En ella se observa la mejora de rendimiento del sistema cuando se utilizan cualquiera de ellas. Además, se ve que la técnica de combinación mediante redes Bayesianas tiene un rendimiento muy similar pero sin presentar los inconvenientes observados en el resto de técnicas estudiadas.

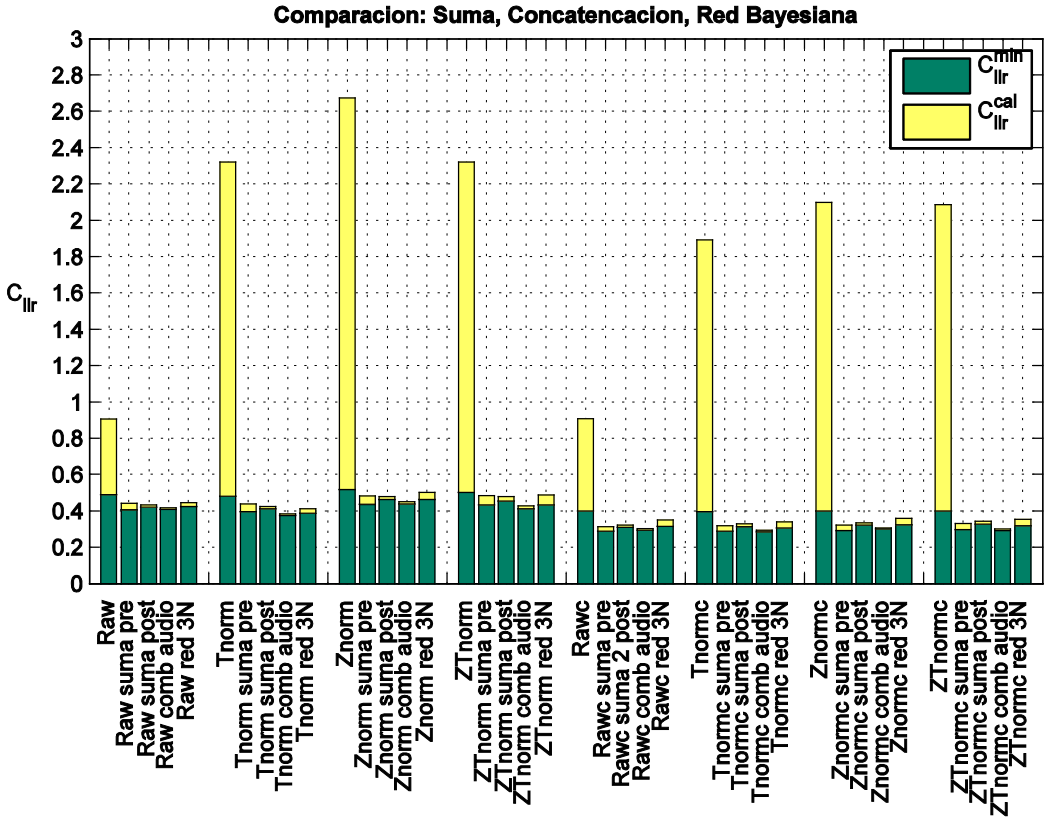


Figura 7.29: *Diagrama de barras representativo del rendimiento del sistema sin combinar (y sin calibrar) y combinado mediante todas las estrategias propuestas: Suma, concatenación y redes Bayesianas. Se muestran los sistemas sin normalización, RAW, y con normalización T-norm, Z-norm, ZT-norm cada uno de ellos sin compensar y compensado en locutor y canal.*

8

Conclusiones y Trabajo futuro

A lo largo de este proyecto se han presentado varias estrategias de combinación de evidencias para sistemas de reconocimiento forense de locutor, basados principalmente en técnicas de combinación a nivel tecnológico y calibración y combinación de evidencias mediante redes Bayesianas. Todas ellas han sido evaluadas sobre la base de datos forense Ahumada III.

En este Capítulo, se procede a analizar de forma global los resultados de cara a extraer conclusiones sobre la mejor técnica de combinación y, en particular, la posibilidad del uso de redes Bayesianas como método de combinación de evidencias.

Adicionalmente se hablará del trabajo futuro, detallando posibles mejoras al trabajo realizado, ampliaciones del mismo, o comentando la posibilidad de continuar evaluando su rendimiento en nuevas bases de datos, más amplias o más realistas.

8.1. Conclusiones

Para solucionar el problema de combinación de evidencias planteado en este proyecto, se han propuesto diferentes técnicas.

En primer lugar se realizó una comparación de diferentes técnicas de combinación de evidencias más inmediatas: suma pre-calibrada y suma post-calibrada (Bayes ingenuo) y concatenación de archivos. Las dos primeras ofrecían mejores resultados en cuanto a discriminación que en el sistema sin combinar, siendo mayor la mejora cuanto mayor era el número de combinaciones. Sin embargo, aparecían pérdidas de calibración que ponían de manifiesto la dependencia entre muestras de audio. Por lo tanto era necesario compensarlas en alguna etapa del proceso para evitar así cometer errores a la hora de interpretar los log-LR de forma probabilística. La última técnica utilizada era la que mejor rendimiento presentaba, sin embargo, tenía el inconveniente que se podrían estar combinando archivos de muy diferentes condiciones haciendo que el modelado de variabilidad y su posterior compensación al utilizar JFA no fuera óptimo.

Con el objetivo de solventar los inconvenientes encontrados en estas técnicas se propuso el uso de redes Bayesianas como método de combinación de evidencias.

Primero se utilizó una red de dos nodos para comprobar la posibilidad del uso de redes Bayesianas como método de calibración, utilizando diferentes métodos de entrenamiento de la red: modelado Gaussiano, GMM y PAV resultando mejores las dos últimas debido a que la

adaptación de los datos era mayor que con el primero, sin embargo, resultan más difíciles de manejar cuando se establece una relación adicional de dependencia con otra evidencia.

Por esto, se implementó una red Bayesiana de 3 nodos, entrenada mediante modelado Gaussiano por ser éste el único en el que existe una solución analítica al problema planteado. Los resultados ofrecidos por este método de combinación, donde se comprueba que no se ha perdido información en el proceso, han permitido concluir que es una técnica adecuada para combinar evidencias de cara a generar el peso de la evidencia de voz final. Este resultado es muy importante debido a las grandes ventajas que ofrece el uso de las redes Bayesianas en el ámbito forense. Una de las más relevantes es, que al ser un modelo gráfico, las dependencias e independencias entre muestras se presentan de un modo visual e intuitivo haciendo que sea más fácil de entender por parte del Tribunal y facilitar así el proceso de la toma de decisión. Además, son escalables por lo que cuando se dispone de más conocimiento, la estructura de la red se puede adaptar para alcanzar una nueva comprensión de las propiedades del dominio.

8.2. Trabajo futuro

Sobre este proyecto es factible realizar una serie de ampliaciones interesantes, sobre todo, aquellas destinadas a seguir aumentando el número de evidencias a combinar, especialmente utilizando redes Bayesianas, que es la estrategia que más ventajas ofrece. También sería interesante utilizar otras técnicas de entrenamiento de la red Bayesiana más complejas como son GMM y PAV. Dichas técnicas, no ofrecen una solución analítica al problema descrito en este proyecto, por lo que se tendrían que utilizar soluciones aproximadas mediante la utilización de algoritmos iterativos.

Por otra parte, también puede ser productivo seguir efectuando pruebas sobre nuevas bases de datos diferentes, de cara a valorar el funcionamiento de los métodos descritos a lo largo de este volumen sobre materiales de diferentes condiciones.

En este proyecto, se han efectuado pruebas sobre la bases de datos Ahumada III. Con ello se ha pretendido valorar el funcionamiento en condiciones forenses reales. No obstante, las muestras fueron obtenidas sobre varones adultos, esto deja fuera de la simulación diferentes perfiles, como niños (perfil quizás menos relevante al tratarse de reconocimiento forense de locutor), y especialmente mujeres de cualquier edad. Sería por tanto de especial interés una prueba de los métodos descritos sobre bases de datos femeninas. También sería interesante en particular una evaluación sobre bases de datos con mayores desajustes: diferentes tipos de ficheros y de diferente duración como por ejemplo la base de datos NIST de altísima variabilidad.

Finalmente, se propone la aplicación de las estrategias de combinación estudiadas en este proyecto, en particular redes Bayesianas, a otros sistemas biométricos como firma, escritura o reconocimiento de cara u otras disciplinas basadas en *scores*, como pueden ser análisis de vidrios, patrones de calzado o balística, los cuales tienen un importante interés en casos forenses.

Referencias

[1] J.P. Campbell: *Speaker verification: A tutorial.* Proceedings of the IEEE, 85: 1437-1462, 1997

[2] D. Meuwly: *Reconaissance de Locuteurs en Sciences Forensiques*: L'apport d'une Approache Automatique. Ph. D. thesis, IPSC-Universite de Lausanne,2001.

[3] J. Gonzalez-Rodriguez, J. Fiérrez-Aguilar, J. Ortega-Garcia y J. J. Lucena-Molina. *Biometric Identification in Forensic Cases According to the Bayesian Approach*. In Proceedings of Biometric Authentication. pp. 177-185, 2002.

[4] F. Taroni, C.G.G. Aitken y P. Garbolino: *De Finetti's subjectivism, the assessment of probabilities and the evaluation of evidence: a commentary for forensic scientists.* Science and Justice, 41, pp. 145-150, 2001.

[5] A. K Jain, A. Ross and S. Prabhakar C. Fredouille: *An introduction to Biometric Recognition*. IEEE Transactions on Circuits and Systems for Video Technology, vol 14 nº 1, January 2004.

[6] R.Clarke: *Human identification for Information Systems: Management Challenges and Public Policy Issues.* Information Technology & People, vol. 7, nº 4, pp. 6-37, 1994.

[7] J.D Woodward: *Biometrics: Privacy's Foe or Privacy'sFriend?* Proc. IEEE Special Issue on Autimated Biometrics, vol.85, nº 9, pp. 1480-1492, 1997.

[8] J. L. Wayman, A. K. Jain, D. Maltoni and D. Maio: *Biometric Systems: Technology,Design and Performance Evaluation.* Springer 2006.

[9] A. K. Jain, P. Flynn y A. Ross: *Handbook of Biometrics*. Springer 2008.

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

[11] A. K. Jain, P. Flynn y A. Ross: *Introduction to Biometrics*. Handbook of Biometrics, pp. 1-22. Springer, 2008.

[12] J. Wayman, A. K. Jain, D. Maltoni, D. Maio. *Biometrics systems. Technology,design and performance evaluation.*Springer, 2005.

[13] Douglas A. Reynolds: *An Overview of Automatic Speaker Recognition Technology*. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002.

[14] A.K. Jain, A Ross y S. Pankanti: *Biometrics: a tool for information security*. IEEE Transactions on Information Forensics and Security, 1, pp. 125 – 143, 2006.

[15] G. Doddington. *Speaker recognition based on idiolectal differences between speakers*. Proc. Eurospeech, pp. 2512-2524, 2001.

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

[17] W.Hess: *Pitch Determination of Speech Signals.* Springer, 1983

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

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

[20] D. A. Reynolds, T. F. Quatieri, y R. B. Dunn. *Speaker verification using adapted gaussian mixture models.* Digital Signal Processing 10, pp. 19-41, 2000.

[21] D. Matrouf y J-F. Bonastre: *Session Effects on Speaker Modeling*. Encyclopedia of Biometrics, pp. 1164-1169, Springer 2008.

[22] J. Ortega-Garcia, S. Cruz-Llanas, J. Gonzalez-Rodriguez: *Quantitative Influence of Speech Variability Factors for Automatic Speaker Verification in Forensic Tasks*. In ICSLP, paper 1062, 1998.

[23] J. Navrátil, G. N. Ramawamy: *The awe and mystery of T-norm*. In Proc. of the European Conference on Speech Communication and Technology, 2003, pp. 2009–2012.

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

[25] R. Cook, I.W. Evett, P.J. Jones, G. Jackson y J.A. Lambert: *A hierarchy of propositions: deciding which level to adress in casework*. Science and Justice, pp 231-239. Vol. 38, 1998

[26] C. Delgado-Romero: *La identificación de locutores en el ámbito forense*. Tesis Doctoral, Universidad Complutense de Madrid, 2001.

[27] F. Taroni y C.G.G. Aitken: *Forensics Science at Trial.* Jurimetrics Journal, 37, pp 327-337, 1997.

[28] B. Robertson y G.A. Vignaux: *Interpreting Evidence. Evaluating Forensic Science in the Courtroom.* Cheicester (Reino Unido): Jonh Wiley & Sons, 1995.

[29] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar y J. Ortega-Garcia: *Forensic Identification Reporting using automatic speaker recognition systems.* International Conference on Information and Communication Technologies: From Theory to Applications, 2008.

[30] E. Itiel Dror, D. Charlton y A. E. Péron: *Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications*. Forensic Science International 156, pp. 74-78, 2006.

[31] M.J. Saks y J.J. Koehler: *The coming Paradifm Shift in Forensic Identification Science.* Science vol. 309, pp 892-895, 2005.

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

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

[34] C. Champod: *The Inference of identity in forensic speaker recognition*. Speech communication, pp. 31, pp. 193-203, 2000.

[35] *U.S. Supreme Court. Daubert vs. Merrel Dow Pharmaceuticals.* Vol. [509 US. 579], 1993.

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

[37] D. Meuwly: *Forensic Evaluation from Biometric Data*. Science & Justice, pp. 205-213, Laussane, 2006.

[38] N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D.A. Van Leeuwen, P. Matejka, P. Scwartz y A. Strasheim: *Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006*. IEEE Transactions on Audio, Speech and Signal Processing, pp. 2072-2084. Vol. 15, 2007.

[39] N. Brümmer y J. Du Preez: *Application independent evaluation of speaker detection*. Computer Speech and Language, 2006.

[40] A. Martin, G. Doddington, T. Kamm, M. Ordowski, y M. Przybocki: *The DET curve in assesment of detection task performance*. Proc. of Eurospeech, pages 1895-1898, 1997.

[41] J. González, J. Fiérrez, D. Ramos, D. García y J. Ortega: *Análisis forense de voces dubitadas en la metodología bayesiana*. II Congreso de la Sociedad Española de Acústica Forense, SEAF, pp. 13-22, Barcelona, España, Abril 2003.

[42] C.E. Shannon: *A mathematical theory of communication*. Bell Sys. Tech. Journal, 1948.

[43] M.I. Jordan: *Learning in Graphicall Models*. MIT Press, Cambridge, MA, 1999.

[44] A.V. Werhli, M. Grzegorczyk y D. Husmeier: *Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks*. Bioinformatics 22, pp. 2523–2531,2006.

[45] J. López, J. García, L. de la Fuente y E.I de la Fuente: *Las redes bayesianas como herramientas de modelado en psicología.* Anales de psicología 23(2), pp. 307-316, 2007.

[46] J. Li, G. Serpen, S.Selman, M. Franchetti, M. Riesen y C. Schneider: *Bayes Net Classifiers for Prediction of Renal Graft Status andSurvival Period.* International Journal of Biological and Biomedical Sciences 2:4, 2006.

[47] A. Walker, B. Pham y M. Moody: *Spatial Bayesian learning algorithms for Geographic information retrieval.* In Proceedings 13th annual ACM international workshop on Geographic information systems, pp. 105-114, Bremen, German, 2005.

[48] L. Ornella y E. Tapia: *Técnicas de Aprendizaje Computacional para Integrar información de Áreas Rurales en un Contexto GIS*. Jornadas de Informática Industrial: Agroinformática - JII 2008. Santa Fe, Argentina, 8-12. ISSN 1850-2849, pp. 168-179. setiembre 2008.

[49] F. Taroni, C. C. G. Aitken, P. Garbolino, A. Biedermann: *Bayesian Networks and Probabilistic Inference in Forensic Science*. Wiley, 2006.

[50] J.H. Wigmore: *The Science of Judicial Proof: As Given by Logic, Psychology, and General Experience and Illustrated in Judicial Trials*, 3rd edn. Little, Brown and Company, Boston, MA, 1937.

[51] B. Robertson y G.A. Vignaux: *Taking fact analysis seriously*. Michigan Law Review 91, pp. 1442–1464, 1995.

[52] D.A. Schum: *Evidential Foundations of Probabilistic Reasoning*. John Wiley & Sons, New York, 1994.

[53] T. Anderson y W. Twining: *Analysis of Evidence: How to do Things with Facts Based on Wigmore's Science of Judicial Proof*. 2nd edn. Northwestern University Press, Evanston, Ill, 1998.

[54] J.L. Cohen: *The Probable and the Provable*. Clarendon Press, Oxford, 1977.

[55] J.L. Cohen: *The difficulty about conjunction in forensic proof*. The Statistician 37, pp. 415-416, 1988.

[56] A.P. Dawid: *The difficulty about conjunction*. The Statistician 36, pp. 91-97, 1987.

[57] H. Katterwe: *True or false*. Information Bulletin for Shoeprint/Toolmark Examiners 9(2), pp. 18-25, 2003.

[58] N. Köller, K. Niessen, M. Riess y E. Sadorf: *Probability Conclusions in Expert Opinions on Handwriting.* Substantiation and Standardization of Probability Statements in Expert Opinions. Luchterhand, München, 2004.

[59] F. Taroni y A. Biedermann: *Inadequacies of posterior probabilities for the assessment of scientific evidence*. Law, Probability and Risk 4, pp. 89-114, 2005.

[60] HUGIN LITE version 7.5, version de evaluación disponible en http://www.hugin.com

[61] NIST. NIST Speech Group website: http://nist.gov/itl/iad/mig/

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

[63] A. Papoulis y S. U. Pillai: Probability, random variables and stochastic processes. Mc Graw Hill 2002, pp. 169-242.

Glosario

APE: Applied Probability of Error

API: Application Programming Interface

ATVS: Área de Tratamiento de Voz y Señales

CMN: Cepstral Mean Normalization

DCT: Discrete Cosine Transform

DET: Detection Error Trade-Off

DTW: Dynamic Time Warping, DTW

ECE: Empirical Cross Entropy

EER: Equal Error Rate

EM: Expectation Maximization

FA: False Accept

FFT: Fast Fourier Transform

FM: Feature Mapping

FR: False Reject

FSR: Forensic Speaker Recognition

FW: Feature Warping

GMM: Gaussian Mixture Models

GSM: Global System for Mobile

HMM: Hidden Markov Model

JFA: Joint Factor Analysis

LPCC: Linear Predictive Cepstral Coding

LR: Likelihood Ratio

MAP: Maximum A Posteriori

MFCC: Mel Frequency Cepstral Coefficients

NIST: National Institute of Standards and Technology

PAV: Pool Adjacent Violators

RASTA: RelAtive SpecTrAl

SRE: Speaker Recognition Evaluation

SVM: Support Vector Machine

TNORM: Test-Normalization

UBM: Universal Background Model

ZNORM: Zero-Normalization

ZTNORM: Zero and Test Normalization

A

Introducción al manejo de redes Bayesianas

A.1. Interfaz gráfica Hugin Expert

A continuación se introducirá la herramienta Hugin Expert, y a través de un tutorial se describirá paso a paso el procedimiento a seguir para crear e instanciar redes Bayesianas mediante su interfaz gráfica.

1. En primer lugar, es necesario diseñar la red para posteriormente construirla.

1.1 Se definen las variables aleatorias para el caso de estudio, es decir, los nodos de la red. En este caso, la red se compone de dos nodos correspondientes al problema de atribución de fuentes:

- El nodo H, que representa la dos posibles hipótesis H_p: El sospechoso es el autor de la actividad criminal y H_d: Otra persona es el autor. En este caso el nodo toma dos posibles valores, por lo que se debe añadir como nodo discreto.
- El nodo E, la evidencia, que en este caso será la puntuación de salida del sistema de reconocimiento de locutor. En este ejemplo el nodo será continuo representado gráficamente con doble borde y la distribución de probabilidad vendrá dada por una distribución Gaussiana. Los dos posibles estados de este nodo son: La toma dubitada proviene del sospechoso (*target*) o la toma dubitada no proviene del sospechoso (*non target*).

1.2 Se especifican las relaciones entre variables, es decir, las flechas. Para este problema el nodo E tiene una dependencia directa con el nodo H

Esta es la ventana principal de Hugin, cuyo panel representa la red actual.

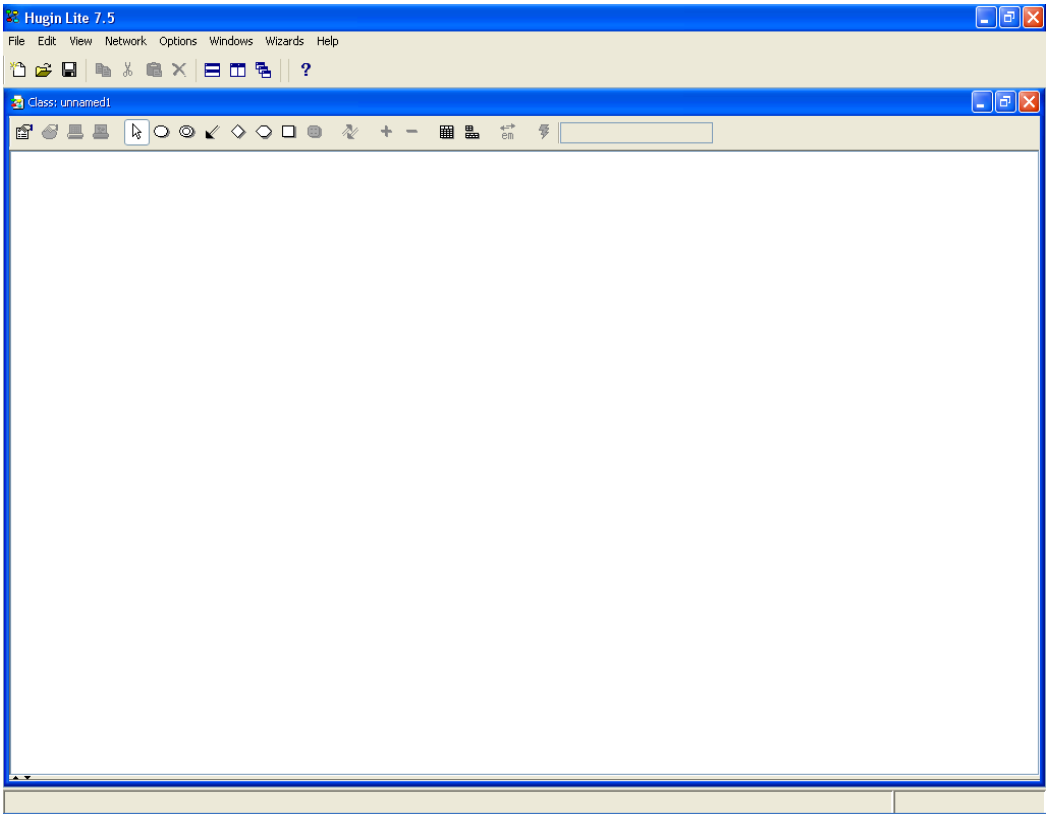


Figura A.1: *Pantalla principal de Hugin Expert.*

Para dibujar la red de ejemplo:

- Se utiliza el botón para añadir un nodo discreto (rojo en la figura A.2), y se dibuja en el panel. Este nodo representará la variable H.
- Se utiliza el botón para crear un nodo continuo (azul en la figura A.2), y se dibuja en el panel. Este nodo representará la variable E.
- Posteriormente, para crear la dependencia entre los nodos H y E, se utiliza el botón representado con una flecha (verde en la figura A.2). Es importante que la flecha tenga la orientación correcta ya que una red Bayesiana es un grafo dirigido. Para ello, se pulsa el botón de unión, se selecciona el nodo H y se arrastra el ratón hasta en nodo E, queda así por tanto definida la dependencia entre ambos nodos.

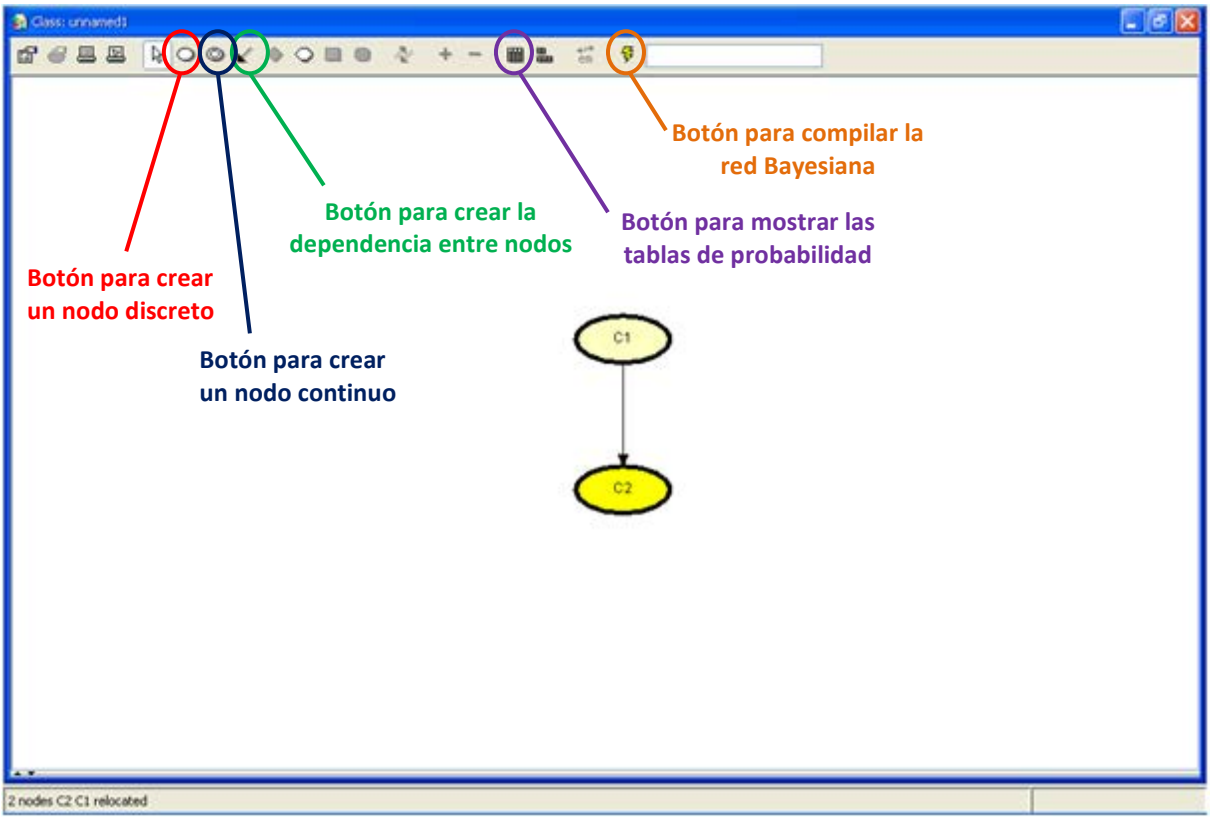


Figura A.2: *Red de ejemplo y descripción de los botones más importantes para la creación y configuración de la red Bayesiana.*

Una vez definida la estructura de la red Bayesiana, se deben configurar las propiedades de los nodos. Para editar estas propiedades, se selecciona el nodo y posteriormente se puede pulsar el botón derecho del ratón, y presionar la opción "*Node Properties*" o presionar CTRL+enter. Aparecerá entonces una ventana como la que representa la figura A.3 (a).

Aquí, se pueden introducir las propiedades del nodo discreto. Es muy útil añadir una etiqueta significativa que represente dicho nodo. En este ejemplo, este nodo representa el nodo H y la etiqueta podría ser: "H – ¿es el sospechoso el autor de la actividad criminal?". Esta etiqueta se mostrará en el panel correspondiente a la red de ejemplo.

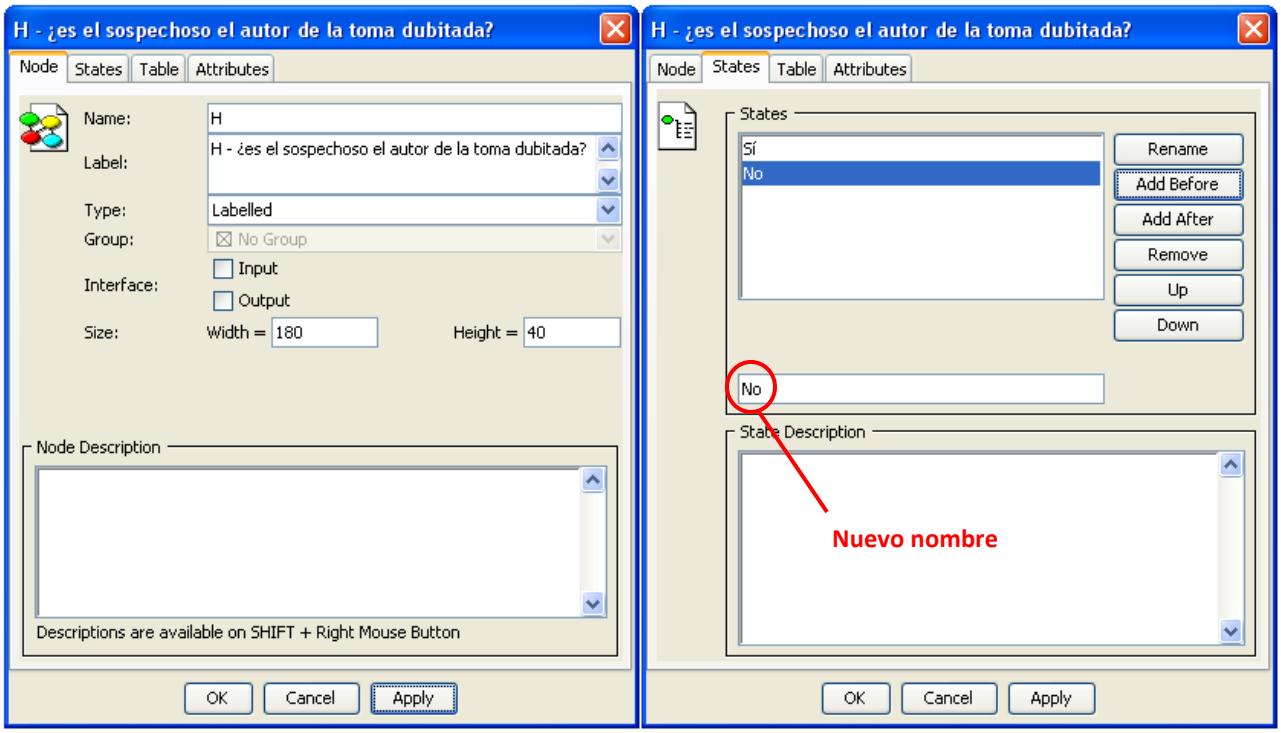


Figura A.3: *Ventanas de configuración de los nodos de la red Bayesiana.*

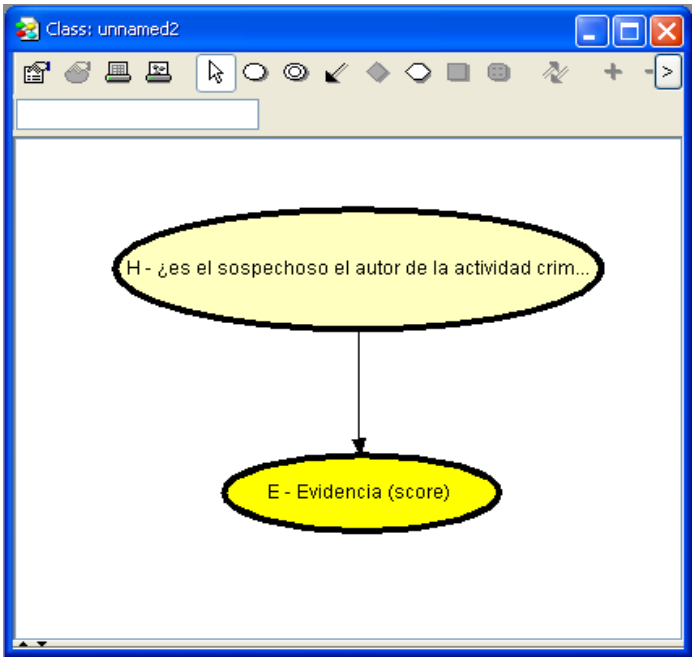


Figura A.4: *Red de ejemplo configurada.*

Se pueden añadir otros atributos desde la pestaña estados (*states*) figura A.3 (b). Por defecto hay dos estados, pero se pueden usar los botones de la derecha para añadir los que sean necesarios. En este caso, sólo se necesitan dos estados, por lo que no hay que añadir ninguno, pero se puede renombrar para darle un nombre más significativo. En el ejemplo, se podrían renombrar como "Sí" y "No" para reflejar su significado. Esto se hace escribiendo el nuevo nombre en el lugar indicado por la figura A.3(b) y presionando cambiar nombre posteriormente.

Para el nodo continuo, E, se procede de la misma manera. En la figura A.4, se muestra la red de ejemplo después de su configuración.

2. En segundo lugar, se introduce la información *a priori.*

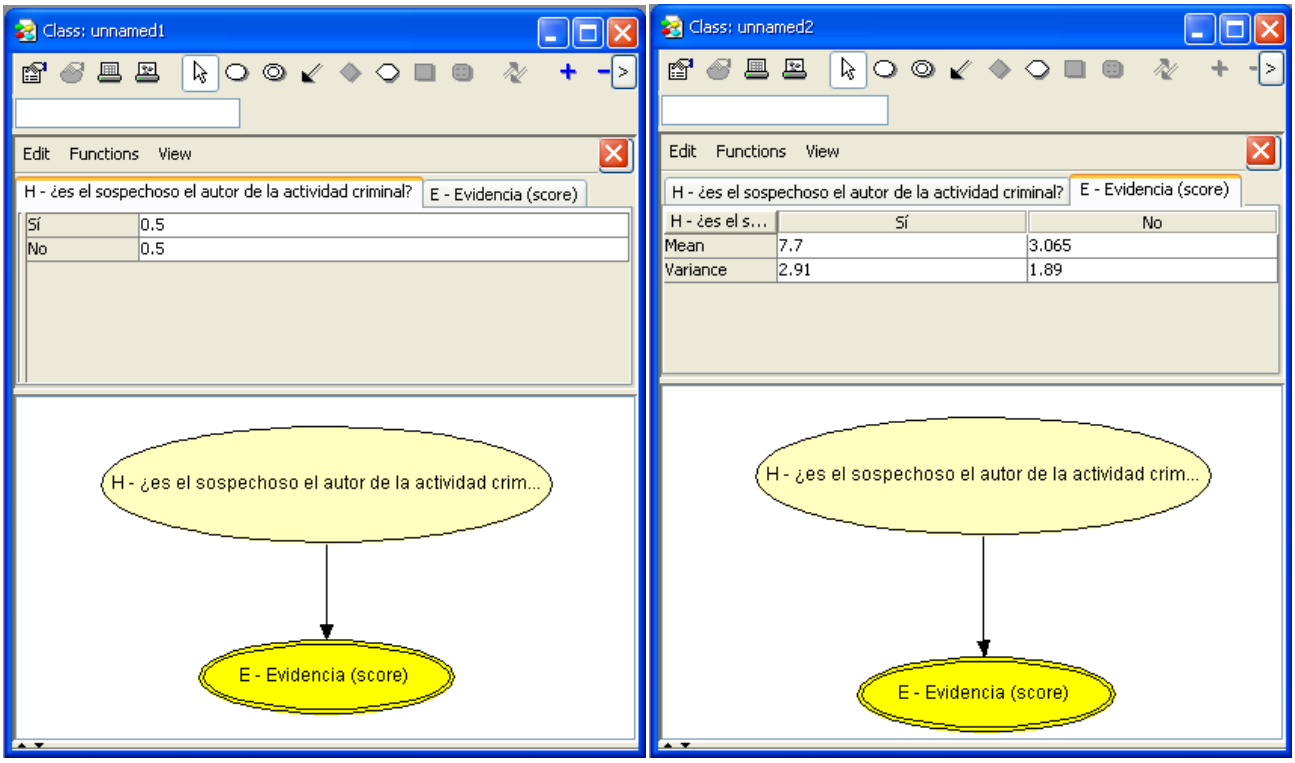
2.1 Se especifican las tablas de probabilidad. Estas tablas son tablas de probabilidad para el nodo padre o tablas de probabilidad condicionada para nodos hijo. Para este caso se tiene que especificar las siguientes probabilidades:

- $P(H_p)$ y $P(H_d)$
- $P(E|H_p)$ y $P(E|H_d)$

Las tablas de probabilidad son lo más importante, y frecuentemente, lo más difícil de una red Bayesiana. Las tablas aparecen en una ventana que puede mostrarse utilizando el botón de la barra de herramientas, (morado en la figura A.2) o pulsando "*View*", y posteriormente "*Show Table Window*". Para que finalmente se muestre la tabla en esta ventana, se debe seleccionar "O*pen Tables*" dentro del mismo menú.

- El nodo H, es un nodo discreto con dos estados, por lo tanto la tabla de probabilidad tendrá 2 valores cuya suma debe ser 1, y para este caso se supone que ambos estados tienen la misma probabilidad de ocurrencia, es decir $P(H_p)$ = $P(H_d)$ = 0.5.
- El nodo E, se ha supuesto que sigue una función densidad de probabilidad normal o Gaussiana para cada posible valor del nodo H, es decir, se supone $P(E|H_p)$ = N(7.7, 2.91) y $P(E|H_d)$ = N(3.065, 1.89).

En la figura A.5, se muestran las tablas de probabilidad para ambos nodos.



(a) *(b)*

Figura A.5: *Tabla de probabilidad del nodo H (a) y nodo E (b).*

Las columnas en las tablas de probabilidad representan los estados del nodo padre y las filas, los estados de los nodos hijo

Una vez configurada la red Bayesiana, se compila pulsando el botón de compilación de la red, (naranja en la figura A.2)

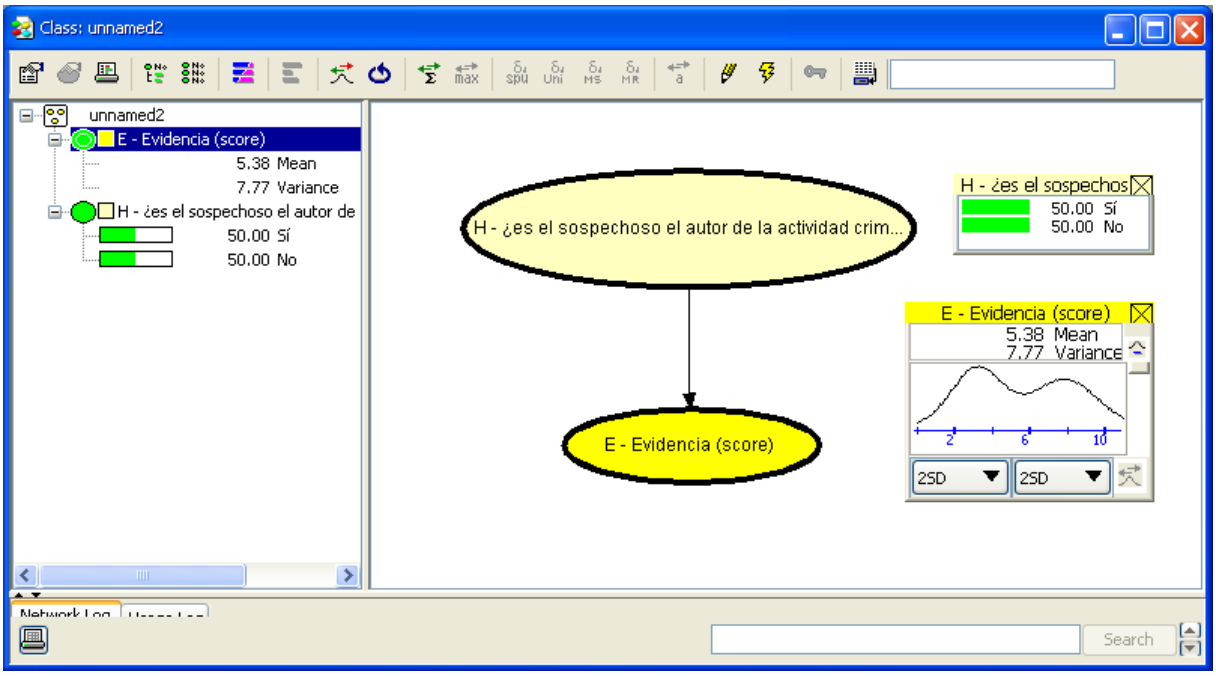


Figura A.6: *Estado de la red antes de introducir la evidencia.*

Hugin utiliza porcentajes para los valores asociados con el estado de cada nodo, en este caso el nodo H, porque es el nodo discreto. La precisión de estos porcentajes se puede cambiar mediante la opción "*Belief Precision*" del menú "*View*".

3. En tercer lugar, se introduce la evidencia, que representa el valor observado de las variables. Como ya se ha especificado anteriormente, en este ejemplo la evidencia será el *score* y el nodo que se instanciará será el nodo E. Para ello, y al ser un nodo continuo, se selecciona el nodo en el panel izquierdo de la ventana (figura A.6), se pulsa el botón derecho del ratón y posteriormente se selecciona "*Insert Continuous Finding*". Se instancia el nodo introduciendo el valor de evidencia observado, en este caso se ha introducido el valor 7, el resultado puede verse en la figura A.7. Para volver al estado inicial del nodo, se procede de la misma manera, pero seleccionando "*Retract Evidence*" en este caso.

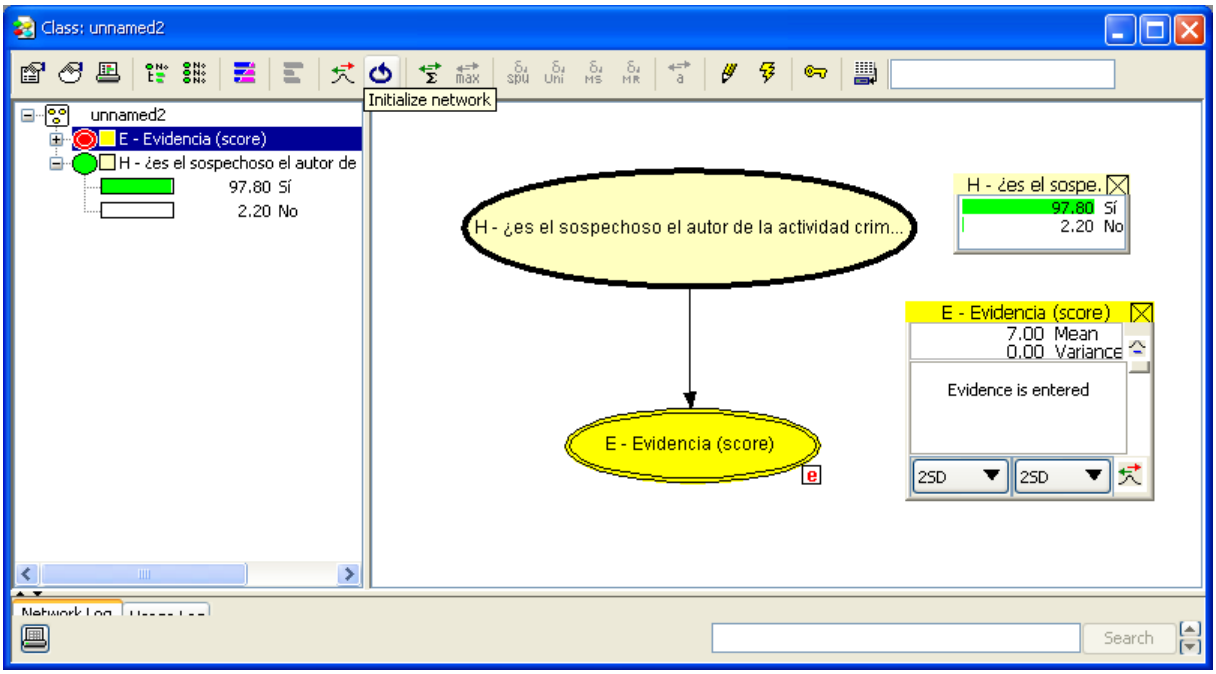


Figura A.7: *Estado de la red Bayesiana tras instanciar el nodo E.*

Para instanciar un nodo discreto, basta con hacer doble click sobre la barra del panel situado a la izquierda de la A.6, una vez instanciado, la barra se vuelve de color rojo. Para volver al estado inicial del nodo, se hace doble click de nuevo, quedando la barra en el color verde inicial. En el ejemplo se instanciará el nodo H para describir la metodología, aunque en este proyecto la instanciación se hará en el nodo E. El resultado de la instanciación puede observarse en la figura A.8.

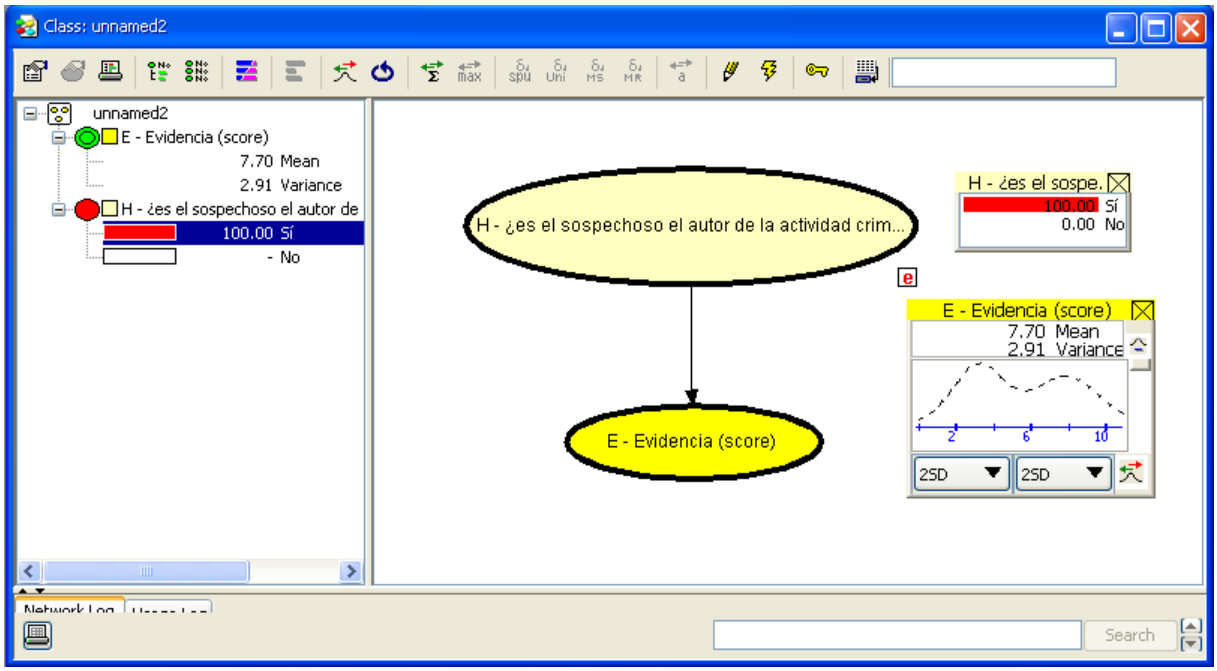


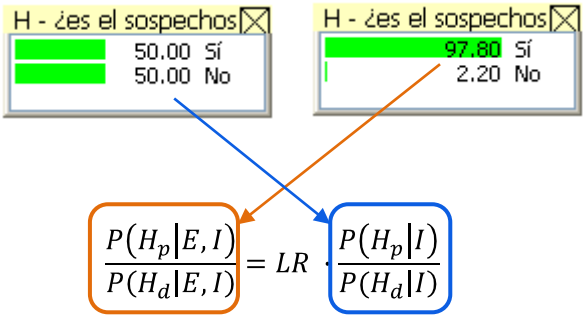
Figura A.8: *Estado de la red Bayesiana tras haber instanciado el nodo H.*

4. Una vez introducida la evidencia, se propaga la red. Hugin realiza la propagación mediante un algoritmo que actualiza las probabilidades *a priori* para las variables no observadas dada la evidencia introducida correspondiente a la variable observada. En este caso, se observa en la figura A.7 que cuando se ha instanciado el nodo E, la probabilidad *a priori* para ambos estados de nodo H, pasa de ser $P(H_p) = P(H_d) = 0.5$, a ser $P(H_p|E) = 0.978$ y $P(H_d|E) = 0.022$.

5. Cálculo de LRs. Como se explicó en el capítulo 4, los LRs se pueden considerar como una medida del peso de la evidencia. La relación entre la información *a priori* y *a posteriori* venía dada por la ecuación 4.3:

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

La información *a priori* y *a posteriori* viene dada por las siguientes tablas antes y después de la instanciación del nodo E respectivamente



Por lo tanto, para calcular el LR basta con dividir las probabilidades *a posteriori*:

$$\frac{0.978}{0.022} = LR \cdot \frac{0.5}{0.5} \approx 45$$

A.2. API C++ Hugin Expert

A continuación se presenta el código fuente generado en C++ para la implementación del mismo ejemplo descrito en la sección A.1mediante la interfaz gráfica de Hugin.

```
# include "hugin"

# include <vector>
# include <iostream>
# include <cmath>
# include <fstream>

using namespace HAPI;
using namespace std;

class BAP {
public:
  BAP ();
protected:

    ContinuousChanceNode* BAP::constructCC (const char *label, const char *name);
    LabelledDCNode* BAP::constructLDC (const char *label, const char *name, size_t n);

    void printBeliefsAndUtilities (Domain*);

    void BAP::buildStructure (IntervalDCNode *H, LabelledDCNode *E);

    void BAP::specifyDistributions (IntervalDCNode *H, LabelledDCNode *E);

    void BAP::buildNetwork ();

    void BAP::insertFinding (double finding);

    Domain *domain;

    ~BAP () { delete domain; }
};

    /** Build a Bayesian network and print node beliefs. */
    BAP::BAP ()
    {
    domain = new Domain ();

    buildNetwork ();
    domain->compile ();


    printBeliefsAndUtilities (domain);

    double finding = 7;
    insertFinding (finding);

    printBeliefsAndUtilities (domain);

    domain->saveAsNet ("RedPrueba.net");
}

void printBeliefsAndUtilities (Domain *domain)
{
    NodeList nodes = domain->getNodes();
    bool hasUtilities = containsUtilities (nodes);
```

```
    for (NodeList::const_iterator it = nodes.begin(); it != nodes.end(); ++it)
    {
        Node *node = *it;

        Category category = node->getCategory();
        char type = (category == H_CATEGORY_CHANCE ? 'C'
                     :category == H_CATEGORY_DECISION ? 'D'
                     :category == H_CATEGORY_UTILITY ? 'U' : 'F');
        cout << "\n[" << type << "] " << node->getLabel()
             << " (" << node->getName() << ")\n";

        if (category == H_CATEGORY_UTILITY)
        {
            UtilityNode *uNode = dynamic_cast<UtilityNode*> (node);
            cout << "  - Expected utility: " << uNode->getExpectedUtility()
             << endl;
        }
        else if (category == H_CATEGORY_FUNCTION)
        {
            try
            {
                FunctionNode *fNode = dynamic_cast<FunctionNode*> (node);
                double value = fNode->getValue ();
                cout << "  - Value: " << value << endl;
            }
            catch (const ExceptionHugin& e)
            {
                cout << "  - Value: N/A\n";
            }
        }
        else if (node->getKind() == H_KIND_DISCRETE)
        {
            DiscreteNode *dNode = dynamic_cast<DiscreteNode*> (node);
            Table* table = dNode->getTable();
            NodeList nodes = table->getNodes();
            vector<size_t> configuration(nodes.size());
            NumberList data = table->getData();

            for (int index = 0; index < table->getSize(); index++)
            {
                table->getConfiguration(configuration, index);
                cout << "P(" << dNode->getName() << " = "  << dNode-
                >getStateLabel(configuration[configuration.size()-1]);
                if (nodes.size() > 1)
                {
                    cout << " |";

                    for (int n = 0; n < configuration.size()-1; n++)
                     {
                         cout << " " << nodes[n]->getName() << "=" <<
                        dynamic_cast<DiscreteNode*>(nodes[n])->getStateLabel(configuration[n]);
                    }
                }
                cout << ") = " << data[index] << endl;
            }
        printNodeMarginals (domain);
        }

        else
        {
            ContinuousChanceNode *ccNode = dynamic_cast<ContinuousChanceNode*> (node);
            cout << "  - Mean : " << ccNode->getMean() << endl;
            cout << "  - SD   : " << sqrt (ccNode->getVariance()) << endl;
        }
    }
}
```

```
/** Are there utility nodes in the list? */
bool containsUtilities (const NodeList& list)
{
    for (size_t i = 0, n = list.size(); i < n; i++)
        if (list[i]->getCategory() == H_CATEGORY_UTILITY)
            return true;

    return false;
}

LabelledDCNode * BAP::constructLDC (const char *label, const char *name, size_t n)
{
    LabelledDCNode *node = new LabelledDCNode (domain);

    node->setNumberOfStates (n);
    node->setLabel (label);
    node->setName (name);

    return node;
}

ContinuousChanceNode * BAP::constructCC (const char *label, const char *name)
{
    ContinuousChanceNode *node = new  ContinuousChanceNode (domain);
    node->setLabel (label);
    node->setName (name);
    return node;
}

/** Build the structure.  */
void BAP::buildStructure (IntervalDCNode *H,  ContinuousChanceNode *E)
{
    E->addParent (H);

    H->setPosition (100, 200);
    E->setPosition (200, 200);
}

/** Specify the prior distribution for H and E nodes. */
void BAP::specifyDistributions (IntervalDCNode *H,  ContinuousChanceNode *E)
{
    /** NODE H */

    Table *table = H->getTable ();
    NumberList data = table->getData ();



    data[0] = 0.5; // A priori probabilities
    data[1] = 0.5;

    table->setData (data);

    /** NODE E */

    CGDistribution *distr = E->getCGDistribution();


    distr->setAlpha(0,7.7); //Mean
    distr->setGamma(0,2.91); //Variance

    distr->setAlpha(1,3.065);
    distr->setGamma(1,1.89);
}
```

```
/** Build the Bayesian network. */
void BAP::buildNetwork ()
{
    domain->setNodeSize (80,40);

    LabelledDCNode *H = constructLDC ("es el sospechosos el autor de la actividad criminal?", "H", 2);
    ContinuousChanceNode *E = constructCC ("Evidencia", "E");

    buildStructure (H,E);

    specifyDistributions (H,E);

}

/** Insert evidence */
void BAP::insertFinding (double finding)
{
    ContinuousChanceNode  *E = dynamic_cast<ContinuousChanceNode*> (domain-
    >getNodeByName("E"));
    DiscreteChanceNode  *H = dynamic_cast<DiscreteChanceNode*> (domain->getNodeByName("H"));


    E->enterValue(finding);
    domain->propagate(H_EQUILIBRIUM_SUM, H_MODE_NORMAL);


    //      E->retractValue ();
    //      domain->initialize();
}

/**
 * Build a Bayesian network, and compute and print the initial node marginals.
 */
int main (int argc, char *argv[])
{
    new BAP ();
    return 0;
}
```

B

Gráficas y Tablas

Este anexo se encuentra en la versión completa de la memoria disponible en
http://atvs.ii.uam.es/listpublications.do

C

Presupuesto

Ejecución Material

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

Gastos generales

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

Beneficio Industrial

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

Honorarios Proyecto

- 1800 horas a 15 € / hora ..... 27000 €

Material fungible

- Gastos de impresión..... 280 €
- Encuadernación..... 200 €

Subtotal del presupuesto

- Subtotal Presupuesto..... 32796 €

I.V.A. aplicable

- 18% Subtotal Presupuesto ..... 5903,28 €

Total presupuesto

- Total Presupuesto ..... 38699,28 €

Madrid, Diciembre de 2011

El Ingeniero Jefe de Proyecto

Fdo.: Eva Barriel Guitián
Ingeniero Superior de Telecomunicación

D

Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización de este proyecto: *Cálculo del peso de la evidencia en casos forenses de reconocimiento automático de locutor en los que existen varias tomas de voz de procedencia desconocida*. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

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

Condiciones generales

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

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

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

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

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

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

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

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.
9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

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

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

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

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

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

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

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

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

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

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

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

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

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

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

Condiciones particulares

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

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

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

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

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

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

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

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

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

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

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

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

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