UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

PROYECTO FIN DE CARRERA

MEJORAS EN EL MODELADO ACÚSTICO PARA RECONOCIMIENTO DE LOCUTOR DEPENDIENTE DE TEXTO

Daniel Hernández López

Septiembre 2008

MEJORAS EN EL MODELADO ACÚSTICO PARA RECONOCIMIENTO DE LOCUTOR DEPENDIENTE DE TEXTO

AUTOR: Daniel Hernández López
TUTOR: Doroteo Torre Toledano

ATVS Biometric Recognition Group
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Septiembre de 2008

PROYECTO FIN DE CARRERA

Título: *Mejoras en modelado acústico para reconocimiento del locutor dependiente de texto*

Autor: D. Daniel Hernández López

Tutor: D. Doroteo Torre Toledano

Tribunal:

Presidente: Joaquín González Rodríguez

Vocal: José Colás Pasamonte

Vocal secretario: Doroteo Torre Toledano

Fecha de lectura:

Calificación:

Resumen:

En este proyecto de fin de carrera se realizan un conjunto de mejoras a un sistema de reconocimiento de locutor dependiente de texto orientado a aplicaciones realistas, es decir con poca cantidad de datos de entrenamiento, siendo estos monosesión. Se emplean técnicas de adaptación al locutor MLLR (Maximum Likelihood Linear Regression) y MAP (Maximum A Posteriori) para Modelos Ocultos de Markov o HMM (Hidden Markov Models) y técnicas de normalización de puntuaciones como T-norm aplicado a nivel de frase fonema y estado.

En la memoria se realiza un repaso a las disciplinas más empleadas en reconocimiento biométrico, se explican los principios de producción de voz y de audición humana y se realiza un amplio repaso al estado del arte en reconocimiento de locutor dependiente de texto. Posteriormente se explican los procedimientos llevados a cabo y se exponen los resultados obtenidos mediante dichos procedimientos. Por último se realiza un análisis del estudio extrayendo conclusiones y proponiendo futuras líneas de investigación.

Los resultados obtenidos en este proyecto de fin de carrera han sido publicados en 2 congresos internacionales en los artículos:

- BioSec Multimodal Biometric Database and its use in Text-Dependent Speaker Recognition Research, publicado en LREC 2008.
- MAP and Sub-Word Level T-Norm for Text-Dependent Speaker Recognition, publicado en Interspeech 2008.

Además un tercer artículo ha sido enviado a congreso y se encuentra a la espera de ser aceptado:

- T-Norm y desajuste Léxico y Acústico en Reconocimiento de Locutor Dependiente de Texto, enviado a las Jornadas de las Tecnologías del Habla 08.

Palabras clave:

Reconocimiento de locutor dependiente de texto, Adaptación MLLR, Adaptación MAP, Alineamiento no completamente forzado, T-norm, T-norm a nivel de fonema, T-norm a nivel de estado.

Abstract:

In this master´s thesis a set of improvements are developed for a realistic-oriented text-dependent speaker recognition system, it means the experiments have been realised with few single-session training data. MLLR (Maximum Likelihood Linear Regression) adaptation and MAP (Maximum A Posteriori) adaptation have been performed to train HMM (Hidden Markov Models) and score normalization techniques like utterance, phoneme and state level T-norm have been performed.

In the report a summary of the most used techniques in biometric recognition has been written, explaining the voice production and human hearing principles and making a deep study of the state o the art in text-dependent speaker recognition. Later the processes and its final scores are explained. Finally, an analysis of the project has been done extracting conclusions and proposing new research lines.

Two conference papers have been published in international congresses as a result of this master´s thesis:

- BioSec Multimodal Biometric Database and its use in Text-Dependent Speaker Recognition Research, publiseh in LREC 2008.
- MAP and Sub-Word Level T-Norm for Text-Dependent Speaker Recognition, published in Interspeech 2008.

A third conference paper has been sent, now it is waiting for acceptance:

- T-Norm y desajuste Léxico y Acústico en Reconocimiento de Locutor Dependiente de Texto, sent to Jornadas de las Tecnologías del Habla 08.

Key Words:

Text-dependent speaker recognition, MLLR adaptation, MAP adaptation, Not-completely forced alignment, T-norm, Phoneme level T-norm, State level T-norm.

Agradecimientos

Quiero dar las gracias primero a toda la gente que me ha ayudado a llegar hasta aquí, hasta el final de los estudios de una Ingeniería. Por ello quiero agradecer a mis padres la actitud que han tenido conmigo desde que empecé la carrera y a mi familia y amigos el apoyo y la comprensión que han mostrado desde el primer día. Sin vosotros no habría llegado hasta aquí. No puedo olvidarme de los compañeros buenos que he tenido a lo largo de estos estudios que me han ayudado y aconsejado sin esperar nada a cambio.

Me siento muy agradecido a mi tutor Doroteo Torre por el buen trato que he recibido de él y por haberme guiado y apoyado durante todo el proyecto. También estoy especialmente agradecido a Joaquín González por haberme dado la oportunidad de realizar mi proyecto de fin de carrera en el ATVS y a Javier Ortega por permitirme entrar en el grupo y conocerlo. El resto de miembros del ATVS me han ofrecido desde que entré a formar parte del grupo su ayuda, consejo y amistad y por ello aunque no les nombre tienen mi más sincera gratitud.

Por último quería dar las gracias a dos miembros del personal de esta escuela que me han aconsejado y ayudado en todo momento, pero que sobretodo siempre han estado dispuestos a escucharme, Maína Aznar y Jesús Bescós.

Este proyecto ha sido realizado en el Área de Tratamiento de Voz y Señales (ATVS) en la Escuela Politécnica superior de la Universidad Autónoma de Madrid. El proyecto ha sido financiado parcialmente por el Ministerio de Ciencia e Innovación a través del proyecto TEC2006-13170-C02-01.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

En los últimos años se dice que hemos entrado en la era las Telecomunicaciones, términos como sociedad de la información o las conocidas TIC (Tecnologías de la Información y Comunicaciones) están a la orden del día. Y es que en muy pocos años este tipo de tecnologías han evolucionado muy rápidamente. Hace unos años nadie pensaría que todos irían por la calle con teléfonos móviles y que podríamos conectarnos a internet desde prácticamente cualquier parte del mundo. La revolución que ha supuesto el hecho de que la gran mayoría de la población occidental tenga acceso a internet ha afectado a la práctica totalidad de los quehaceres cotidianos. Hoy en día se puede hacer la compra al supermercado sin movernos de casa, podemos realizar cualquier tipo de compra, compartir contenidos, enviar correos de forma instantánea o mantener una videoconferencia con las antípodas en tiempo real.

La libertad de movimientos que todo esto conlleva, arrastra inevitablemente consigo nuevas formas de estafa y delincuencia. Es por ello por lo que han proliferado los sistemas de seguridad basados en reconocimiento biométrico. Los sistemas basados en clave eran relativamente fáciles de asaltar con las técnicas de desencriptado existentes. Pero un sistema de seguridad basado en un parámetro biométrico es mucho más difícil de asaltar con éxito. Por ello poco a poco se han ido imponiendo progresivamente este tipo de sistemas para garantizar la seguridad en transacciones económicas, intercambios de información o bien para tener acceso a información reservada o bienes preciados.

Entre las muchas aplicaciones en las que tiene cabida el reconocimiento biométrico, tal vez es el reconocimiento de locutor y de habla el más versátil. Esto es porque puede ser empleado tanto para reconocimiento biométrico como para reconocimiento de instrucciones.

Entre las muchas aplicaciones que tiene el reconocimiento de voz podríamos destacar la banca telefónica, donde un impostor con los datos de otra persona podría tener control sobre sus cuentas, con una aplicación de reconocimiento de locutor el banco garantiza un nivel muy superior de seguridad, ya que el impostor no podrá realizar ningún tipo de transacción sin la voz del dueño de las cuentas. Otra aplicación son los servicios de conversación guiada, hoy en día en determinados sitios se pueden comprar las entradas para un evento deportivo o cultural haciendo una llamada, sin necesidad de que exista un operador humano en el otro extremo de la comunicación. En entornos domóticos o para discapacitados, se puede conseguir dar instrucciones a un sistema automático o a un software para que realice una determinada función con sólo pronunciar la instrucción. Y por supuesto en entornos de seguridad o control de acceso la voz es un sistema muy cómodo y natural de identificación.

En este proyecto se profundiza en el reconocimiento de locutor dependiente de texto. En un sistema de estas características el individuo para identificarse deberá aportar una clave mediante la voz, con lo que no sólo deberá conocer la clave, sino que también deberá tener la voz de la persona como la que pretende identificarse.

1.2 Objetivos

El principal objetivo de este proyecto es mejorar un sistema de reconocimiento de locutor dependiente de texto basado en Modelos Ocultos de Markov. El proyecto estudia la forma de mejorar un sistema de estas características en dos aspectos fundamentales:

- El modelado acústico y la adaptación al locutor de los modelos independientes del locutor.
- El tratamiento de los resultados para obtener una lectura de los mismos que hagan el sistema más eficiente.

El proyecto se centra en mejorar el sistema para aplicaciones realistas, en las que los datos disponibles para el entrenamiento de un modelo de locutor sean limitados y monosesión. Esto es debido a que en una aplicación real, con la que el usuario se sienta cómodo y con la que esté dispuesto a cooperar, no podemos pedir al usuario que repita muchas frases, ni que realice varias sesiones, debe ser un proceso corto y cómodo.

Esta limitación debida a las necesidades de la aplicación es importante, ya que cuanto menor es la cantidad de datos para entrenar más difícil resulta la tarea de reconocimiento. Además si las locuciones que tenemos son de una única sesión, nuestros modelos no recogerán la variabilidad que se produce en la voz debida al uso de distintos canales (teléfono, grabación microfónica, etc) o a distintos estados del locutor (resfriados, distintos estados de ánimo, etc).

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos:

1. Introducción: motivación y objetivos del proyecto.
2. Sistemas de reconocimiento biométrico: repasa la estructura y los diferentes tipos de sistemas de reconocimiento biométrico.
3. Sistemas de reconocimiento de locutor: realiza una introducción a la naturaleza de la señal de voz y del oído humano y explica los diferentes tipos de sistemas de reconocimiento de locutor.
4. Estado del arte en reconocimiento de locutor dependiente de texto: realiza un repaso a las principales técnicas usadas en esta disciplina, así como los retos y soluciones planteados.
5. HMMs: repasa los conceptos matemáticos en los que se basa el proyecto
6. Bases de datos utilizadas: describe las bases de datos que se han empleado para realizar los experimentos.
7. Sistema desarrollado: explica el sistema de partida y las mejoras realizadas al mismo
8. Experimentos realizados: muestra los resultados obtenidos con cada mejora realizada.
9. Conclusiones y trabajo futuro: Evalúa los resultados obtenidos y propone nuevas líneas de investigación y mejora.

2 Sistemas de reconocimiento biométrico

2.1 Introducción al reconocimiento biométrico

Un sistema de reconocimiento biométrico es un sistema automatizado de reconocimiento humano basado en las características físicas y comportamiento de las personas. Es un sistema que reconoce a la persona basado en quién es la persona, no importando lo que la persona está llevando o lo que la persona conoce. Cosas que una persona puede llevar, así como llaves y tarjetas de identificación, pueden ser perdidas, sustraídas o duplicadas. Cosas que una persona conoce, tales como palabras clave (passwords) y códigos, pueden ser olvidados, sustraídos o duplicados. El lugar de ello un sistema de reconocimiento biométrico se fija en quién es la persona, basándose en una característica humana que no puede ser perdida, olvidada, sustraída o duplicada.

2.1.1 Estructura de un sistema de reconocimiento biométrico

En un sistema de reconocimiento biométrico intervienen una serie de etapas desde que la persona susceptible de reconocimiento se expone al sistema hasta que el sistema le reconoce, acepta o rechaza.

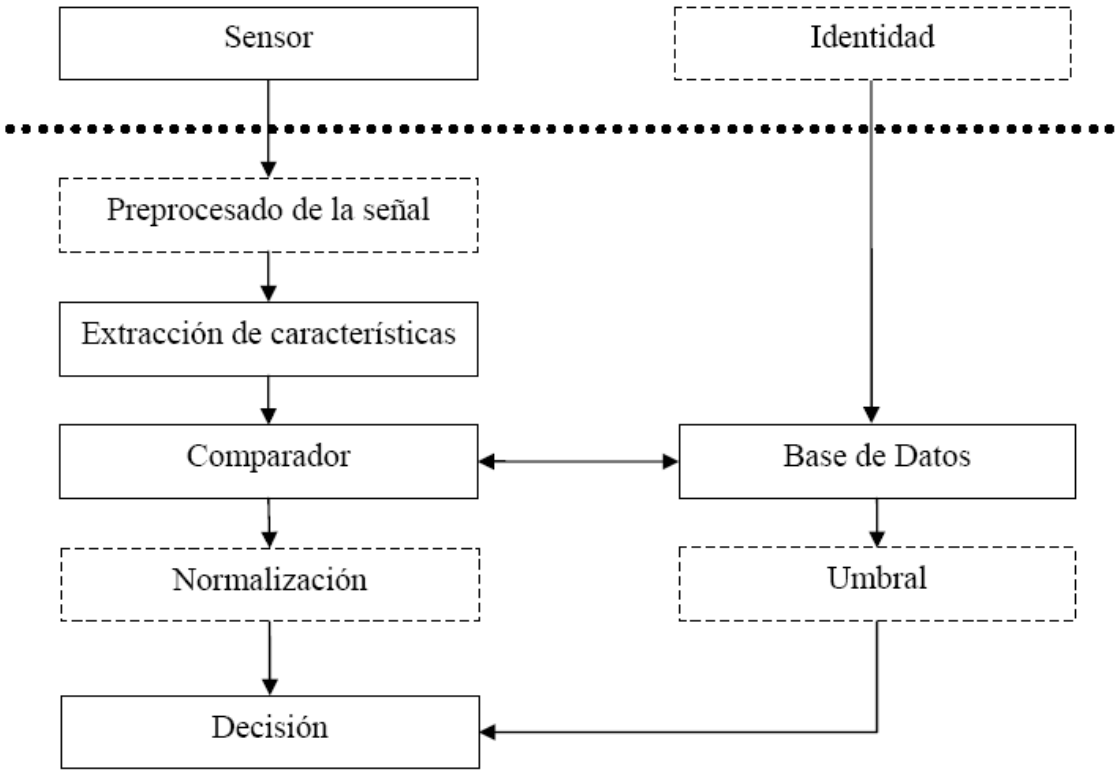


Figura 1: Arquitectura de un sistema de reconocimiento biométrico

La primera etapa es la transducción, es decir convertir la señal física en una señal eléctrica. En el caso de la voz la transducción la realiza un micrófono, transformando la señal acústica en señal eléctrica. Para el caso de la huella dactilar es un sensor lo que recibe la

yema del dedo y proporciona una imagen de las crestas y valles de la huella. Para otros rasgos como la cara o el iris es algún tipo de cámara lo que obtiene una imagen del rasgo bajo estudio.

Una vez se tiene una señal eléctrica representativa del rasgo a reconocer debemos digitalizar dicha señal eléctrica y codificarla de forma que el sistema de reconocimiento biométrico pueda evaluarla de forma cuantitativa, a este proceso se le denomina parametrización.

Las dos etapas descritas anteriormente se emplean tanto para la fase de registro como para la fase de reconocimiento o test. En la fase de registro el usuario aporta sus datos biométricos al sistema, estos son parametrizados y utilizados para generar un modelo estadístico o patrón mediante un proceso denominado entrenamiento. En el entrenamiento se construye un patrón representativo del rasgo del usuario a partir de las muestras aportadas en la fase de registro. Cuanto mayor sea la cantidad de datos más variabilidad del rasgo será recogida, consecuentemente el reconocimiento posterior será mejor. La variabilidad es el hecho de que un mismo rasgo no puede ser adquirido dos veces de forma exactamente igual. Por ejemplo una misma frase pronunciada por un mismo locutor a través de un teléfono móvil o a través de un teléfono fijo producirán parámetros distintos (variabilidad intercanal), esa misma frase pronunciada por el locutor afónico también producirá parámetros distintos (variabilidad intersesión).

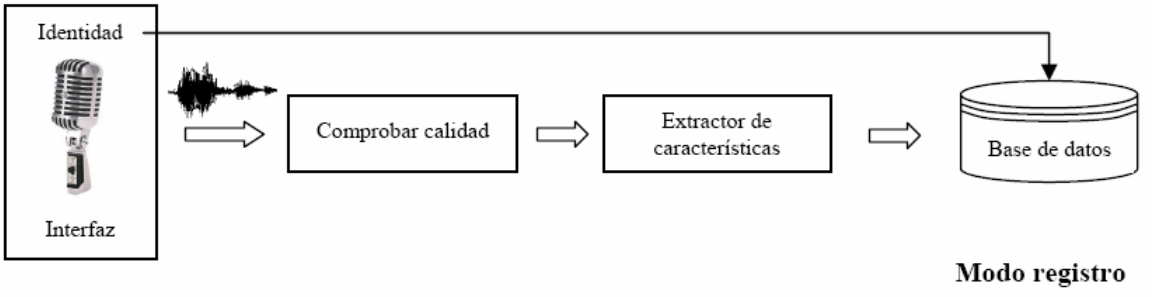


Figura 2: Estructura de un sistema de reconocimiento biométrico en modo registro, figura adaptada de [1]

Una vez tenemos el patrón del rasgo del usuario entrenado, éste es almacenado en una base de datos para posteriormente emplearlo en la fase de reconocimiento. Para determinar la identidad de un sujeto que realiza un intento (en inglés trial) de identificación lo que haremos será comparar el rasgo parametrizado con el patrón almacenado de la identidad pretendida. Ésta es la fase de reconocimiento y como resultado de la misma se obtiene una puntuación (en inglés score). Cuando se compara una muestra de un usuario con el patrón del mismo usuario al intento se le denomina target, si la muestra y el patrón son de usuarios distintos al intento se le denomina non-target.

La puntuación será comparada en la fase de decisión con un umbral fijado previamente, si la puntuación supera el umbral se considerará que el usuario que ha generado la muestra y el que ha generado el patrón son el mismo y consecuentemente hay coincidencia (en inglés match).

2.1.2 Evaluación de un sistema de reconocimiento biométrico

Para evaluar un sistema de reconocimiento biométrico se debe lanzar un elevado número de intentos tanto target como non-target y analizar cómo responde el sistema ante los mismos. El comportamiento del sistema dependerá fundamentalmente del valor del umbral. En las siguientes gráficas se pueden apreciar las funciones de densidad de probabilidad para las puntuaciones de intentos target y non-target dado un sistema genérico, así como la probabilidad de falsa aceptación y falso rechazo en función del umbral:

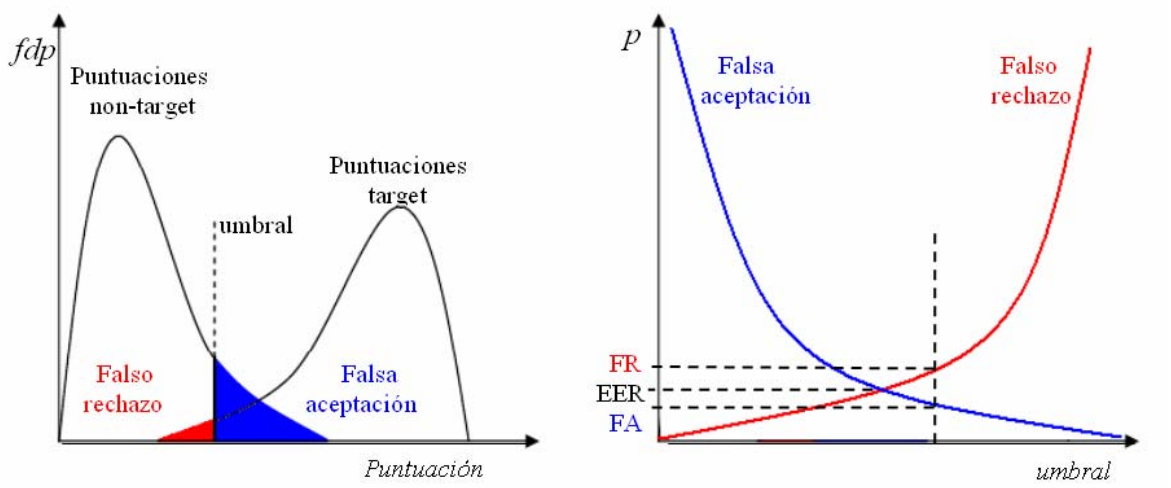


Figura 3: curvas fdp de las puntuaciones target y non-target (izquierda) y probabilidad de FA y FR en función del umbral (derecha)

Como se puede ver en la gráfica de la izquierda al elegir un umbral se estarán clasificando una proporción de los intentos de usuario genuino (target) como impostores y una parte de los intentos de usuario impostor (non-target) como genuino. La tasa de usuarios genuinos clasificados como impostores se denomina tasa de Falso Rechazo (FR) y la tasa de usuarios impostores clasificados como genuinos se denomina tasa de Falsa Aceptación (FA). En la figura de la derecha se puede ver como varía la probabilidad de FR y FA en función del umbral establecido. Si se pone un umbral muy alto no se correrá riesgo de FA, por lo tanto no accederán al sistema impostores, pero entonces un usuario genuino tendrá que realizar varios intentos para acceder al sistema y puede que no lo consiga. Si se pone un umbral muy bajo entonces los usuarios genuinos accederán al sistema sin problemas pero puede que también acceda algún impostor. Por lo tanto existe un compromiso entre FR y FA que se debe valorar en función de la aplicación que se le vaya a dar al sistema. Cuando se tiene la misma tasa de FA y FR nos encontramos ante la Tasa de Igual Error o EER (en inglés Equal Error Rate), que es la tasa que se suele utilizar para comparar sistemas.

Para una visión más amplia del comportamiento del sistema podemos enfrentar la probabilidad de FA y FR en una gráfica, de este modo veremos que probabilidad de FA y FR tenemos para cada umbral escogido, esto es lo que se llama curva ROC (Receiver Operating Curve). Otra representación, más utilizada si cabe, es la curva DET (Detection Error Tradeoff) que es la equivalente a una curva ROC pero con la escala de los ejes cambiada (Figura 4).

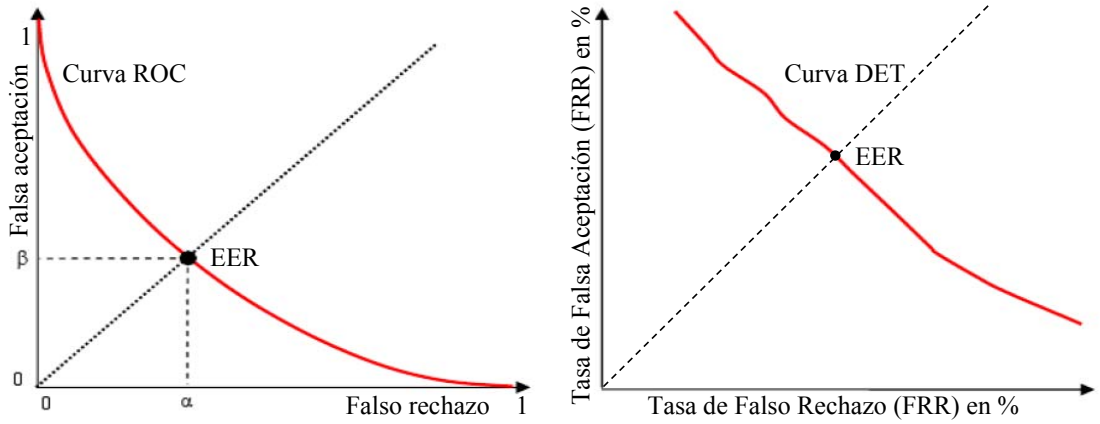


Figura 4: Curvas ROC (izquierda) y DET (derecha)

2.2 Tipos de sistemas de reconocimiento biométrico

A continuación se explicarán los distintos tipos y modos de funcionamiento de un sistema de reconocimiento biométrico, así como las aplicaciones que estos tienen.

2.2.1 Identificación

Un sistema de reconocimiento biométrico funciona en modo identificación cuando realiza una comparación de uno a varios. Esto quiere decir que compara los datos biométricos introducidos en la fase de test con todos los modelos existentes en la base de datos (o con un subconjunto especificado de ellos) con la finalidad de encontrar entre esos modelos la identidad del sujeto.

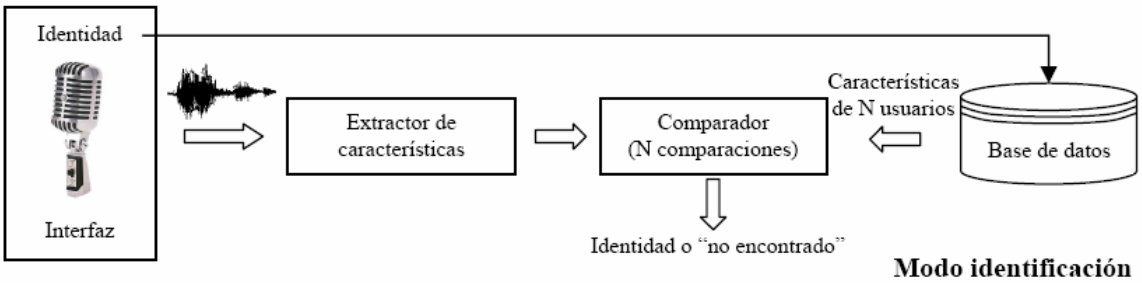


Figura 5: Estructura de un sistema de reconocimiento biométrico en modo identificación, figura adaptada de [1]

Este modo de funcionamiento dará como salida la identidad o el conjunto de identidades más probables del individuo. Aún así es posible que la identidad devuelta por el sistema no sea nadie, es decir, que el modelo que mejor representa al individuo es aquel que representa a los individuos que no están en la base de datos. También puede ser que la identidad más probable tenga aún así una puntuación baja y arroje dudas sobre si es la verdadera identidad del individuo.

Las aplicaciones que este modo de funcionamiento puede tener están enmarcadas sobretodo en el terreno de la seguridad, especialmente en la búsqueda de sujetos. Por

ejemplo, si en el escenario de un crimen se encuentran huellas dactilares se pueden cotejar con una base de datos en modo identificación con el objetivo de encontrar la identidad asociada a esa huella. El tipo de sistemas descritos en el ejemplo se denominan sistemas automáticos de identificación de huella dactilar o AFIS (en inglés Automatic Fingerprint Identification System). Recientemente se han desarrollado también sistemas automáticos de identificación de voz o ASIS (Automatic Speech Identification System).

2.2.2 Verificación

Otro modo de funcionamiento de un sistema de reconocimiento biométrico es el modo verificación. Este modo hace una comparación uno a uno. En este modo el usuario que facilita sus datos biométricos facilita también una identidad pretendida, así el sistema busca en la base de datos el modelo de dicha identidad para enfrentarlo a los datos biométricos facilitados.

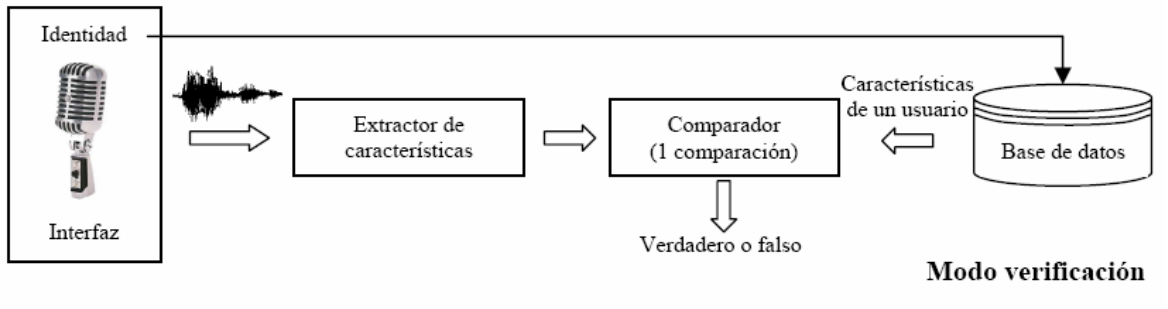


Figura 6: Estructura de un sistema de reconocimiento biométrico en modo verificación, figura adaptada de [1]

En este caso el sistema devolverá una respuesta binaria, o hay coincidencia o no la hay. Por tanto el sistema únicamente verifica la identidad proporcionada por el usuario. Las aplicaciones de este modo de funcionamiento van enfocadas mayoritariamente al control de acceso a sistemas.

2.2.3 Sistemas multimodales

Un sistema multimodal [2] es aquel que basa su decisión, ya sea en modo identificación o en modo verificación, en la comparación de más de un rasgo físico. Por ejemplo un sistema multimodal puede ser uno que combine firma con huella dactilar, o iris con voz. De hecho la combinación puede ser de utilizar un rasgo para identificación y un segundo rasgo para verificación. Así por ejemplo se podría emplear reconocimiento de iris para identificación, y una vez se tiene la supuesta identidad del sujeto realizar una verificación mediante voz o huella dactilar o ambos.

Un buen ejemplo de sistema multimodal es el nuevo DNI que se está implantando en España. Este documento combina una clave llamada firma digital que sólo el usuario conoce con huella dactilar, imagen frontal de la cara y firma offline en una tarjeta con un chip. De este modo con este tipo de identificación podemos validad nuestra identidad de distintas maneras o como fusión de las mismas.

2.2.4 Fusión de sistemas

Fusionar sistemas biométricos consiste en fusionar los resultados de dos o más sitemas biométricos. Un sistema multimodal puede ser una fusión de dos sistemas de reconocimiento biométricos de varios rasgos distintos (fusión multimodal). Pero también se pueden fusionar sistemas basados en el mismo rasgo.

Por ejemplo se puede fusionar un sistema de reconocimiento de locutor dependiente de texto basado en GMMs con otro basado en HMMs. De este modo se tendrá un sistema con fusión multinivel o con fusión de algoritmos. Si se eligen bien las técnicas que vamos a utilizar, la fusión multinivel puede hacer que el sistema mejore mucho, pero para que esto suceda las técnicas fusionadas tienen que ser complementarias, es decir que una de ellas supla las carencias de la otra.

Por otra parte se pueden fusionar sistemas a nivel de decisión. Varios sistemas pueden tomar decisiones distintas de aceptar o rechazar a un usuario, en este caso se puede elegir la opción mayoritaria o estudiar la forma de ponderar las decisiones de los distintos sistemas.

Puede realizarse también fusión a nivel de sensor, por ejemplo se pueden fusionar varios sistemas de captura de imagen facial en 2 dimensiones para obtener un sistema de reconocimiento de cara en 3 dimensiones. También se puede fusionar un sistema que emplee un sensor de huella dactilar óptico con otro que use un sensor de huella dactilar térmico.

Otra técnica de fusión sería emplear un mismo sistema, por ejemplo de reconocimiento de iris o de huella dactilar, y fusionar los resultados obtenidos para distintas características o instancias, fusionando el resultado obtenido con el iris izquierdo con el resultado obtenido con el iris derecho o fusionando los resultados obtenidos para huella de los dedos índice, anular y pulgar, por poner otro ejemplo.

2.3 Los rasgos

Hay dos tipos de rasgos biométricos principalmente, los rasgos fisiológicos o estáticos y los rasgos de comportamiento o dinámicos. Los rasgos fisiológicos son implícitos al individuo, como por ejemplo la huella dactilar, la cara, la retina, las venas del dorso de la mano, la geometría de la mano o el iris, no requieren ningún comportamiento por parte del individuo para su identificación. Los rasgos dinámicos o conductuales como la voz, la forma de andar, dinámica de tecleo o firma o escritura manuscrita, sí requieren de un comportamiento o actuación por parte del individuo. En la siguiente tabla podemos ver los rasgos más empleados:

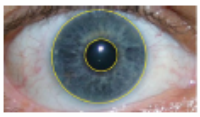
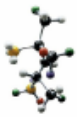
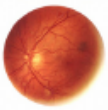
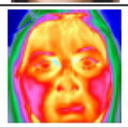
ADN		Iris	
Dinámica de tecleo		Olor	
Escáner de retina		Oreja	
Firma		Rostro	
Forma de caminar		Termograma facial	
Geometría de la mano		Venas de la mano	
Huella dactilar		Voz	

Tabla 1: Principales rasgos biométricos

Los rasgos biométricos deben tener en mayor o menor medida las siguientes características para ser considerados identificativos [3] [4]:

- Universalidad (U): todas las personas deben poseerlo.
- Distintividad (D): un mismo rasgo para dos personas diferentes no puede ser idéntico. Un rasgo epigenético es aquel que es independiente de la genética, es decir que dos hermanos gemelos deberían presentar dos formas distintas para ese mismo rasgo.
- Estabilidad (E): los rasgos deben permanecer idénticos en el tiempo.
- Mensurabilidad (M): deben permitir ser evaluados cuantitativamente.
- Rendimiento (R): el sistema de identificación que los utilice debe ser preciso, emplear pocos recursos y no verse afectado por los factores ambientales o de trabajo que lo rodean.
- Aceptabilidad (A): Debe tener un alto grado de aceptación en la sociedad.
- Seguridad (S): El sistema debe ser seguro, y difícil de engañar de forma fraudulenta.

A continuación podemos ver en la Tabla 2 una valoración del nivel (alto, medio o bajo) en el que encontramos estas características en los diferentes rasgos:

Rasgo	U	D	E	M	R	A	S
Voz	M	B	B	M	B	A	A
Huella	M	A	A	M	A	M	M
Firma	B	B	B	A	B	A	A
Iris	A	A	A	M	A	B	B

Tabla 2: Propiedades de los distintos rasgos biométricos, tabla adaptada de [5]

2.3.1 Voz

La voz es el principal instrumento de comunicación de los seres humanos, y es probablemente el rasgo más aceptado socialmente. La voz cambia con el tiempo pero de forma muy progresiva a partir de la adolescencia. Como rasgo conductual (reflejado en la prosodia) es relativamente fácil de imitar, pero es muy seguro como rasgo fisiológico. Las técnicas que se emplean han evolucionado mucho a lo largo del tiempo pero tienen una base muy sólida que ha demostrado buenos resultados. Este es el rasgo del que hablaremos en profundidad a lo largo del proyecto.

Anualmente se realizan evaluaciones competitivas a cargo del NIST (National Institute of Standards and Technology) de reconocimiento de locutor (SRE, Speaker Recognition Evaluation) [6].

2.3.2 Huella Dactilar

Actualmente supone más del 40% del mercado del reconocimiento biométrico. La huella dactilar es permanente desde el séptimo mes fetal hasta la descomposición post-mortem. Sin embargo puede verse alterada con relativa facilidad a causa de cortes o quemaduras. Históricamente ha sido aceptado como un símbolo de identidad (desde la antigua China) y durante el siglo pasado ha sido el rasgo biométrico más empleado en investigaciones policiales. Hace tiempo que existen extensas bases de datos de huella dactilar y AFIS (Automatic Fingerprint Identification System) o sistemas automáticos de identificación de huella dactilar. Los principales problemas que presenta son que es relativamente fácil atacar un sistema con huellas de goma, aunque ya hay sistemas de detección de vida, así como la falta de interoperabilidad entre distintos tipos de sensores.

Cada 2 años se organizan evaluaciones competitivas [7] denominadas FVC (en inglés, Fingerprint Verification Competition).

2.3.3 Firma

Hay dos tipos de reconocimiento de firma manuscrita, on-line y off-line. La diferencia está en que la firma manuscrita on-line es adquirida con un bolígrafo especial sobre una superficie preparada para poder medir presión, altitud, velocidad, ángulo…, mientras que la firma manuscrita off-line simplemente se basa en la imagen de la firma, siendo ésta mucho más fácil de falsificar. La firma es un símbolo de identidad propia y voluntaria y

refleja muchos aspectos de la personalidad del individuo. Está muy aceptado socialmente, su adquisición es poco invasiva y cada vez hay más dispositivos capaces de medir parámetros de firma on-line sin que estén especialmente diseñados para ello, como móviles, PDAs, Tablet PCs…, aunque entre ellos suelen presentar variabilidad. La principal ventaja de la firma on-line frente a la firma off-line es que un falsificador puede imitar la forma de la firma pero muy difícilmente imitará la dinámica de la firma.

2.3.4 Iris

Según Flom y Safir [8] "No hay 2 iris iguales", y es verdad, puesto que es un rasgo epigenético. Desde la niñez es un rasgo que prácticamente no varía, aunque se pierde algo de pigmentación en la vejez. Es un rasgo muy distintivo y con un gran poder discriminante, pero tiene un gran inconveniente, y es que su adquisición resulta bastante incómoda para el usuario debido a la iluminación necesaria para una buena toma. Por ello requiere de un alto grado de cooperación por parte del usuario.

Cada año el NIST organiza evaluaciones competitivas ICE (Iris Challenge Evaluation) [9].

2.3.5 Otros rasgos

Existen además muchos otros rasgos biométricos, que por sus características están menos desarrollados, pero que también merecen una mención por su emergente implantación en el mercado. De ellos, probablemente el más importante sea la cara por lo fácil que resulta obtener una imagen de la misma. También se realiza reconocimiento biométrico con las radiografías de piezas dentales, que aunque no son perennes pueden valer como reconocimiento complementario, es una técnica muy usada en reconocimiento biométrico de víctimas de incendio. La forma de la oreja también se utiliza para el reconocimiento biométrico, su característica más importante es que mediante técnicas térmicas puede extraerse a través del pelo, es invariante y no cambia cuando se habla o se gesticula. Otra técnica es la extracción del patrón vascular de la retina, aunque es muy invasiva, se suele utilizar para reconocimiento de ganado. Otros rasgos como la forma de andar son útiles cuando no se tiene otra cosa que una grabación del sujeto andando, aunque su uso no está muy extendido. La termografía facial presenta la ventaja de que no es en absoluto invasiva y el pelo no dificulta su extracción. El reconocimiento de olor no está especialmente desarrollado pero puede servir de complemento a la huella dactilar para la detección de vida.

Por último hay parámetros que, aunque no pueden ser considerados parámetros biométricos por no cumplir las características anteriormente mencionadas, sí pueden servir como complemento para aportar fiabilidad al reconocimiento. Estos parámetros son la altura, el color de ojos o el peso.

2.4 Repercusión social

Un sistema de reconocimiento biométrico triunfará en la medida que la sociedad quiera emplearlo. Desde ese punto de vista parece que los sistemas menos invasivos son los socialmente más aceptados. También existen distintos entornos en los que un sistema de reconocimiento biométrico puede resultar más o menos atractivo. Por ejemplo es difícil que alguien se moleste porque le pidan que firme en un supermercado al pagar la factura. Sin embargo puede haber gente que se sienta molesta si al hacerse el DNI le piden que deje sus huellas dactilares porque puede sentirse vigilado de alguna forma.

El reconocimiento biométrico es una disciplina que despierta recelos entre algunas personas y admiración y confianza en otras. Ambas posiciones son comprensibles desde cierto punto de vista. En la sociedad del bienestar en la que vivimos a todo el mundo le gusta sentirse seguro. Cualquiera que tenga ahorros en el banco quiere estar seguro de que no se los robarán, si hacemos una compra por internet queremos garantías y si tenemos una casa no queremos sufrir robos. Para evitar todo este tipo de situaciones incómodas, la biometría resulta ser un importante aliado.

El problema viene cuando las técnicas de reconocimiento biométrico son empleadas injustificadamente o con una justificación un tanto dudosa, especialmente cuando su uso puede suponer una invasión de nuestra intimidad. Así como las técnicas de reconocimiento biométrico pueden suponer una de las mejores garantías para mantener nuestra privacidad, si el sistema biométrico es asaltado con éxito nuestra privacidad, datos personales y bienes más preciados pueden quedar completamente al descubierto. Alguien que pudiera asaltar un sistema de reconocimiento biométrico podría realizar en nuestro nombre cualquier tipo de actividad, como por ejemplo retirar nuestros ingresos de una cuenta bancaria, además de cometer delitos en nuestro nombre.

Por otra parte, un sistema de reconocimiento biométrico usado de forma desmedida puede tener controladas prácticamente todas nuestras actividades, las llamadas telefónicas pueden ser seguidas reconociendo nuestra voz y con que exista una cámara en la calle podemos ser reconocidos por un sistema de reconocimiento facial. Lo que afortunadamente ocurre es que para que este tipo de actividades se den tiene que darse un claro abuso de autoridad si es que fuera un cuerpo de seguridad quien lo hiciera, o bien un enorme despliegue de medios prácticamente inabordable, en el caso de una iniciativa privada. En cualquier caso, hoy por hoy parece más un argumento de una novela de ciencia ficción que una realidad. De modo que de momento parece evidente que los sistemas de reconocimiento biométrico pueden aportarnos más cosas positivas que negativas.

3 Sistemas de reconocimiento de locutor

3.1 Características de la voz

Para entender el funcionamiento de un sistema de reconocimiento de locutor, primero se debe entender la naturaleza física del rasgo que se empleará para realizar el reconocimiento. Por ello se debe conocer cómo se produce la voz y cuáles son sus características que permiten distinguir a un locutor de otro.

3.1.1 La señal de voz

El proceso de generación de voz comienza cuando el diafragma comprime los pulmones y genera una diferencia de presión entre los pulmones y el exterior, suministrando así un flujo de aire. Desde un punto de vista de señal este flujo se comporta como una fuente de ruido blanco de pequeña amplitud.

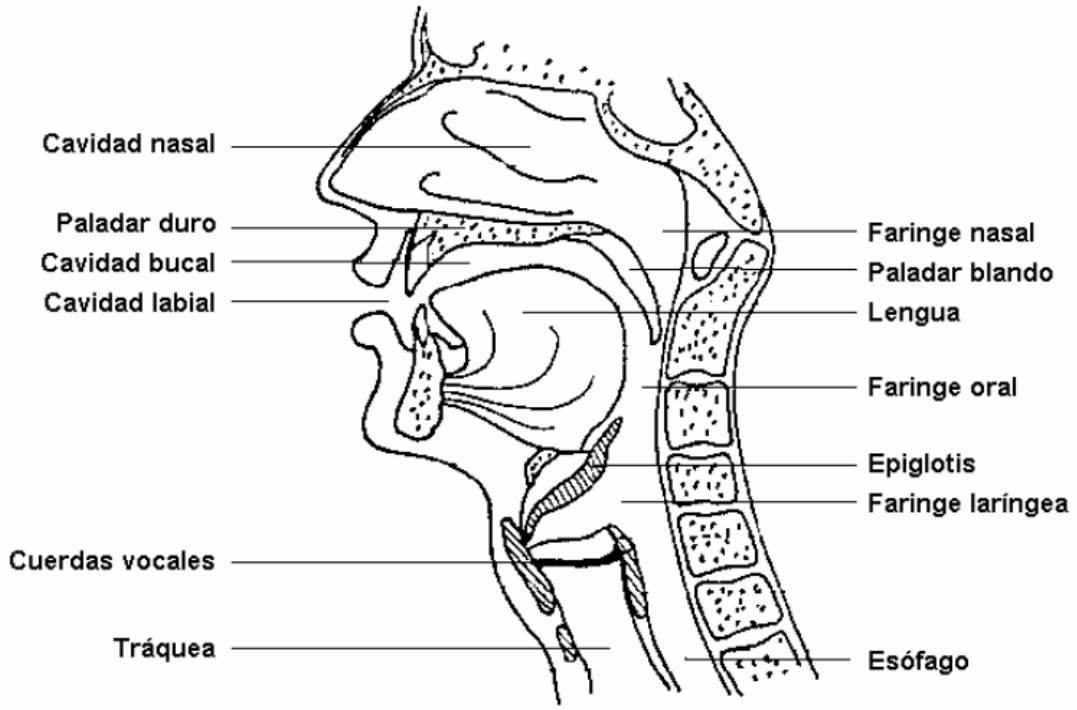


Figura 7: El tracto vocal

Tras atravesar la tráquea, el aire llega a la laringe donde hace vibrar las cuerdas vocales. La laringe es el órgano donde propiamente se produce la voz, tanto su tono fundamental como sus armónicos. Las cuerdas vocales están formadas por cuatro repliegues divididos en repliegues superiores y repliegues inferiores. Los superiores son conocidos por cuerdas falsas y no son de gran relevancia durante el proceso de generación de voz. Los inferiores son las verdaderas cuerdas vocales. Entre ambas se encuentra un pequeño espacio vacío conocido como la glotis.

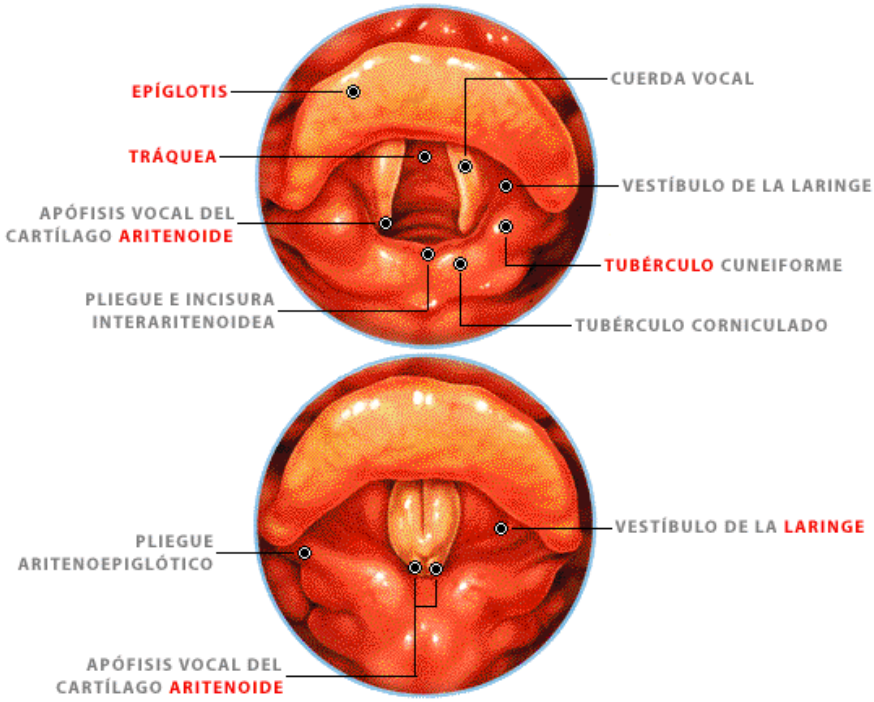


Figura 8: Las cuerdas vocales

En la respiración, las cuerdas vocales se abren y permiten el paso del aire sin hacer gran resistencia. Sin embargo cuando se emiten sonidos las cuerdas vocales inferiores se juntan. Pueden constreñirse parcialmente, o bien vibrar como lo haría un oscilador, generando una onda de presión sinusoidal conocida como tono o frecuencia fundamental. Las cuerdas vocales inferiores son responsables de las siguientes características del sonido:

- Los sonidos sonoros o vocálicos se producen cuando las cuerdas vocales se aproximan y vibran. En caso contrario, cuando estas solo se cierran parcialmente y se abren abruptamente, los sonidos producidos son conocidos como sonidos sordos.
- La frecuencia a la que vibran las cuerdas vocales es conocida como frecuencia fundamental, tono, o pitch. Esta vibración provoca una onda sonora, compuesta por el tono fundamental y unos armónicos. Posteriormente, estos armónicos son filtrados en las cavidades laríngea, bucal y nasal (cavidad naso-buco-faríngea) donde se forma el timbre del sonido.
- La intensidad de la señal de voz depende de la energía del flujo de aire al pasar hacia las cuerdas vocales y del tipo de sonido: sordo o sonoro. Típicamente los sonidos sonoros presentan mayor amplitud que los sordos.

Cuando hablamos ni los valores de energía ni los de frecuencia fundamental son constantes, sino que describen contornos. La forma en la que se producen estos contornos y la frecuencia con la que aparecen contornos específicos es un parámetro discriminante entre locutores ya que estos son causados por variaciones de longitud, espesor y tensión de las cuerdas vocales, y dependen tanto de sus características morfológicas como de su comportamiento. Cuando estas se cierran, se corta el flujo de aire, cuando se abren se libera la onda de presión acumulada. El grado en el que este efecto se repite está relacionado con el tipo de fonación.

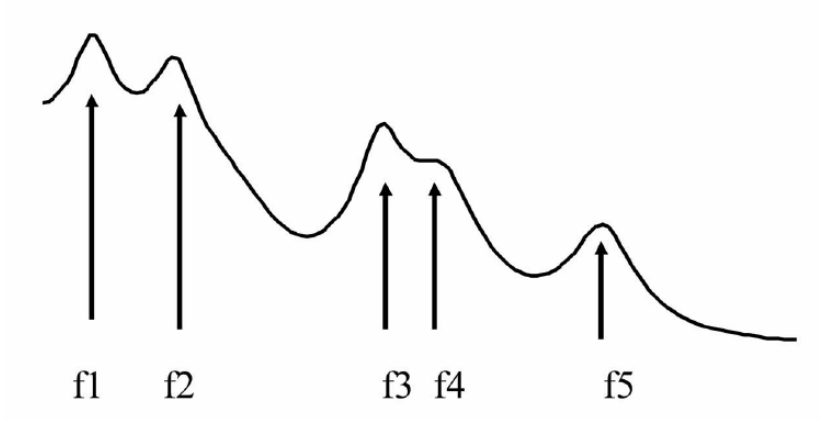


Figura 9: Los formantes en la envolvente espectral

La señal resultante de atravesar las cuerdas vocales, pasa por la laringe, donde sufre una modificación en la caja de resonancia formada por la cavidad naso-buco-faríngea. En dicha cavidad es donde se amplifica la voz y esta adquiere una envolvente espectral característica que se conoce con el nombre de timbre. Las resonancias que se producen en la cavidad naso-buco-faríngea tienen su energía concentrada alrededor de unas frecuencias específicas del sonido y la persona. Estas frecuencias de resonancia son conocidas como formantes.

Los formantes dependen casi exclusivamente del tamaño y forma de la cavidad naso-buco-faríngea y son más fuertes (distinguibles) para sonidos sonoros (vocales y consonantes sonoras) que para las consonantes no sonoras. El concepto de los formantes es de extraordinaria importancia en cualquier ámbito de procesado de voz, ya que en ellos está concentrada también una gran cantidad de información del mensaje contenido y del locutor.

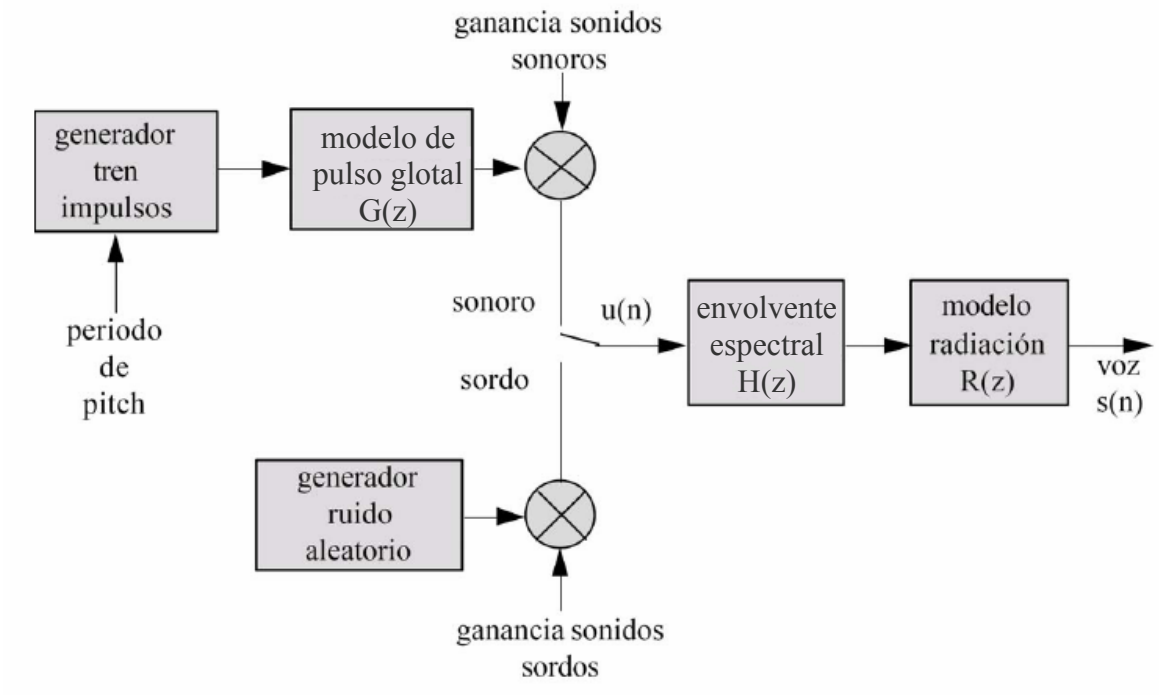


Figura 10: Esquema del modelo simplificado de producción de voz

Por tanto, el espectro de la señal de voz, en cada instante se compone fundamentalmente de 2 componentes claramente diferenciadas: la envolvente espectral y la estructura fina. La estructura fina de la señal de voz se debe a la excitación de las cuerdas vocales y no es muy discriminatoria entre locutores si se compara con la envolvente espectral. La envolvente espectral depende casi exclusivamente de las características morfológicas de cada individuo y del tipo de sonido. En la Figura 10 se puede ver el modelo simplificado de producción de voz como combinación de un conjunto de sistema lineales.

En realidad, la mayor parte de la información de locutor no solo depende de las características instantáneas del timbre, sino que también depende de la forma en la que el timbre varia durante en la transición entre sonidos, lo cual se conoce como coarticulación. La coarticulación es el resultado del funcionamiento coordinado de todo el aparato fonador.

3.1.2 El sistema auditivo

El sistema auditivo recoge las ondas, las amplifica, las convierte en impulsos eléctricos que son enviados hacia el cerebro. También sirve para ubicarse espacialmente e informa de los movimientos corporales en 3 dimensiones. El oído se compone de oído externo, oído medio y oído interno.

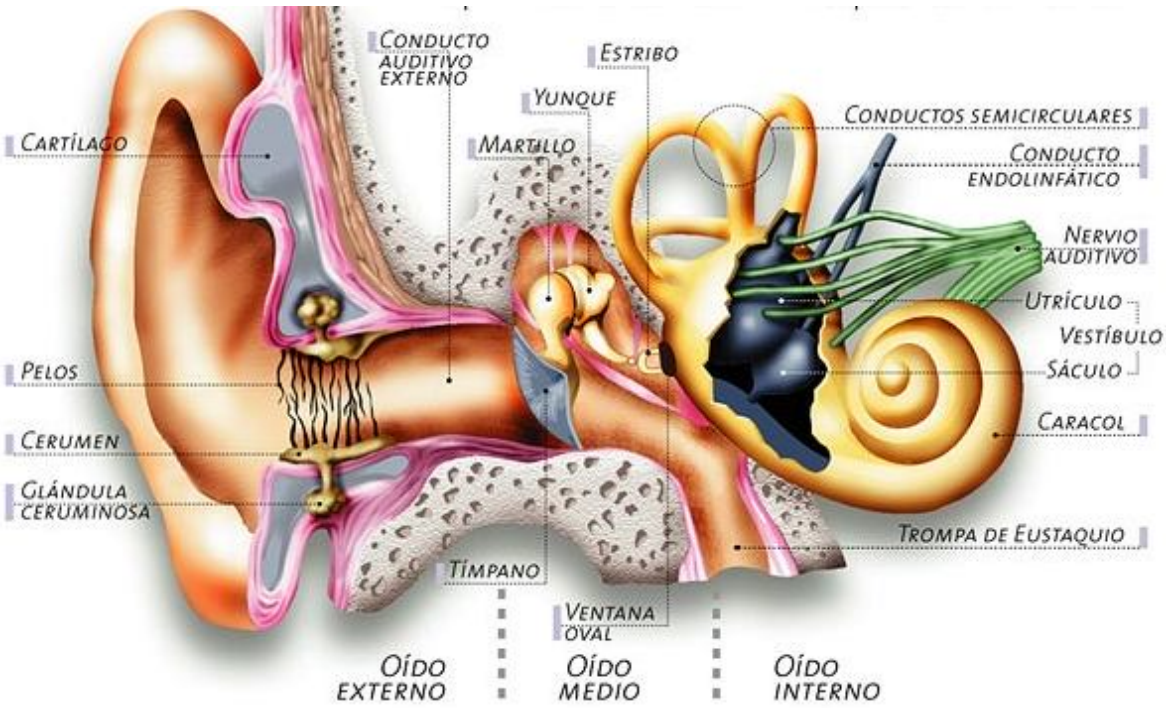


Figura 11: El oído humano

El oído externo se compone de pabellón, conducto auditivo y Tímpano. La función principal del pabellón es captar el sonido y dirigirlo hacia el conducto auditivo, también capta la dirección del sonido. El conducto auditivo conduce las ondas hacia la membrana timpánica, las amplifica y actúa de acoplador de impedancias. Funciona como resonador, amplificando las frecuencias de la banda critica que contienen la inteligibilidad del habla. El tímpano es un transductor, convierte la energía sonora en energía mecánica.

El oído medio tiene como función principal propagar la energía hacia el oído interno. Se compone de martillo, yunque, estribo (que forma la cadena de huesecillos), lenticular y la trompa de Eustaquio. El tímpano vibra a la misma intensidad y frecuencias que llegan del oído externo, esta información es transmitida a los 3 huesesillos, los cuales tienen la función de mantener un equivalente de energía entre el aire que se encuentra en el oído externo y el liquido que se encuentra en el oído interno, y vibran la misma intensidad y frecuencia que el tímpano. La trompa de Eustaquio, actúa de ecualizador de presión atmosférica, de ella depende mantener el equilibrio. La trompa de Eustaquio tiene una membrana que sirve como relguador atmosférico, por eso cuando cambia la presión atmosférica, los oídos se tapan, esto es una protección para evitar que el tímpano se reviente.

El oído interno convierte la energía en pulsos eléctricos y estos son captados por el nervio auditivo que los envía hacia el cerebro para interpretarlos como habla, música o ruido. El estribo hace contacto con el lenticular, que es un hueso de forma oval, que a su vez hace contacto con la ventana oval (membrana). Del otro lado de la ventana oval se encuentra el caracol que tiene 3 cavidades llenas de líquido, en uno de estos canales se encuentra la membrana Basilar donde se despliega el órgano de Corti. Este órgano tiene una franja de células muy sensibles que tienen unas microvellosidades conocidas como células ciliadas, conectadas al nervio auditivo. El líquido del caracol se mueve, este hace que se mueva la membrana basilar, las células de esta membrana son estimuladas y éstas hacen que se muevan las células ciliadas, a su vez éstas generan impulsos eléctricos los cuales son mandados al nervio auditivo, el cual recolecta esta información y la envía hacia el cerebro.

La frecuencia es determinada por la zona de células que se mueven a lo largo del órgano de Corti y la amplitud está determinada por la cantidad de células ciliadas que se muevan. La membrana Basilar se va estrechando conforme nos adentramos en el caracol, pero este estrechamiento no es lineal. El caracol o cóclea actúa como un analizador de espectros, como un banco de filtros, pero estos filtros no están distribuidos a lo largo de la frecuencia de forma lineal (debido a la forma en que la membrana Basilar se estrecha). Es por esto por lo que percibimos las distancias tonales en octavas (múltiplos potencia de 2 de la frecuencia de referencia).

3.1.3 Parametrización de la señal de voz

La parametrización que se suele utilizar en los sistemas de reconocimiento de locutor tiene mucho que ver con el apartado que acabamos de ver, se conoce como MFCC (en inglés Mel-Frequency Cepstral Coefficients). Los parámetros extraídos están basados en coeficientes cepstrales extraídos a partir de un banco de filtros de frecuencias Mel. Un banco de filtros de frecuencias Mel está basado en la percepción acústica del oído humano (no lineal), así que los filtros tienen una separación de la frecuencia central que va creciendo conforme la frecuencia es mayor (en realidad esta distancia es constante e igual a 100 Hz hasta 1KHz), además el ancho de banda de los filtros también crece conforme aumenta la frecuencia. Los filtros son triangulares y el ancho de banda se calcula de manera que la frecuencia de corte inferior coincida con la frecuencia central del filtro de frecuencia central inmediatamente inferior, la frecuencia de corte superior coincidirá con la frecuencia central del filtro de frecuencia central inmediatamente superior.

El sistema enventana la señal de audio con una ventana de Hamming en ventanas de 25 ms de audio con un solape de 15 ms. A cada una de las ventanas resultantes se le aplica la transformada rápida de Fourier o FFT (en inglés Fast Fourier Transform), resultando un vector de coeficientes que son el módulo de la FFT de la ventana.

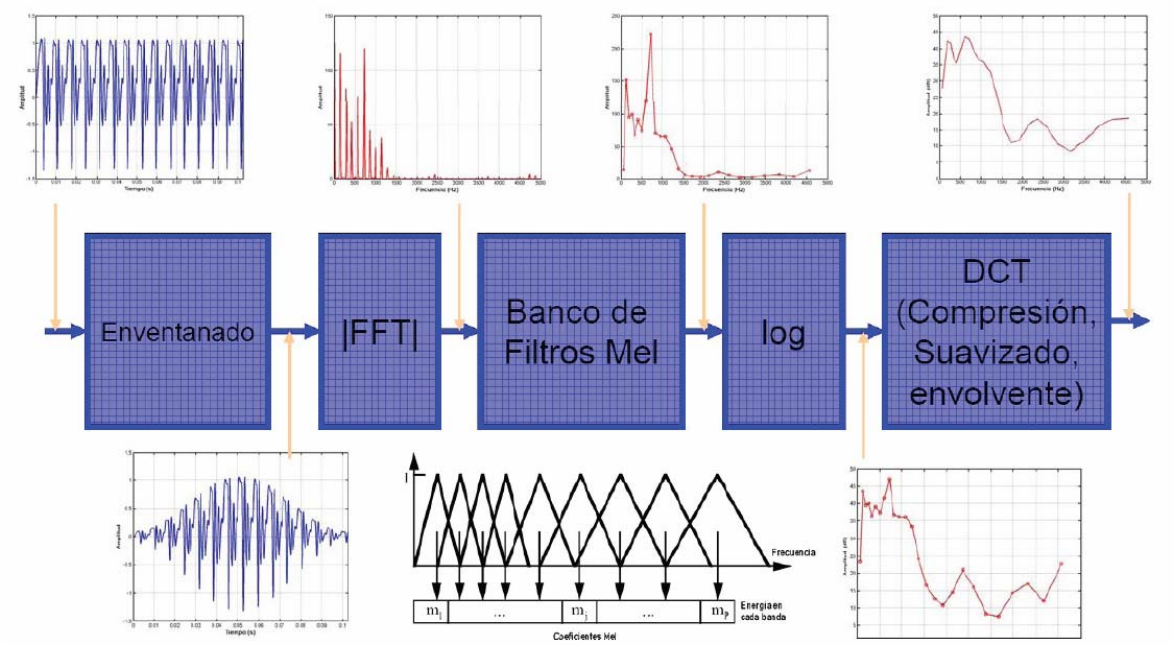


Figura 12: Esquema de la parametrización MFCC

La señal en el dominio transformado es filtrada mediante un banco de filtros de frecuencia Mel. Posteriormente se calcula el logaritmo de la energía de la señal filtrada con cada uno de los filtros, teniendo tantos coeficientes como filtros que son almacenados en un vector. Finalmente a dicho vector se le aplica la transformada discreta del coseno o DCT (en inglés Discrete Cosine Transform), obteniendo el vector de parámetros final.

Para recoger el comportamiento dinámico de los coeficientes MFCC se suele tener en cuenta, además la diferencia entre el vector MFCC siguiente y el vector MFCC previo al vector bajo análisis, obteniendo así los coeficientes Δ. Si además queremos observar como varían los coeficientes Δ se puede volver a repetir la resta entre el vector Δ siguiente y el vector Δ previo al vector Δ bajo estudio, obteniendo así un vector de coeficientes Δ Δ.

3.2 Tipos de reconocimiento de voz

Las tecnologías de reconocimiento de voz pueden ser empleadas en multitud de aplicaciones. En este proyecto nos centraremos en el reconocimiento de locutor pero merece la pena reseñar algunas de las demás aplicaciones que este tipo de tecnologías nos ofrecen.

La primera de ellas es el reconocimiento de habla espontánea. Su objetivo es convertir en texto lo que el locutor dice de forma automática. Las aplicaciones que esto tiene van desde editores de texto, proporcionando una interfaz cómoda y rápida al usuario, hasta aplicaciones orientadas a discapacitados, que mediante la voz pueden acceder a programas informáticos que de otra manera serían inaccesibles para ellos, pasando por aplicaciones

domóticas o cualquier tipo de mecanismos susceptibles de recibir instrucciones, de esta manera ya no es necesario teclear la instrucción o desplazarse hasta un terminal para activar la función mediante un click de ratón o presionar un botón, siempre que tengamos un micrófono en la habitación o un micrófono inalámbrico con nosotros podremos ejecutar comandos u órdenes.

Otra aplicación relacionada con la primera es la denominada Word Spotting, cuyo objetivo es buscar determinadas palabras en grabaciones de voz. Puede ser utilizado, como en el caso anterior, para dar instrucciones a un PC o algún tipo de automatismo, pero también puede ser empleado para realizar búsquedas de contenidos, por ejemplo en internet, o para indexar y clasificar grabaciones de audio que contengan voz.

También hoy en día se invierten muchos esfuerzos en reconocimiento de idioma. Empleado en combinación con otras disciplinas puede ser empleado en aplicaciones de aprendizaje de una lengua, como fase previa al reconocimiento de habla espontánea, para indexación y clasificación de grabaciones de voz, como complemento a sistemas de traducción automática de idioma, para segmentar una conversación de varias personas en varios idiomas o para realizar búsquedas de grabaciones en dicho idioma.

En este proyecto nos centraremos en el reconocimiento de locutor, es decir en la verificación de la identidad de una persona a través de su voz. Sus aplicaciones son fundamentalmente la seguridad, control de paso a lugares de acceso restringido o identificación de un locutor presente en una grabación. Existen fundamentalmente 2 tipos de reconocimiento de locutor, reconocimiento de locutor independiente de texto y reconocimiento de locutor dependiente de texto.

3.2.1 Reconocimiento de locutor independiente de texto

Esta disciplina consiste en el reconocimiento de un locutor a partir de una grabación que se tenga suya. No se sabe lo que está diciendo ni en las grabaciones en las que se tiene conocimiento de su identidad ni en las que no se sabe. Por lo tanto se debe realizar un modelo genérico de su voz.

Con las grabaciones de voz que se tengan de un sujeto de identidad conocida se entrena un patrón genérico universal (representa a una amplia variedad de locutores) modificándolo para que dicho modelo refleje las características distintivas del sujeto. Para realizar el entrenamiento, en reconocimiento de locutor independiente de texto, la técnica más popular hasta ahora son los modelos de mezclas de Gaussianas o GMMs (del inglés Gaussian Mixture Models). Cuanta más cantidad de habla utilicemos para el entrenamiento mejor entrenado estará el sistema y mejor reconocerá al locutor.

Una vez entrenado el modelo del locutor podemos realizar pruebas de reconocimiento de locutor. En estas pruebas se enfrenta una grabación con la voz de un individuo al modelo previamente entrenado representativo de la identidad pretendida, enfrentamiento del cual obtendremos una puntuación que de momento consideraremos provisional. Después se realiza otro enfrentamiento de la grabación, pero esta vez con el modelo universal, obteniendo una segunda puntuación provisional. La puntuación final será aquella resultante de la resta de las 2 anteriores.

3.2.2 Reconocimiento de locutor dependiente de texto

Esta disciplina se diferencia de la anterior en que para este caso conocemos el contenido léxico de la frase, tenemos una trascripción textual de lo que dice el locutor tanto en las locuciones de entrenamiento como en las locuciones de test. Este hecho nos permitirá realizar el reconocimiento a un nivel de detalle mayor, esto significa que podremos realizar el reconocimiento de locutor palabra a palabra o incluso fonema a fonema. La ventaja que esto aporta es que podemos tener tantos modelos por locutor como fonemas haya y todos ellos entrenados y representativos del locutor.

Para aprovechar esta ventaja de la que disponemos en reconocimiento de locutor dependiente de texto es habitual utilizar Modelos Ocultos de Makov o HMMs (en inglés Hidden Markov Models). Esta técnica permite modelar, además de las distintas partes de un fonema (estados), las transiciones entre los distintos estados, siendo cada uno de estos estados, a su vez, un GMM. Estos modelos son entrenados con grabaciones de voz además de con las trascripciones de texto de lo que se está diciendo en la grabación, de esta forma cada fragmento de audio sirve para entrenar un modelo en particular.

De la misma manera en la fase de test se enfrenta una locución con una trascripción asociada a los modelos del locutor pretendido, así enfrentamos cada fragmento de la locución al modelo correspondiente del locutor por una parte y al modelo correspondiente universal (independiente del locutor) por otra parte. A continuación se restan las puntuaciones obtenidas con el modelo dependiente del locutor e independiente del locutor para cada enfrentamiento, obteniendo una puntuación final para cada fragmento de audio. Finalmente se realiza un promedio de las puntuaciones de todos los fragmentos de audio analizados para obtener la puntuación final de la frase.

4 Estado del arte en reconocimiento de locutor dependiente de texto

4.1 Introducción

El reconocimiento de locutor dependiente de texto es una disciplina de la Biometría que por el momento ha suscitado menos interés investigador comparada con reconocimiento de locutor independiente de texto. Esto es debido probablemente a la ausencia de evaluaciones competitivas, como sí ocurre para el caso de reconocimiento de locutor independiente de texto [10]. Sin embargo por sus características técnicas y su funcionalidad es una disciplina que tiene mucho potencial en aplicaciones comerciales. Esto es debido a que es capaz de conseguir porcentajes aceptables de EER (Equal Error Rate) con muy poca cantidad de audio para entrenamiento (menos de 8 segundos), mientras que para Reconocimiento de locutor independiente de texto para un mismo EER se necesita mucho más (más de 30 segundos).

En la siguiente sección analizaremos los principales elementos que componen un sistema de reconocimiento de locutor dependiente de texto. Posteriormente veremos las principales restricciones que sufren estos sistemas. Por último, veremos algunos de los resultados más relevantes en cada uno de los aspectos vistos.

4.2 Principales elementos de un sistema de reconocimiento de locutor dependiente de texto

En esta sección vamos a ir viendo una por una las diferentes etapas que componen un sistema de reconocimiento de locutor dependiente de texto, así como las principales técnicas utilizadas en cada una de estas etapas para luego poder analizar (en siguientes secciones) qué problemas y resultados hay debidos a cada una de las etapas y cómo influyen éstos en la precisión o exactitud del sistema y su comportamiento.

4.2.1 Parametrización

Lo primero que tenemos que hacer con el audio disponible para ser objeto de reconocimiento es parametrizarlo para que las herramientas matemáticas de reconocimiento puedan trabajar con este material. Hay varias formas de parametrizacón de la voz, basadas en análisis frecuencial o basadas en coeficientes de predicción lineal. Por otra parte debido a que la voz puede venir de distintas fuentes o canales y con ello el sistema de reconocimiento varía su comportamiento (variabilidad intercanal) existen diferentes técnicas de compensación de canal.

Las técnicas de parametrización más utilizadas son MFCC y LPCC. MFCC (Mel Frequency Cepstral Coeffiecients) [11] se basa en un sistema de filtros similar al que efectúa el sistema auditivo humano. LPCC (Linear Predicctive Cepstral Coefficients) modela la voz mediante una predicción lineal. Recientemente la técnica Dynamic Features ha demostrado dar buenos resultados [12] aunque su uso no se ha extendido todavía.

En compensación de canal las técnicas más utilizadas han sido mucho tiempo Feature Mapping [13], y Speaker Model Synthesis [14] (en menor medida), aunque a día de hoy Factor Analisys [15] es la técnica que da mejores resultados.

4.2.2 Modelado acústico

La principal diferencia entre reconocimiento de locutor dependiente de texto e independiente de texto está en que en reconocimiento de locutor dependiente de texto se realiza un reconocimiento conociendo las palabras que se están diciendo. Por ello podemos realizar un reconocimiento por partes de la locución, es decir podemos realizar reconocimiento palabra a palabra o fonema a fonema por poner un par de ejemplos.

Debido a la existencia de esta posibilidad, típicamente la herramienta estadística utilizada para este tipo de reconocimiento son los modelos ocultos de Markov [16] mientras que para reconocimiento de locutor independiente de texto se utilizan típicamente modelos de mezclas gaussianas o GMM [17] y SVM (en inglés Support Vector Machine) [18], que han demostrado producir excelentes resultados. En [19] prueban un sistema dependiente de texto en el que mezclan SVM y HMM con excelentes resultados. Otros tipos de modelado utilizados en los principios del reconocimiento de locutor dependiente de texto son el alineamiento temporal dinámico o DTW (en inglés Dynamic Time Warping) [20] y redes neuronales [21].

4.2.3 Puntuación

La forma de puntuar en reconocimiento de locutor dependiente de texto depende de si se trata de identificación o verificación. Para verificación, que es el caso que nos interesa, la puntuación se basa en el enfrentamiento entre 2 hipótesis, que son que la frase haya sido pronunciada por el locutor genuino o que no haya sido pronunciada por el locutor genuino.

$$L(X \mid \lambda) = \log\left(p\left(X \mid \lambda\right)\right) - \log\left(p\left(X \mid \overline{\lambda}\right)\right)$$

Donde X representa los vectores de parámetros extraídos de la locución y λ representa el modelo del locutor genuino (normalmente un modelo HMM). Para obtener el modelo del locutor genuino se realiza un entrenamiento (que veremos en el siguiente punto). Las puntuaciones de las distintas partes de la locución se obtienen mediante el algoritmo de Viterby. $\overline{\lambda}$ Representa el modelo de la hipótesis de que el locutor no sea el genuino, para obtener este modelo se utilizan típicamente 2 técnicas: seleccionar puntuaciones de una cohorte de impostores (con sus respectivos modelos) y mezclarlas o bien crear un único modelo de impostor universal (Universal Background Model, UBM) y generar una única puntuación de impostor. Hay diferentes formas y criterios de generar este UBM, que iremos viendo en las siguientes secciones.

4.2.4 Entrenamiento del modelo de locutor

Para tener la posibilidad de verificar la identidad de un locutor se necesita tener un modelo matemático que, con la locución parametrizada, devuelva una puntuación que luego poder evaluar. Dichos modelos parten de un modelo genérico que puede ser el UBM, que debe ser entrenado para representar al locutor en particular que luego se utilizará para puntuar y verificar la identidad de dicho locutor. Para adaptar el modelo independiente de locutor (UBM) al modelo dependiente de locutor hay 2 técnicas utilizadas principalmente, Reestimación Baum-Welch y Adaptación (lo veremos a continuación).

Tradicionalmente se ha utilizado la Reestimación de Baum-Welch para entrenar los modelos de locutor a partir del modelo independiente de locutor, es un proceso basado en el algoritmo Expectation-Maximization. Recientemente, sin embargo se ha demostrado que funciona mejor la adaptación MLLR (Maximum Likelyhood Linear Regression) [22], consistente en maximizar la verosimilitud a través de técnicas de regresión lineal. Comparando los resultados para distintas cantidades de audio para entrenamiento funciona mejor MLLR que Baum-Welch especialmente para entrenamientos con poca cantidad de audio para entrenar.

4.2.5 Normalización de puntuaciones

Una vez hemos obtenido las puntuaciones como está explicado en el punto 4.2.3 todavía podemos ir más allá y mejorar los resultados finales si realizamos normalizaciones de los mismos, consiguiendo así que nuestros resultados sean más robustos.

Un caso de normalización un tanto especial es la denominada H-norm (Handset Normalization) [23], consistente en normalizar el resultado obtenido en la puntuación de la locución, restándole la media de las puntuaciones obtenidas para modelos del mismo locutor pero para diferentes canales y dividiendo el resultado entre la desviación típica que presentan dichos canales entre sí. De esta forma obtenemos una puntuación más robusta ante los resultados de diferentes canales, compensando así la variabilidad intercanal.

$$L_{H-norm}\left(X \mid \lambda, H\right) = \frac{L(X \mid \lambda) - \mu_H(\lambda)}{\sigma_H(\lambda)}$$

Otra normalización utilizada habitualmente es la denominada T-norm (Test Normalization) [24], que consiste en normalizar el resultado obtenido con un locutor con los resultados obtenidos por una cohorte de impostores. Se coge la puntuación obtenida con el modelo del locutor a verificar, se le resta la media de las puntuaciones obtenidas con los modelos de la cohorte de impostores y se divide el resultado entre la desviación típica que presentan las puntuaciones de la cohorte de impostores.

$$L_{T-norm}\left(X \mid \lambda, T\right) = \frac{L(X \mid \lambda) - \mu_T(X)}{\sigma_T(X)}$$

El problema que tiene esta normalización es que consume mucho tiempo de procesador, ya que requiere el reconocimiento de la locución por parte del modelo del locutor a verificar, además del reconocimiento por parte de todos los modelos de la cohorte de impostores.

Una última técnica utilizada para normalización de puntuaciones es Z-norm (Zero Normalization). Consiste en enfrentar el modelo del locutor a un conjunto de locuciones non-target. Con las puntuaciones obtenidas se calcula la media y la desviación típica. En la etapa de test, a la puntuación obtenida frente al archivo de test se le resta la media de las puntuaciones obtenidas de los enfrentamientos non-target, dividiendo el resultado por la desviación típica de dichas puntuaciones non-target.

$$L_{Z-norm}(X \mid \lambda, Z) = \frac{L(X \mid \lambda) - \mu_Z(Z)}{\sigma_Z(Z)}$$

Este tipo de normalización resulta más eficiente computacionalmente que T-norm en etapa de test, ya que el cálculo de la media y la varianza puede realizarse de forma previa.

Por otra parte hay otra forma de mejorar los resultados de las puntuaciones que es fusionar puntuaciones obtenidas para una misma locución con diferentes técnicas, de este modo las posibles carencias que pueda tener una técnica pueden ser subsanadas con otra y viceversa [25].

4.2.6 Adaptación de modelos de locutor

Una vez se ha entrenado un modelo de locutor a partir de las locuciones de entrenamiento se puede seguir mejorando la precisión del modelo con sucesivas locuciones, esto es lo que se llama adaptación (no confundir con adaptación MAP o MLLR). Hay 2 formas de realizar adaptación, supervisada y no supervisada:

- La adaptación supervisada consiste en que, después de tener el modelo entrenado con locuciones de entrenamiento, con cada nueva locución del locutor se sigue entrenando el modelo. Esto puede hacerse por ejemplo si se tienen locuciones de test de un locutor cuya identidad conocemos a ciencia cierta. Este conocimiento puede tenerse porque haya alguien supervisando la prueba de test, porque posea o sepa algo identificativo o porque haya pasado previamente otro tipo de prueba de reconocimiento biométrico.

- En la adaptación no supervisada sin embargo no se sabe a priori la identidad del locutor, por ello se tiene que establecer algún tipo de criterio para decidir si se sigue entrenando el modelo del locutor o no con la locución de test entrante. En la mayor parte de experimentos de este tipo dicho criterio consiste en la evaluación del fichero de test y en función de la puntuación obtenida considerar si la locución es o no del locutor [26]. En el caso de que la puntuación supere lo establecido en el criterio de aceptación, es decir que consideramos que la locución de test es del locutor pretendido, entonces utilizamos dicha locución para entrenar y seguir mejorando el modelo del locutor.

4.3 Limitaciones de los sistemas de reconocimiento de locutor dependiente de texto

Como en todas las tecnologías, el ámbito de reconocimiento de locutor dependiente de texto se enfrenta a limitaciones y dificultades derivadas de la naturaleza de la tarea en sí. El ejemplo más claro es la poca cantidad de locuciones de entrenamiento que podemos pedir a un usuario sin que éste se sienta molesto o decida abandonar. Además de este ejemplo tenemos otros como la elección del conjunto de frases a utilizar tanto para entrenamiento como para test, la variabilidad inter-canal o la elección del umbral de aceptación para las locuciones de test y adaptación.

4.3.1 Limitaciones debidas a la tecnología

Entre las principales limitaciones de los sistemas de reconocimiento de locutor dependiente de texto se encuentra la cantidad de datos de entrenamiento de que disponemos. En una aplicación realista no podemos permitirnos tener al usuario mucho tiempo grabando locuciones de entrenamiento porque se convierte en una tarea aburrida para el usuario y corremos el riesgo de que decida abandonar la tarea. Debido a que la sesión de entrenamiento tiene que ser corta nos vemos obligados a seleccionar cautelosamente las locuciones que queremos grabar en dicha sesión.

Hay varios factores que entran en juego en la selección del léxico a utilizar para entrenamiento. Tal vez el más importante de ellos sea la relación que tendrá más tarde con el léxico utilizado en test. Un problema propio de este tipo de sistemas es que debido a la poca cantidad de frases de entrenamiento disponemos también de poca cantidad de léxico entrenado, con lo cual para tener efectividad en el reconocimiento no podremos desviarnos mucho del léxico de entrenamiento en el léxico de test. De hecho en muchos casos el léxico de entrenamiento y test es coincidente porque las frases también lo son, lo cual mejora notablemente los resultados. Sin embargo llevar este caso al extremo y realizar los test siempre con la misma locución puede dar muy buenos resultados pero también hacer que el sistema sea muy vulnerable a grabaciones.

Otro problema técnico se nos presenta cuando realizamos la sesión de entrenamiento en un canal y luego realizamos sesiones de test en otro canal. A este caso de le denomina intento de canal cruzado y ocurre en el 25-50% de las llamadas en los sistemas a través de canal telefónico. Según [27], el efecto que esto tiene llega a duplicar o incluso cuadruplicar la EER.

También hay que tener en cuenta que conforme pasa el tiempo el modelo de locutor se puede quedar obsoleto debido a varios factores. Uno de ellos es el envejecimiento del propio locutor, ya que a lo largo de largos periodos de tiempo y debido a cambios morfológicos en nuestra cavidad bucal nuestra voz cambia, claro que este es un problema menor en periodos de tiempo cortos o medios. Otro factor que influye en el envejecimiento del modelo es la evolución de las tecnologías de telecomunicación. Por ejemplo en una verificación de locutor a través del móvil el sistema no funcionará igual si hemos grabado las sesiones de entrenamiento con una tecnología (por ejemplo GSM) y realizamos el reconocimiento con llamadas a través de un terminal de la misma tecnología que si realizamos el reconocimiento con una llamada a través de un sistema más moderno (por ejemplo UMTS).

Por último hay que notar que la actitud del locutor a la hora de realizar el entrenamiento no tiene porqué ser la misma que cuando pronuncia una frase para reconocimiento. Probablemente durante el entrenamiento el locutor pronuncie claramente y en voz alta la locución. Sin embargo la pronunciación que utilice para luego reconocimiento puede ser mucho más natural y por lo tanto menos nítida, por ello el modelo de locutor no funciona todo lo bien que funcionaría si la actitud fuera la misma.

4.3.2 Limitaciones de los sistemas comerciales

En los sistemas experimentales podemos permitirnos cierto tipo de licencias a la hora del diseño, ya que éste está orientado a evaluar cambios y mejoras tecnológicas. Sin embargo cuando entramos en un ámbito más comercial hay que cambiar un poco la forma de actuar debido a que el sistema tiene que funcionar correctamente ante posibles ataques de impostores (por ejemplo mediante grabación) y además tiene que tener una interfaz cómoda para el usuario, sin largas sesiones de entrenamiento.

Por este tipo de motivos es muy importante diseñar buenos sistemas de reconocimiento a nivel algorítmico, pero también a nivel de gramáticas y léxico. En [28] nos muestran los principales problemas que ocasiona un mal diseño del léxico a utilizar en entrenamiento y test. El primero de ellos es que para que el sistema sea cómodo tiene que tener locuciones cortas tanto para train como para test. Por otra parte intentar personalizar el léxico para locutores en concreto puede complicar la tarea. Por ejemplo si el sistema de autenticación consiste en decir nombre y apellido del locutor y hay muchos locutores la tarea de buscar la coincidencia entre lo que el locutor ha dicho y los posibles nombres de la lista puede ser muy compleja. Además hay que diseñar un sistema que contemple errores por parte del locutor a la hora de introducir su locución y sepa resolverlos y prevenirlos con una correcta interfaz sencilla de utilizar. Por último y como ya hemos comentado antes hay que tener en cuenta que si el léxico utilizado en entrenamiento difiere del de test los resultados serán notablemente peores que si son parecidos o iguales.

Otro problema que puede surgir al intentar evitar problemas con diferencias de léxico es que el sistema sea fácilmente accesible mediante una grabación. Para evitar este tipo de problema una posible solución es que las frases de test sean aleatorias, lo cual genera un problema de diferencia de léxico. Para solucionar este problema de diferencia de léxico podemos utilizar un léxico basado en dígitos La pega es que este léxico basado en dígitos será muy pobre porque tendrá poca variedad de fonemas, con lo que será difícil de ampliar posteriormente, sin embargo puede dar buenos resultados dentro de esta limitación. Una posible solución en sistemas por canal telefónico para evitar el fraude por grabación es crear una lista de números seguros de cada impostor y sólo aceptar locuciones de alguno de dichos números. Otra solución es que el usuario tenga que hacer uso de algún tipo de conocimiento secreto para acceder al sistema de verificación (p.e. marcar un código secreto en el teléfono desde el que llama).

Un problema que se plantea es qué frase o qué tipo de frase pedimos para verificar al locutor. Parece que lo más lógico sería una frase que no fuera siempre la misma, a ser posible aleatoria, pero para que esto funcione bien su contenido léxico tiene que estar previamente entrenado, para ello el léxico de entrenamiento tiene que ser flexible, por eso se suelen utilizar de forma tan frecuente los léxicos basados en dígitos.

Otro de los problemas que se plantea consiste en establecer el punto de operación del sistema, es decir la puntuación a partir de la cual el locutor se considera válido. Esta decisión dependerá de las características del entorno donde se vaya a utilizar el sistema, del nivel de seguridad requerido y del tipo de léxico que utilicemos en test y entrenamiento. En [29] intentan establecer un umbral de forma automática en función de como vaya creciendo el sistema. Tengamos en cuenta que en un sistema comercial, en la entrega no tiene ningún locutor entrenado, por ello no podemos utilizar el mismo umbral que si tuviéramos a todos lo usuarios entrenados, éste tiene que ir adaptándose. No obstante es conveniente realizar pruebas antes con bases de datos para darle al cliente curvas y especificaciones de funcionamiento. Otra manera de afrontar este problema de establecer el umbral es calibrar el sistema para que devuelva tasas de verosimilitud (Likelihood Ratio, LR) calibradas sobre las que se pueden establecer los umbrales de forma automática a partir de las probabilidades a priori y de los costes asociados a la aplicación. Sin embargo, la calibración no se ha usado todavía extensamente en reconocimiento de locutor dependiente de texto. Si lo ha sido en reconocimiento de locutor independiente de texto.

Por último, el problema que presentan los sistemas comerciales es que no son fáciles de actualizar. No se guardan las locuciones de entrenamiento de los locutores, sino los modelos generados, por lo tanto si se encuentra una mejora en el algoritmo de entrenamiento y se quiere actualizar el sistema hay que volver a repetir el entrenamiento. Un ejemplo de esto es que si en un principio entrenamos con nombre y apellido ya no podremos cambiarlo más tarde sin volver a entrenar el sistema desde el principio.

4.4 Algunos resultados interesantes

A continuación presentamos algunos resultados interesantes que han sido obtenidos en distintos grupos de investigación de todo el mundo. Estos resultados confirman las limitaciones y retos de los sistemas de reconocimiento de locutor dependiente de texto y abren posibilidades sobre las futuras líneas de investigación en el tema.

4.4.1 Extracción de características

Algunos de los resultados que veremos a continuación son para reconocimiento de locutor independiente de texto, pero deberían servirnos para hacernos una idea de qué método funcionaría mejor en reconocimiento de locutor dependiente de texto.

En [30] realizan un estudio sobre el impacto de los codecs en la precisión de los sistemas de reconocimiento. En este estudio transmitían una misma serie de locuciones por distintos estándares (GSM, G.729 y G723.1) y luego con los resultados ponía a prueba el sistema de reconocimiento de locutor dependiente de texto. Llegaron a la conclusión de que cuanto mayor es la tasa de bits transmitida mayor es la precisión del sistema. La otra conclusión del estudio es que con las nuevas tecnologías la SNR detectada disminuye conforme pasa el tiempo. Así, de forma global podemos decir que cada nuevo periodo de tiempo en la tecnología de telecomunicaciones hace que la SNR suba y consecuentemente los sistemas de reconocimiento funcionen mejor.

Hasta ahora las técnicas más utilizadas en extracción de características para reconocimiento de locutor han sido curiosamente MFCC y LPCC. La curiosidad viene de que en principio estas técnicas fueron diseñadas para reconocedores fonéticos en vez de reconocimiento de locutor y sin embargo han dado muy buenos resultados en reconocimiento de locutor. Sin embargo también se han utilizado otras tecnologías, en [31] extraen los parámetros mediante un sistema de redes neuronales y consiguen una mejora relativa del 28%, eso sí el sistema es independiente de texto. En [32] realizan el análisis frecuencial con wavelets en vez de con las herramientas de Fourier habituales y consiguen una mejora relativa de entre 15 y 27%. Sin embargo pese a que estas técnicas han demostrado buenos resultados no han tenido aceptación entre la comunidad científica, debido probablemente a lo específico de su campo de actuación, recordemos que con MFCC y LPCC también podemos realizar reconocimiento de habla espontánea, por lo que son técnicas con mayor polivalencia en todo el área de análisis de voz.

Por otra parte debemos tener en cuenta la variabilidad de canal, no es lo mismo realizar una tarea de reconocimiento si todos los locutores utilizan el mismo canal que si no es así. La técnica con mejores resultados en este campo ha sido Channel Factors sobre los modelos de locutor GMM. Sin embargo en [33] han demostrado que aplicando esta misma técnica sobre los parámetros en lugar de sobre los modelos los resultados son similares. Esto es muy relevante, ya que implica que una compensación tan eficiente puede ser hecha sobre los parámetros y con ello podríamos utilizar cualquier técnica posteriormente para el entrenamiento, test y normalizaciones, sin necesidad de restringirnos a GMM. Para el caso que nos ocupa que es reconocimiento de locutor dependiente de texto esto es especialmente importante, ya que se suelen utilizar HMM en lugar de GMM. Hay otros estudios [34] que se centran en compensar la variabilidad inter-sesión, que dependerá, además de del canal, de otros factores propios del locutor. Para modelar esta variabilidad incluyen un término en el GMM que pretende modelar la sesión. De esta forma el GMM definitivo queda como un GMM del locutor al que se le suma una constante de baja dimensionalidad que representa la sesión.

De los estudios realizados para compensación de canal se derivan estudios que aplican dichas técnicas a la variabilidad inter-locutor. En [15] aplican Factor Análisis para estimar la variabilidad inter-locutor, consiguiendo una mejora relativa con respecto a su sistema base de entre el 10 y el 15%, de hecho muestran que esta mejora es comparable a la obtenida de la fusión de varios sistemas.

4.4.2 Impacto del léxico en la fiabilidad del sistema

Entramos en un terreno ampliamente estudiado en el área de reconocimiento de locutor, está demostrado [35] que la efectividad del sistema de reconocimiento de locutor depende en gran medida del léxico de entrenamiento (por la riqueza de éste) y sobre todo del parecido de éste con el léxico de test. En [35] nos ofrecen la siguiente tabla en la que podemos observar dicha dependencia. Podemos obtener mejoras relativas de hasta un 50% cuando, utilizando un sistema de reconocimiento con un léxico constituido por dígitos, el orden de los dígitos es el mismo.

Entrenamiento\Test	E	S	R	pR	N
E	6.16%	10.2%	13.2%	10.4%	36.2%
S	11.6%	5.05%	14.2%		39.3%
R	10.7%	9.43%	11.5%	10.0%	36.4%
pR	10.0%		11.3%	8.05%	35.6%
N	38.9%	39.7%	39.3%	39.1%	10.6%

Tabla 3: Resultados obtenidos como combinación de léxicos en entrenamiento y test, expresados en EER [35]

Donde:

E	Cuenta del 1 al 9
S	9 dígitos aleatorios (los mismos para entrenamiento que para test)
R	Secuencia de dígitos aleatoria
pR	Secuencia de dígitos pseudo-aleatoria
N	Nombre y apellido

Tabla 4: Significados de las abreviaturas de la Tabla 3

Mirando la Tabla 3 podemos sacar varias conclusiones, si nos fijamos en la diagonal principal vemos que la combinación que peor puntúa es la que utiliza tanto para entrenamiento como para test secuencias aleatorias de números, aún teniendo el mismo léxico en entrenamiento y test. La conclusión que se extrae de esto es que el orden de los dígitos tiene una gran influencia en la falta de precisión del sistema debida a falta de coincidencia en el contenido de la locución. Este fenómeno se puede volver a observar si nos fijamos en los resultados obtenidos entrenando con una secuencia en orden de números y reconociendo con una secuencia aleatoria o pseudo-aleatoria del mismo léxico, donde obtenemos mejores resultados es en la secuencia pseudo-aleatoria. Por otra parte podemos observar que de forma genérica el léxico que peor funciona es el consistente en nombre y apellido, esto es debido a que el tiempo empleado en decir este tipo de locución tiene una media de 0,98 segundos mientras que por ejemplo el tiempo medio de locución de una secuencia ordenada de dígitos tiene una media de 3,97 segundos.

Las conclusiones por tanto que se extraen de esta tabla es que conservar el orden de los dígitos aumenta la precisión del sistema y que cuanto más cortas son las locuciones, mayor es la influencia de que no haya coincidencia léxica.

En [36] llegan a la conclusión de que la falta de coincidencia en el léxico introduce un error similar a la falta de coincidencia de canal y bastante mayor al error introducido por SNR.

Tipo de desajuste	EER (%)
Sin desajuste	7.02
SNR	7.47
Canal	9.76
Léxico (2 dígitos en común)	8.23
Léxico (1 dígito en común)	13.4
Léxico (0 dígitos en común)	36.3

Tabla 5: EERs obtenidas con diferentes tipos de desajuste, extraido de [36]

Es por tanto un objetivo prioritario alcanzar sistemas robustos ante la falta de coincidencia del léxico. De hecho en [28] afirman que conforme avancemos en robustez frente a la falta de coincidencia léxica más se acercará el campo de reconocimiento de locutor dependiente de texto al de reconocimiento de locutor independiente de texto.

4.4.3 Diseño del modelo universal

El diseño del modelo acústico independiente de locutor juega un papel importante en la fiabilidad final del sistema de reconocimiento de locutor ya que es con las puntuaciones obtenidas con este modelo con las que se comparan las puntuaciones obtenidas con los modelos entrenados para cada locutor. Entrenar un UBM (Universal Background Model) con el mismo contenido léxico que se va a utilizar en entrenamiento de locutor hizo en [37] que el EER de un sistema de reconocimiento con léxico del tipo My voice is my password (MVIMP) descendiera de 16,3 a 11,8%.

Otra idea que ha demostrado buenos resultados consiste en personalizar un modelo universal para cada locutor. En [38] construyen un UBM para cada locutor como la mezcla de una selección de locutores impostores, siendo el criterio de selección el parecido de la sesión de entrenamiento de los locutores con la del locutor genuino. Los resultados fueron buenos pero el problema que tiene este tipo de entrenamiento para el UBM es que no es realista para una aplicación comercial, debido a la cantidad de tiempo y espacio en memoria que consumiría realizar este modelo UBM para el caso de que haya una cantidad importante de usuarios en el sistema.

Por último reseñar un artículo [39] en el que se construye el UBM como un HMM de menor complejidad que el utilizado para los locutores a partir de las locuciones de entrenamiento de todos los locutores para así tener un modelo más genérico, y si se quiere difuminado, de todas las locuciones de entrenamiento, teniendo luego el modelo del locutor como un HMM completo de mayor nivel de complejidad.

4.4.4 T-norm

T-norm mejora notablemente los resultados obtenidos a través de la normalización de las puntuaciones. Son varios los estudios que han indagado en las mejoras que puede aportar T-norm y cómo explotar éstas. En [40] se demuestra que T-norm es sensible a las frases que utilicemos para entrenar a los impostores de la cohorte y del parecido que tengan estas frases con las frases de entrenamiento de los locutores genuinos. El estudio se ha realizado con una base de datos descrita en [41] donde los locutores tienen etiqueta de género y canal y las cohortes de impostores son construidas en función del género y canal de la locución del locutor genuino.

Por otra parte hay un estudio muy interesante de T-norm que prueba a normalizar los mismos locutores con cohortes de impostores (de los mismos locutores) entrenadas de forma rica y pobre. Decimos que una cohorte de impostores está entrenada de forma pobre si dicho entrenamiento recoge poca variedad léxica (en el caso que nos ocupa esto serían locuciones del 1 al 9, ya que todas las locuciones tienen el mismo contenido) y decimos que un entrenamiento es léxicamente rico cuando recoge bastante variedad léxica (en el caso que nos ocupa consistiría en series de 9 dígitos sin orden). Según este estudio T-norm

realizado con cohortes entrenadas de forma rica da mejores resultados que las realizadas de forma pobre. Incluso se puede apreciar que en T-norm realizado con cohortes de impostores entrenadas de forma pobre el resultado puede empeorar.

Tipo Experimento	Sin T-Norm	T-norm pobre	T-norm rico
E (entrenamiento y test)	17.10%	14.96%	14.74%
S (entrenamiento y test)	14.44%	16.39%	10.42%

Tabla 6: Tasas de FR para FA=1%

Donde:

E	Cuenta del 1 al 9
S	9 dígitos aleatorios (los mismos para entrenamiento que para test)

Tabla 7: Significados de las abreviaturas de la Tabla 6

Tal vez otra manera de plantearse la normalización sea hacer dicha normalización previa a la obtención de puntuaciones como es el caso de [42]. En este trabajo normalizan el alineamiento temporal de la frase. Concretamente se realiza un alineamiento temporal con el modelo independiente de locutor que será usado posteriormente tanto por el modelo del locutor como por el UBM.

4.4.5 Entrenamiento

En toda tarea de reconocimiento de patrones (y por supuesto en reconocimiento de locutor) probablemente la parte más importante es el entrenamiento de los modelos. En reconocimiento de locutor siempre se ha tenido el problema de que para tener bien entrenado un modelo es necesario tener muchos datos para entrenarlo, es decir mucha cantidad de voz del locutor a entrenar. En reconocimiento de locutor independiente de texto este problema es especialmente grave, ya que las técnicas utilizadas para modelado acústico del locutor (GMM), pese a tener un funcionamiento excepcional, necesitan mucha cantidad de habla del locutor para dar buenos resultados. La ventaja que presentan los HMMs en reconocimiento de locutor dependiente de texto es que no necesitan tanta cantidad de datos para entrenar, ya que se entrenan modelos de fonemas, que son muy específicos, y además se le dice en todo momento que fonema se tiene que reconocer. De este modo se puede decir que el reconocimiento de locutor dependiente de texto necesita menos cantidad de audio del locutor para entrenar, no por la disciplina en sí, sino por la tecnología que utiliza y los datos que se le proporcionan.

Para realizar el entrenamiento de locutores en reconocimiento de locutor dependiente de texto se utilizan principalmente tres técnicas, reestimación Baum-Welch, adaptación MLLR y adaptación MAP. La reestimación de Baum-Welch se trata de un caso particular del algoritmo Expectation-Maximization, que busca maximizar la verosimilitud aplicado a los HMMs, el problema es que únicamente garantiza alcanzar un máximo local en una superficie de optimización extraordinariamente compleja. La adaptación MLLR demuestra funcionar mejor que la Reestimación de Baum-Welch [22].

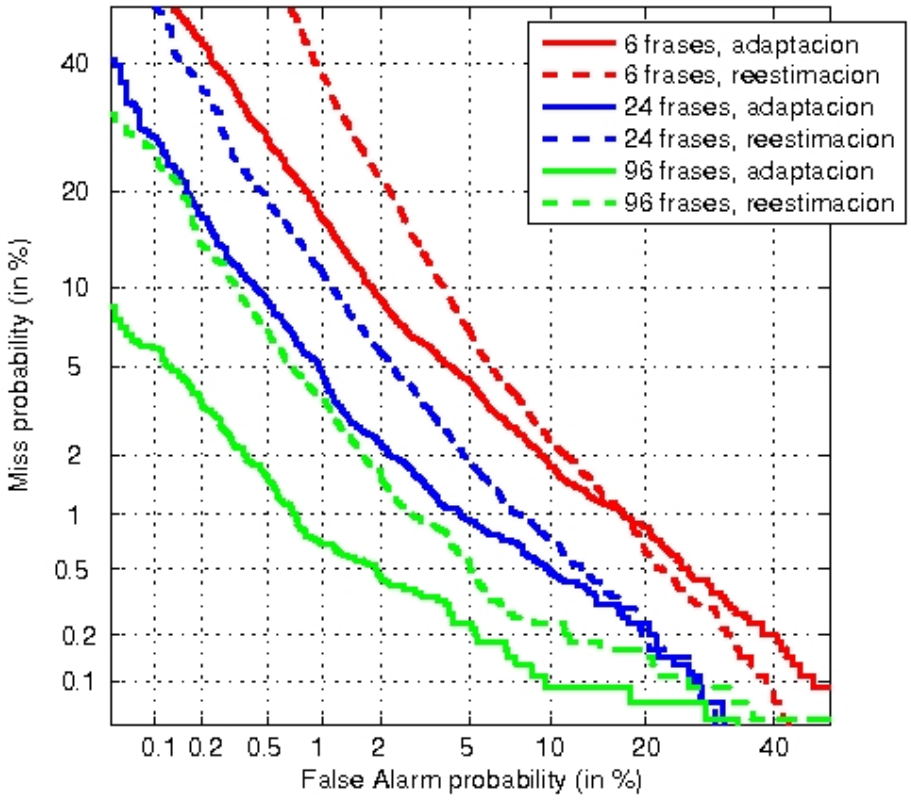


Figura 13: Comparación entre reestimación Baum-Welch y adaptación MLLR para diferentes cantidades de frases de entrenamiento, extraido de [22]

La adaptación MLLR funciona bien comparada con otras técnicas cuando tenemos pocos datos de entrenamiento, cuando disfrutamos de cantidades abultadas de datos para entrenamiento la opción que normalmente se escoge es MAP, menos efectivo para pocos datos de entrenamiento.

4.4.6 Adaptación de los modelos de locutor

Como hemos comentado antes, debido a la poca cantidad de datos de entrenamientos de que disponemos en las aplicaciones comerciales de reconocimiento de locutor dependiente de texto, la adaptación se presenta como la gran oportunidad para obtener mejores resultados. Fundamentalmente hay 2 tipos de adaptación, supervisada y no supervisada:

- En la adaptación supervisada el sistema tiene conocimiento de la identidad del locutor que realiza la identificación. Puede ser mediante otro tipo de verificación biométrica o por la posesión de algún tipo de dispositivo físico (tarjeta) o conocimiento (palabra clave o pin). Lo bueno que tiene la adaptación supervisada es que en cada iteración de adaptación introducimos información nueva nunca vista antes, en forma de variabilidad temporal o bien en forma de nuevos fonemas o palabras. Vale para entornos experimentales, para ver como evoluciona el sistema con adaptación, también vale para sistemas en los que haya dos factores de identificación (huella dactilar, firma…), siendo uno de ellos ajeno a la locución

utilizada para adaptar el modelo acústico del locutor.

- En la adaptación no supervisada [43] el sistema no tiene conocimiento de la identidad del locutor, por lo tanto deberá utilizar algún tipo de criterio para establecer la validez del archivo para adaptar el modelo acústico del locutor. En este tipo de adaptación podemos encontrarnos con 2 tipos de errores fundamentalmente en función de lo exigente que sea el sistema para aceptar las locuciones entrantes para adaptar el modelo acústico del locutor. Si el nivel de exigencia es muy bajo, es posible que el sistema acepte como válidas locuciones pertenecientes a impostores, con lo que la adaptación que se haría no haría que el sistema cometiera menos errores sino todo lo contrario. Por otra parte si el umbral exigido por el sistema para considerar la locución válida para adaptar es muy alto, hay poco riesgo de que el sistema acepte a impostores, pero las locuciones del locutor genuino admitidas no tendrán información adicional para el sistema, con lo cual no introducimos variabilidad y la información introducida no tendrá nada de nueva.

Hay experimentos para text-independent [17] en los que se ha permitido que las medias, varianzas y pesos de las mezclas puedan ser variables en función de la etapa de adaptación en la que nos encontremos. Esta técnica se llama Variable Rate Smoothing (VRS). Se puede comprobar que a partir de cierto número de etapas de adaptación esta técnica ya no mejora de forma considerable los resultados. Esto es debido a que VRS está pensado para mejorar los resultados ante ausencia de datos conocidos, es decir, con poca cantidad de habla. Conforme van sucediendo etapas de adaptación tenemos más datos conocidos, lo que hace que esta técnica vaya dejando de tener utilidad de forma progresiva, entre otras cosas por su propia naturaleza, que consiste en dar diferentes pesos en función de la etapa de adaptación en la que nos encontremos. Pese a todo en las primeras etapas de adaptación la mejora conseguida con VRS es importante, por ello deberíamos considerar la posibilidad de utilización en dependiente de texto [44], porque disponemos de pocos datos, de forma similar a las primeras etapas de adaptación.

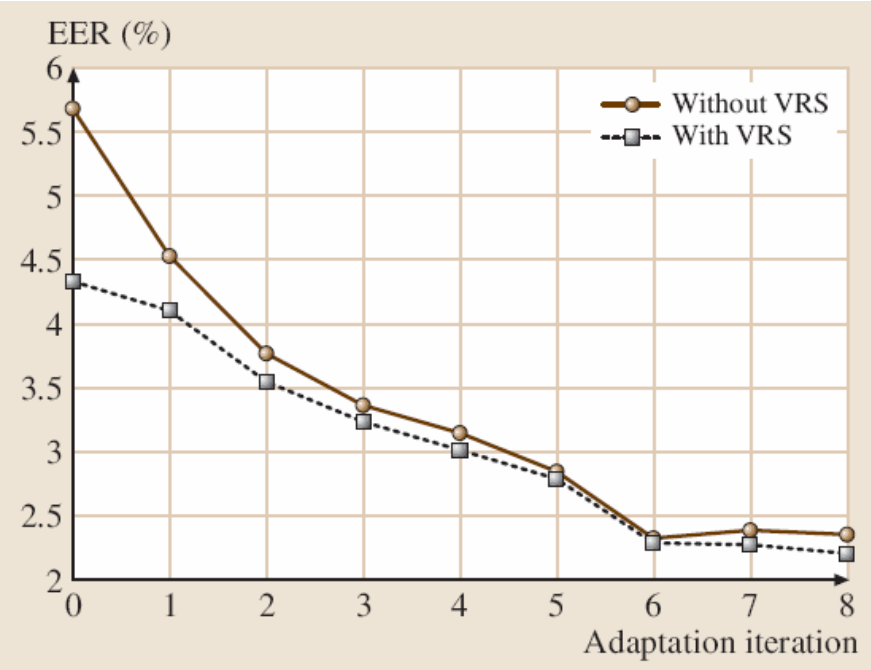


Figura 14: Comparación de EERs obtenidas con y sin VRS, extraído de [45]

En reconocimiento de locutor dependiente de texto se utilizan mucho los léxicos basados en dígitos. Esto es por una sencilla razón, los dígitos conforman un reducido conjunto de palabras y pueden combinarse en infinidad de formas distintas resultando locuciones naturales que no suenen extrañas. Consecuentemente se han realizado multitud de experimentos con locuciones de dígitos. Este tipo de locuciones permiten cierto grado de libertad a la hora de experimentar. Se ha demostrado que dependiendo del tipo de locuciones que usemos para entrenar tendremos resultados mejores o peores en test. La gráfica que a continuación se presenta muestra dos experimentos en los que hemos realizado el reconocimiento con secuencias pseudo-aleatorias de números, la diferencia está en que en uno de los experimentos se ha realizado el entrenamiento con secuencias pseudo-aleatorias de números y en otro experimento el entrenamiento se ha realizado con secuencias de números del 1 al 9 en orden. Podemos ver que si entrenamos con secuencias de dígitos pseudo-aleatorias el reconocimiento es mejor en las primeras iteraciones debido a que el sistema está entrenado con secuencias más parecidas a las de prueba que en el otro caso. Sin embargo conforme van transcurriendo iteraciones de adaptación se observa que ambas curvas convergen. Este fenómeno es debido a que las locuciones de adaptación aportan la nueva información, antes no conocida, con las nuevas locuciones, y esta nueva información hace que al cabo de unas pocas iteraciones ambos sistemas estén entrenados con el mismo léxico. En esta gráfica podemos ver el alto potencial que tiene la adaptación en el reconocimiento de locutor dependiente de texto.

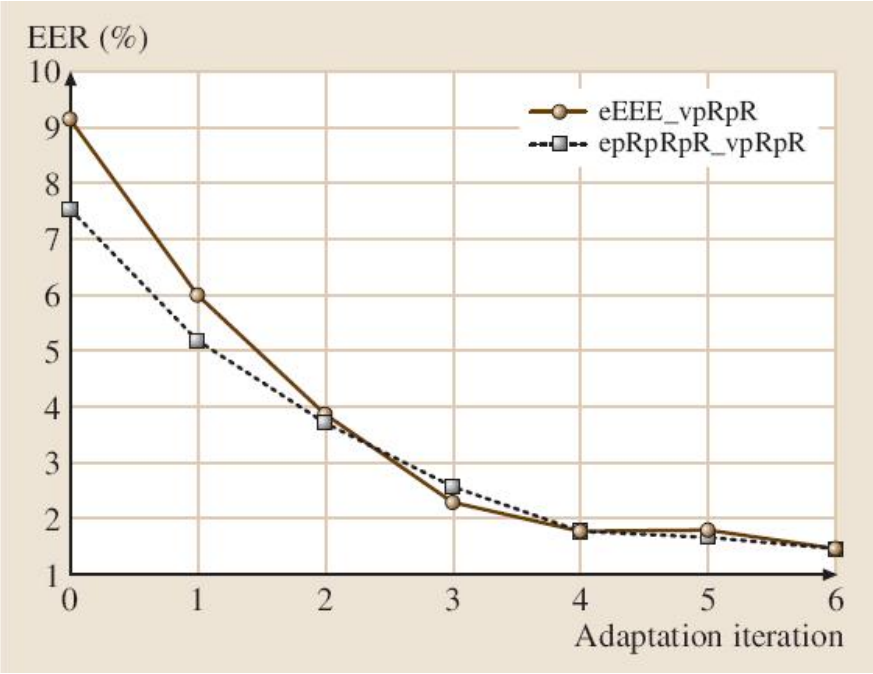


Figura 15: Comparación de EERs obtenidas para diferentes tipos de léxico empleados en entrenamiento y verificación, en función de la etapa de adaptación, extraído de [45]

Otro de los fenómenos observados en adaptación es que aunque la EER disminuya de forma global, cuando fijamos un punto de operación (por ejemplo fijamos una tasa de Falsa Aceptación) la tasa de Falso Rechazo aumenta conforme vamos iterando en adaptación. En [46] y [47] resuelven este problema con el algoritmo FCDT (Frame Count Dependent Thresholding), el cual pone el umbral de adaptación en función de la cantidad de voz que contenga la locución, es decir pondera menos una locución con, por ejemplo, 3

segundos de habla neta que una locución que contenga 10 segundos de habla neta. De esta manera se evita el efecto anteriormente comentado, manteniendo constante el Falso Rechazo para una tasa de Falsa Aceptación dada. En [46] se realizó un experimento con una FA objetivo de 0.525% y tras adaptar con FCDT se obtuvo una FA media de 0.855% con una varianza de 0.671%.

Otro problema, planteado en [48], es que cada locutor tiene sus propias características y debería tener su propio umbral para adaptar, pero esto no es viable en un sistema comercial, de modo que hay que buscar un umbral óptimo para el sistema para todos los locutores todas las veces. De nuevo, para resolver este problema la calibración del sistema resultaría muy útil.

Un problema que nos encontramos en dependiente de texto, que no existe en independiente de texto es que conforme adaptamos los modelos, los impostores cada vez puntúan más alto [26]. Esto parece ser debido a la influencia del léxico restringido, que hace que el sistema tienda a comportarse como un reconocedor fonético. Por esto, parece que lo más apropiado es combinar la adaptación con FCDT la T-norm.

4.4.7 Protección contra grabaciones

Si ya de por si es difícil diseñar un sistema de reconocimiento de locutor dependiente de texto con un buen compromiso entre FA y FR en el punto de trabajo deseado, además debemos tener en cuenta lo robusto que pueda ser nuestro sistema frente a ataques, no de impostores, sino de grabaciones de la voz del propio locutor genuino. Por una parte si la grabación no es muy buena tampoco debería suponer demasiado problema por el hecho de que la variabilidad de canal se encargaría de descartar la grabación, pero si la calidad de la grabación es similar a la del entrenamiento entonces nos encontramos ante un problema. De todas formas hay que tener en cuenta que hay que realizar un cierto esfuerzo tecnológico para poder llevar a cabo esta falsificación con éxito. Además la violación del sistema mediante grabaciones solo será efectiva en sistemas que pidan siempre la misma locución, y esto es fácil de solucionar implementando un sistema text-prompted (en el cual se muestra al locutor el texto que debe pronunciar) que pida una clave distinta (preferiblemente aleatoria) cada vez.

Hay estudios que se han hecho con el propósito de poner a prueba sistemas text-pompted [49]. En estos estudios, con una pequeña cantidad de voz disponible se realiza un modelo del locutor para posteriormente sintetizar el texto que se desee en voz mediante un sistema de conversión de texto en habla (TTS, Text To Speech). La locución de impostor funciona bien y es capaz de romper la seguridad de un sistema, eso si, siempre que haya realizado el mismo modelado que el usado para el sistema de impostor, lo cual parece poco realista poder conocer e implementar. Otro método [50] consiste en hacer transformaciones de voz, pero este sistema de asalto requiere todavía más información sobre el sistema que se ataca, además de que es una técnica compleja, por lo que por el momento no parece realista considerarla como un posible método de asalto al sistema.

4.4.8 Generación de impostores

Una vez hemos diseñado un sistema de reconocimiento de locutor dependiente de texto tendremos que evaluarlo. Si queremos que el sistema vaya incluido en algún tipo de aplicación comercial deberemos proporcionar con el mayor detalle posible las especificaciones de FA, FR y EER junto con la curva DET. Hasta aquí no debería existir ningún tipo de inconveniente. El problema viene cuando nos piden por ejemplo el FR para una FA de 1%±0.3%. Para poder dar ese dato con semejante precisión necesitaríamos más de 3000 intentos de acceso al sistema por parte de impostores. Por parte de locutores genuinos, si éstos utilizan regularmente el sistema, no debería suponer un inconveniente tan grave como para los impostores generar tantos intentos de acceso.

La solución a este problema pasa por tener claves de acceso iguales para varios locutores, para así poder utilizarlas para impostar a otros mediante el algoritmo Round-Robin. El problema es que si no lo hacemos así la diferencia léxica será muy alta y la FA obtenida estará por debajo de la FA real. En [40] resuelven este problema bajando al fonema como unidad fundamental, de modo que crean intentos de acceso de impostor a partir de fonemas de otros locutores, ya que es más fácil conseguir encontrar un fonema en un impostor que una palabra completa.

5 HMM

5.1 Introducción

En este capítulo se explicarán los conceptos matemáticos asociados al modelado del locutor. Para realizar el reconocimiento del locutor se necesita una herramienta que modele cada uno de los fonemas empleados por el locutor, esta herramienta son los Modelos Ocultos de Harkov (en inglés Hidden Markov Models o HMMs). Los HMMs constituyen la técnica de modelado de voz más empleada en reconocimiento de habla y reconocimiento de locutor dependiente de texto, desde que en la década de los 80 sustituyeron a la técnica de Alineamiento Temporal Dinámico (en inglés Dynamic Time Warping o DTW).

Un HMM es una máquina de estados finita, en la que las observaciones son una función probabilística del estado. Esto quiere decir que el modelo es un proceso doblemente estocástico formado por un proceso estocástico oculto no observable directamente, que corresponde a las transiciones entre estados, además de un proceso estocástico observable cuya salida es la secuencia de vectores espectrales. Las observaciones son los vectores de parámetros acústicos y los estados suelen modelar eventos sonoros de menor duración que un fonema. Las densidades de probabilidad de cada estado, las probabilidades de transición y la secuencia de observaciones son parámetros conocidos del sistema, mientras que la secuencia de estados que el modelo de Markov ha seguido para generar esa secuencia de observaciones permanece desconocida para el usuario. Para reconocimiento de locutor dependiente de texto, dado que se proporciona una trascripción, se sabe que las transiciones solo se podrán realizar al mismo estado o a estados siguientes. Esto quiere decir que una vez se ha salido de un estado no se puede volver a él. A este tipo de HMMs se les llama modelos de Bakis o de izquierda a derecha.

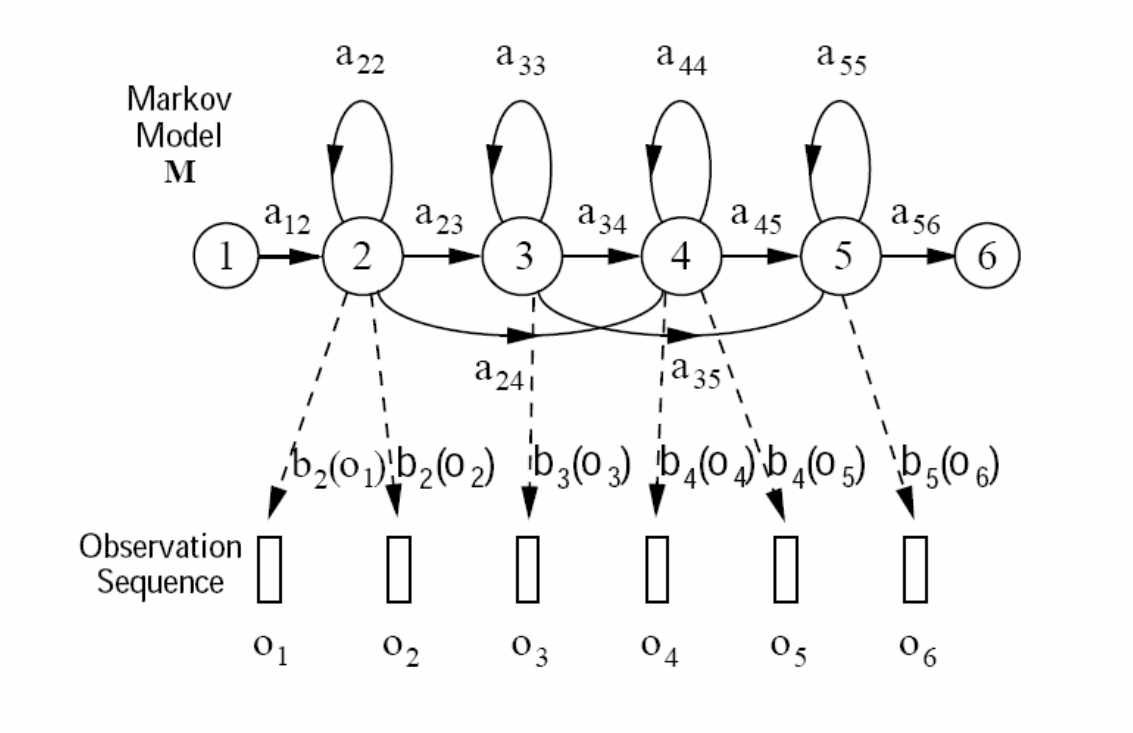


Figura 16: El modelo de generación de Markov

En la Figura 16 se pueden observar los elementos que definen un HMM:

- El número de estados del modelo N, donde q_t denota el estado en el instante de tiempo t. Los HMMs que vamos a utilizar están compuestos por 3 estados. $S=\{s_1,s_2,...,s_N\}$.
- La dimensión del conjunto de observaciones distintas de salida M, es decir el tamaño del alfabeto $V=\{v_1,v_2,...,v_M\}$.
- La distribución de probabilidad de transición entre estados $A=\{a_{ij}\}$: $aij=p(q_t=s_j|q_{t-1}=s_i)$ $1\leq i,j\leq N$.
- La distribución de probabilidades de emisión de símbolos entre estados $B=\{b_j(k)\}$: $b_j(O_k)=p(O_k|q_t=s_j)$ $1\leq j\leq N$, $1\leq k\leq M$, donde O_k es un símbolo perteneciente a V.
- Distribución del estado inicial $\pi=\{\pi_i\}$: $\pi_i=p(q_0=s_i)$ $1\leq i\leq N$.

Con todo esto, un HMM se describe como $\lambda=\{A,B,\pi\}$.

5.2 GMM

Cada estado del HMM es, a su vez un Modelo de Mezclas de Gaussianas (en inglés Gaussian Mixture Model o GMM). Los GMM son modelos estadísticos, que explotan las características espectrales de la voz para discriminar a los locutores.

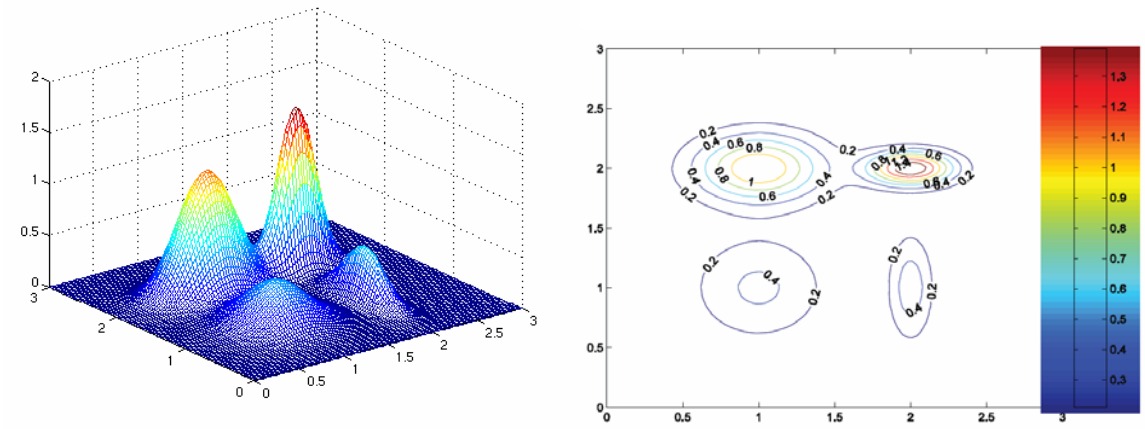


Figura 17: GMM bidimensional de 4 Gaussianas

Un GMM modela la distribución de probabilidad de las observaciones de un fragmento de audio de un determinado locutor a partir de un modelo de suma de G Gaussianas. Cada una de las Gaussianas se caracteriza por su peso w_i, su vector de medias μ_i y su matriz de covarianzas Σ_i, así $\theta=\{w_i,\mu_i,\Sigma_i\}$ es la definición del modelo donde, $i=1\ldots G$ y G es el número de mezclas.

Ante una observación desconocida x, el modelo GMM asigna una puntuación relacionada con su verosimilitud. Es decir, una puntuación relacionada con la probabilidad de que el locutor del modelo haya generado la observación.

$$p(x \mid H_0) = \sum_{i=1}^{G} w_i \cdot p_i(x)$$

Como se puede observar analizando la fórmula, se trata de la suma ponderada de las *G* densidades componentes. La verosimilitud sobre cada Gaussiana viene dada a su vez por:

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

El teorema de Bayes demuestra que la decisión óptima no viene dada únicamente por la probabilidad de que el locutor haya generado la observación $p(x|H_0)$, sino por un cociente de probabilidades:

$$\frac{p(x \mid H_0)}{p(x \mid H_1)} \begin{array}{l} \geq umbral\ aceptar\ H_0 \\ < umbral\ rechazar\ H_0 \end{array}$$

Donde $p(x|H_1)$ es la probabilidad de que el locutor no haya generado la observación y sin embargo, haya sido cualquier otro locutor.

Para estimar $p(x|H_1)$ se hace uso de los que se conoce como modelos UBM (Universal Background Model). Un UBM es un modelo GMM estándar pero que ha sido entrenado a partir de observaciones de todos los locutores (o un conjunto representativo de los mismos). Dicho de otra forma un modelo UBM en el ámbito de los HMMs se corresponde con un estado de un HMM del modelo independiente del locutor.

Los UBM estiman la densidad de probabilidad de las observaciones, sobre todo el conjunto del habla humana. Por tanto, la verosimilitud frente al UBM mide la probabilidad de que la observación haya podido ser generada por una persona cualquiera.

5.3 Problemas planteados para HMM

Hay 3 problemas que es necesario resolver para que los HMMs tengan utilidad en aplicaciones reales:

1. Problema de evaluación de la probabilidad
2. Problema de encontrar la secuencia de estados óptima
3. El problema de entrenamiento de un modelo

5.3.1 Problema 1: Problema de evaluación de la probabilidad

Dada una secuencia de observación $O=\{O_1,O_2,...,O_T\}$ y un modelo $\lambda=\{A,B,\pi\}$, ¿cómo se calcula $p(O \mid \lambda)$, la probabilidad de la secuencia de observación? Si es posible calcular esta probabilidad, entonces se podría calcular para todos los modelos y escoger aquel para el cual la probabilidad sea mayor.

La manera más directa de solucionarlo sería enumerando todas las posibles secuencias de estados de longitud T que generen la secuencia de observación O y sumando sus probabilidades según el teorema de la Probabilidad Total:

$$p(O \mid \lambda) = \sum_{Q} p(O \mid Q, \lambda) \cdot p(Q \mid \lambda) \tag{1}$$

Para ello se considera una determinada secuencia de estados: $Q=\{q_1,q_2,...,q_T\}$ donde q_1 es el estado inicial. La probabilidad de la secuencia de observación O dada la secuencia de estados Q es:

$$p(O \mid Q, \lambda) = \prod_{t=1}^{T} p(O_t \mid q_t, \lambda)$$

Donde se asume independencia estadística de las observaciones. Por lo tanto se obtiene:

$$p(O \mid Q, \lambda) = b_{q1}(O_1) \cdot b_{q2}(O_2) \cdots b_{qT}(O_T) .$$

Por otra parte la probabilidad de la secuencia de estados Q se puede expresar como:

$$p(Q \mid \lambda) = \pi_{q1} \cdot a_{q1q2} \cdot a_{q2q3} \cdots a_{qT-1qT}$$

Esto se interpreta como la probabilidad del estado inicial, multiplicada por las probabilidades de transición de un estado a otro. Sustituyendo los dos términos anteriores en la Ecuación 1 se obtiene la probabilidad de la secuencia de observación:

$$p(O \mid \lambda) = \sum_{Q} p(O \mid Q, \lambda) \cdot p(Q \mid \lambda) = \sum_{q1,q2,...,qT} \pi_{q1} \cdot b_{q1}(O_1) \cdot a_{q1q2} \cdot b_{q2}(O_2) \cdots a_{qT-1qT} \cdot b_{qT}(O_T)$$

El resultado se puede interpretar como que inicialmente, en el tiempo t=1 nos encontramos en el estado q_1 con probabilidad π_{q1} y generamos el símbolo O_1 con probabilidad $b_{q1}(O_1)$.Al avanzar el reloj al instante t=2 se produce una transición al estado q_2 con probabilidad $a_{q1,q2}$ y generamos el símbolo O_2 con probabilidad $b_{q2}(O_2)$. Este proceso se repite hasta que se produce la última transición del estado q_{T-1} al estado q_T con probabilidad $a_{qT-1,qT}$ y generamos el símbolo O_T con probabilidad $b_{qT}(O_T)$.

Llegados a este punto se puede ver que el coste computacional es muy alto, $2T \cdot N^T$ operaciones, lo que implica un orden de $O(N^T)$. Afortunadamente existe una manera más eficiente de llegar al mismo resultado. La clave está en guardar los resultados intermedios y utilizarlos para los posteriores cálculos de la secuencia de estados. A este algoritmo se le denomina el Algoritmo de Avance (En inglés Forward). El primer paso es definir la variable hacia delante como $\alpha_t(i)=p(O_1,O_2,...,O_t,\ q_t=S_i|\lambda)$. Esta variable corresponde con la probabilidad de que el modelo λ se encuentre en el estado i habiendo generado la secuencia parcial $O_1,O_2,...,O_t$ hasta el instante de tiempo t. $\alpha_t(i)$ se puede calcular por inducción siguiendo los siguientes pasos:

Inicialización:

$$\alpha_1 = \pi_i \cdot b_i(O_1),\ 1 \le i \le N$$

En este paso se inicializan las probabilidades hacia delante como la probabilidad conjunta del estado S_i y la observación inicial O_1.

Inducción:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) \cdot a_{ij} \right] \cdot b_j(O_{t+1}) \text{ , } 1 \le t \le T-1 \text{ , } 1 \le j \le N$$

La expresión entre corchetes representa la probabilidad de alcanzar el estado S_j en el instante de tiempo t+1 partiendo de todos los estados posibles S_i en el instante t habiendo observado hasta el instante t la secuencia parcial $O_1,O_2,...,O_t$. Si multiplicamos ahora dicho término por la probabilidad de observar O_{t+1} se obtiene $\alpha_{t+1}(j)$.

Finalización:

$$p(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

El cálculo de $p(O|\lambda)$ final se realiza sumando todas las variables hacia delante $\alpha_T(i)$ en el instante final T. Esto es así ya que por definición $\alpha_T(i)$ es igual a la probabilidad conjunta de haber observado la secuencia $O_1,O_2,...,O_T$ y encontrarnos en el estado Si: $\alpha_T(i)=P(O_1,O_2,...,O_T, q_T = S_i \mid \lambda)$, con lo que si sumamos dicha probabilidad para todos los estados posibles obtenemos la probabilidad esperada $p(O|\lambda)$. La complejidad de este algoritmo comparado con la manera directa de calcular $p(O|\lambda)$ es mucho menor y se encuentra en el orden de $O(N^2 \cdot T)$, con lo que el ahorro computacional es claro.

5.3.2 Problema 2: Problema de encontrar la secuencia de estados óptima

Decodificar un HMM consiste en encontrar la secuencia de estados óptima, dada una secuencia de observación. La resolución de este problema resulta muy importante para tareas de segmentación y reconocimiento de voz. A diferencia del problema 1 para el que se puede dar una solución exacta, existen diferentes maneras de resolver este problema. La razón es que la definición de secuencia óptima no es única, sino que existen varios criterios de optimización.

El criterio más extendido es el que utiliza el algoritmo de Viterbi, que trata de encontrar la mejor secuencia de estados, es decir, maximizar la probabilidad $p(q|O,\lambda)$ o lo que es equivalente, maximizar $p(O,q|\lambda)$. En la práctica este método también se puede utilizar para evaluar HMMs.

Para encontrar la mejor secuencia de estados $Q=\{q_1,q_2,...,q_T\}$ para una secuencia de observación dada $O=\{O_1,O_2,...O_T\}$ definimos la variable:

$$\delta_t(i) = \max_{q1,q2,...,qt-1} p\left[q_1\, q_2 ...\, q_t = i, O_1\, O_2\, ...\, O_t \mid \lambda \right]$$

Representa la secuencia de estados con mayor probabilidad en el instante t que acaba en el estado S_i y que ha generado las t primeras observaciones.

A continuación se sigue un proceso de inducción similar al algoritmo Avance-Retroceso (en inglés Forward-Backward), con la excepción de que en vez de tomar la suma de las probabilidades de los diferentes caminos que acaban en un mismo estado, el algoritmo de Viterbi selecciona y recuerda el mejor camino.

Inicialización:

$$\delta_1(i) = \pi_i \cdot b_i(O_1)\,,\ 1 \le i \le N$$
$$\phi_1(i) = 0$$

Inicialmente se define la probabilidad $\delta_1(i)$ como la probabilidad de encontrarse en el estado S_i en el instante t=1 multiplicada por la probabilidad de generar el símbolo O_1. El vector ϕ, en el que se va a almacenar el argumento que maximiza $\delta_t(j)$ para cada valor de t y de j, toma inicialmente el valor 0.

Recursión:

$$\delta_t(j) = \max_{1 \le i \le N}[\delta_{t-1}(i) \cdot a_{ij}]\, b_j(O_t)\,,\ 2 \le t \le T,\ 1 \le j \le N$$
$$\phi_t(j) = \arg\max_{1 \le i \le N}[\delta_{t-1}(i) \cdot a_{ij}]\,,\ 2 \le t \le T,\ 1 \le j \le N$$

Finalización:

$$p^* = \max_{1 \le i \le N}\left[\delta_T(i)\right]$$
$$q_T^* = \arg\max_{1 \le i \le N}[\delta_T(i)]$$

La iteración del punto 3 se termina cuando se han generado las T observaciones.

Backtracking:

$$q_t^* = \phi_{t+1}(q_{t+1}^*)\,,\ t = T-1,\ T-2,\ \ldots,\ 1$$

En este último paso se reconstruye la secuencia de estados partiendo desde el estado final hasta llegar al principio.

5.3.3 Problema 3: Entrenamiento de un modelo

El último y más complicado de los 3 problemas plantea cómo se deben ajustar los parámetros del modelo $\{A,B,\pi\}$ para maximizar la probabilidad de la secuencia de observación dado el modelo $p(O|\lambda)$. El principal inconveniente es que no existe ningún método analítico conocido que maximice el conjunto de parámetros a partir de los datos de entrenamiento. Se puede resolver, sin embargo, utilizando un procedimiento iterativo como el algoritmo de Baum-Welch, también conocido como el algoritmo de Avance-Retroceso. Este algoritmo usa los mismos principios que el algoritmo EM (Expectation-Maximization). El procedimiento consiste en actualizar los pesos de forma iterativa para poder explicar mejor las secuencias de entrenamiento observadas.

Antes de describir formalmente el algoritmo de Baum-Welch es necesario definir la probabilidad hacia atrás de manera similar a como se definió la probabilidad hacia delante $\beta_t(i)=p(O_{t+1},O_{t+2},...,O_T\ \ q_t=S_i|\lambda)$. $\beta_t(i)$ es en este caso la probabilidad de generar la

observación parcial $O=\{O_{t+1},O_{t+2},...,O_T\}$ desde el instante t+1 hasta el instante final T dado que el modelo se encuentra en el estado S_i en el instante de tiempo t. $\beta_t(i)$ se puede calcular por inducción como sigue.

Inicialización:

$$\beta_T(i)=1,\ 1 \le i \le N$$

Recursión:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} \cdot b_j(O_{t+1}) \cdot \beta_{t+1}(j)\ ,\ t=T-1,T-2,...,1,\ 1 \le i \le N$$

La relación entre α y β adyacentes se puede observar mejor en la siguiente figura. α se calcula recursivamente de izquierda a derecha mientras β se calcula recursivamente de derecha a izquierda.

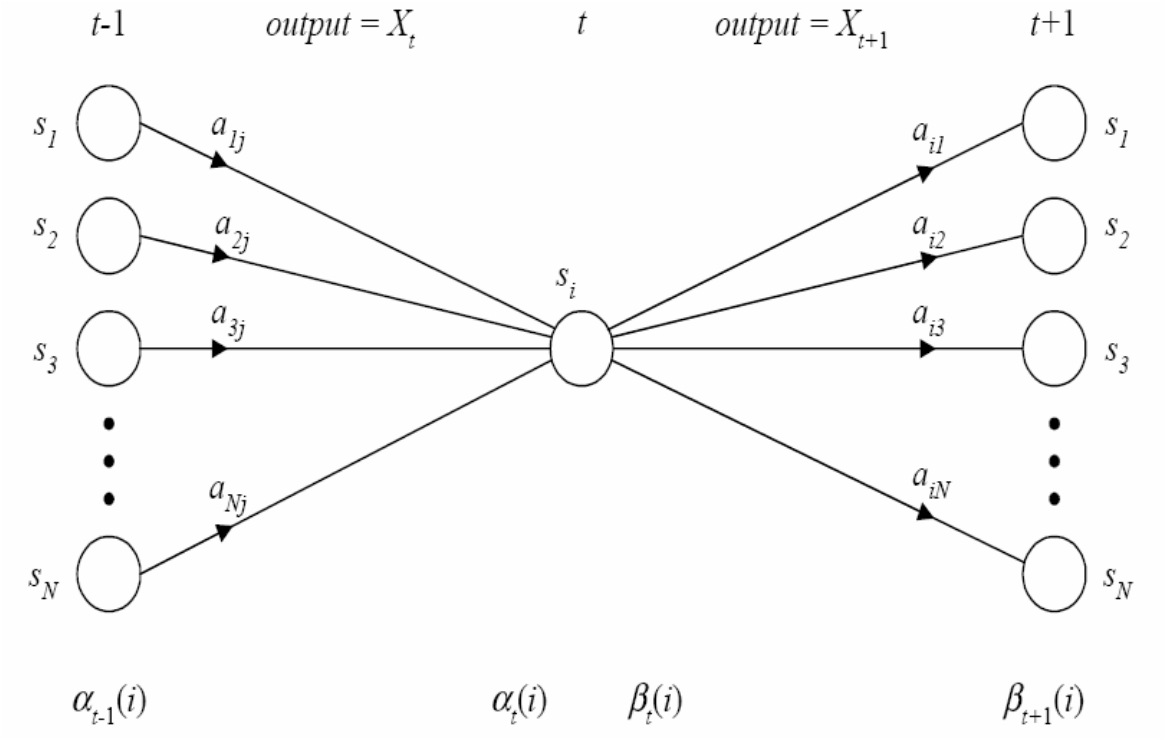


Figura 18: La relación entre α_{t-1} y α_t y β_{t-1} y β_t en el algoritmo Forward-Backward [51]

A continuación definimos la variable $\gamma_t(i,j)$, que representa la probabilidad de realizar una transición del estado S_i al estado S_j en el instante de tiempo t dado el modelo y dada la secuencia de observación, es decir:

$$\gamma_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda) = \frac{P(q_t = S_i, q_{t+1} = S_j, O \mid \lambda)}{P(O \mid \lambda)} = \frac{\alpha_t(i) \cdot a_{ij} \cdot b_j(O_{t+1}) \cdot \beta_{t+1}(j)}{\sum_{k=1}^{N} \alpha_T(k)}$$

Este resultado se puede ilustrar mejor con la siguiente figura:

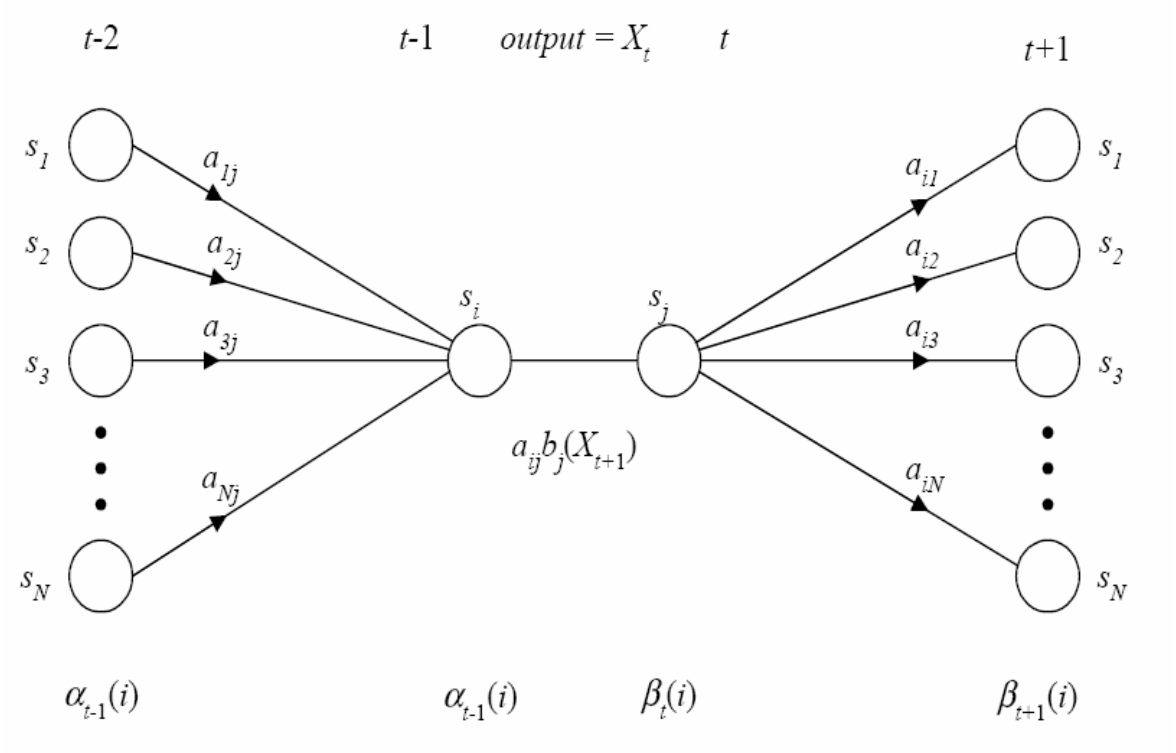


Figura 19: Ilustración de las operaciones necesarias para el cálculo de $\gamma_t(i,j)$, [51]

Es posible refinar iterativamente el vector de parámetros del HMM $\lambda=\{A,B,\pi\}$ si se maximiza la probabilidad de la observación, $p(O|\lambda)$, en cada iteración. Para ello denotamos como $\hat{\lambda}$ al nuevo vector de parámetros calculado a partir del vector de parámetros λ, obtenido en la iteración anterior. De acuerdo con el algoritmo EM, esto es equivalente a maximizar la siguiente función Q:

$$Q(\lambda,\hat{\lambda}) = \sum_{S1,S2,\ldots,SN} \frac{P(O,S\mid\lambda)}{P(O\mid\lambda)} \log P(O,S\mid\hat{\lambda})$$

Donde $p(O,S|\lambda)$ y $\log p(O,S|\hat{\lambda})$ se definen como sigue:

$$p(O,S\mid\lambda) = \prod_{t=1}^{T} a_{t-1t} b_t(O_t)$$

$$\log p(O,S\mid\hat{\lambda}) = \sum_{t=1}^{T} \log a_{t-1t} + \sum_{t=1}^{T} \log b_t(O_t)$$

Por lo tanto la ecuación inicial se puede rescribir de la siguiente manera:

$$Q(\lambda,\hat{\lambda}) = Q_{ai}(\lambda,\hat{a}_i) + Q_{bj}(\lambda,\hat{b}_j)$$

Donde:

$$Q_{ai}(\lambda,\hat{a}_i)=\sum_i\sum_j\sum_t\frac{P(O,q_{t-1}=i,q_t=j\mid\lambda)}{P(O\mid\lambda)}\log\hat{a}_{ij}$$
$$Q_{bj}(\lambda,\hat{b}_j)=\sum_j\sum_k\sum_{t\in o_t=V_K}\frac{P(O,q_t=j\mid\lambda)}{P(O\mid\lambda)}\log\hat{b}_j(V_k) \quad (2)$$

Como hemos separado la función en tres términos independientes, se puede maximizar $Q(\lambda\mid\hat{\lambda})$ maximizando cada uno de los términos por separado, sujeto a las siguientes restricciones:

$$\sum_{j=1}^{N}a_{ij}=1\quad\forall\, i$$
$$\sum_{k=1}^{M}b_j(V_k)=1\quad\forall\, i$$

Además, los términos en las Ecuaciones 2 tienen todos la siguiente forma:

$$F(x)=\sum_i y_i\log x_i$$

Donde:

$$\sum_i x_i=1\,.$$

Haciendo uso de los multiplicadores de Lagrange, se demuestra que la función $F(x)$ toma su valor máximo en:

$$x_i=\frac{y_i}{\sum_i y_i}$$

A partir de todo esto, se obtienen las estimaciones de los parámetros del modelo HMM:

$$\hat{a}_{ij}=\frac{\frac{1}{P(O\mid\lambda)}\sum_{t=1}^{T}P(O,q_{t-1}=i,q_t=j\mid\lambda)}{\frac{1}{P(O\mid\lambda)}\sum_{t=1}^{T}P(O,q_{t-1}=i\mid\lambda)}=\frac{\sum_{t=1}^{T}\gamma_t(i,j)}{\sum_{t=1}^{T}\sum_{k=1}^{N}\gamma_t(i,k)}$$
$$\hat{b}_j(V_k)=\frac{\frac{1}{P(O\mid\lambda)}\sum_{t=1}^{T}P(O,q_t=j\mid\lambda)\cdot\delta(O_t,V_k)}{\frac{1}{P(O\mid\lambda)}\sum_{t=1}^{T}P(O,q_t=j\mid\lambda)}=\frac{\sum_{t\in O_t=V_k}\sum_i\gamma_t(i,j)}{\sum_{t=1}^{T}\sum_i\gamma_t(i,j)}$$

La probabilidad inicial $\hat{\pi}_i$ se puede derivar como un caso especial de la probabilidad de transición. Sin embargo, $\hat{\pi}_i$ se suele fijar para la mayoría de aplicaciones de voz, por ejemplo $\hat{\pi}_i=1$ para el estado inicial.

Al observar las ecuaciones anteriores, se puede ver que la primera corresponde con el cociente entre el número medio de transiciones del estado *i* al estado *j* y el número medio

de transiciones desde el estado i. La segunda ecuación se puede interpretar también como el cociente entre el número medio de veces que el símbolo V_k se emite desde el estado j y el número medio de veces que se emite un símbolo desde el estado j.

De acuerdo con el algoritmo EM, el algoritmo de reestimación de Baum-Welch garantiza una mejora monótona en la probabilidad en cada iteración hasta que ésta converge en un máximo local. El algoritmo se puede resumir en los siguientes pasos:

1. Inicialización: Se escoge una estimación inicial del modelo λ.
2. Paso E: Se calcula la función auxiliar $Q(\lambda, \hat{\lambda})$ a partir de λ.
3. Paso M: Se calcula $\hat{\lambda}$ de acuerdo con las ecuaciones de reestimación para maximizar la función auxiliar Q.
4. Iteración: λ pasa a tomar el valor de $\hat{\lambda}$ y se repite el algoritmo desde el paso 2 hasta que converge.

5.4 Adaptación MLLR (Maximum Likelihood Linear Regression)

La adaptación MLLR [The HTK Book, 2005] es una adaptación lineal, realiza una serie de transformaciones para reducir las diferencias entre el modelo independiente del locutor inicial y las locuciones de entrenamiento de cada locutor. Hay 2 formas de realizar adaptación MLLR, en la primera de ellas se adaptan las medias y varianzas de todas las Gaussianas del estado que adaptamos de forma global, se denomina MLLR global:

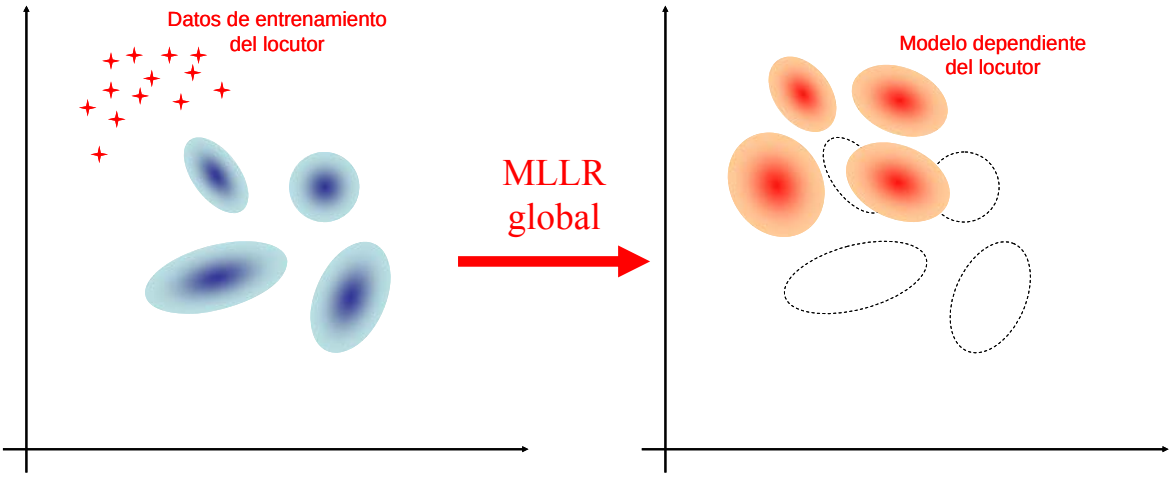


Figura 20: Comportamiento de las Gaussianas en la adaptación MLLR global

Otra forma es dividir las Gaussianas del estado que adaptamos en clases de regresión, para posteriormente adaptar cada una de las clases de regresión:

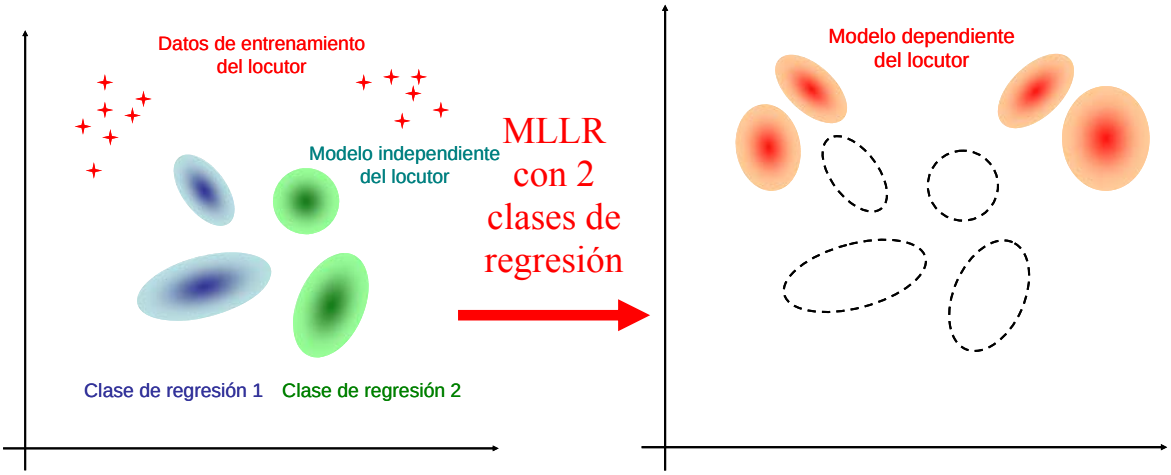


Figura 21: Comportamiento de las Gaussianas en la adaptación MLLR con 2 clases de regresión

5.4.1 Transformación global

El efecto de estas transformaciones es desplazar las medias y modificar las varianzas del sistema inicial, de manera que crezca la probabilidad de que cada estado del HMM inicial genere las locuciones. Las matrices de transformación se obtienen mediante la técnica EM (Expectation-Maximization). El nuevo vector de medias viene dado por:

$$\hat{\mu} = W\xi$$

Donde W es la matriz de transformación de dimensiones $n_x(n+1)$ (n es la dimensión de los datos) y ξ es el vector de medias extendido $\xi = [1, \mu_1, \mu_2, \ldots, \mu_n]^T$. Por lo tanto, W se puede descomponer en $W=[b,A]$, siendo A una matriz de transformación $n_x n$ y b un vector de bias.

Existen dos formas de realizar la adaptación de las varianzas. La primera es:

$$\hat{\Sigma}_m = B_m^T H_m B_m$$

Donde H_m es la transformación lineal a estimar y B_m es la inversa del factor de Choleski de Σ_m^{-1}, de manera que:

$$\Sigma_m^{-1} = C_m C_m^T$$

$$B_m = C_m^{-1}$$

Esta forma de transformación resulta en una matriz de covarianzas completa efectiva, siempre que la matriz de transformación H_m esté completa a su vez. Esto hace que el cálculo de puntuaciones sea altamente ineficiente.

Con la segunda y más eficiente manera de realizar la transformación de la matriz de covarianzas ésta se obtiene de la siguiente manera:

$$\hat{\Sigma} = H\Sigma H$$

Donde H es la matriz de transformación de covarianzas $n_x n$. Este tipo de transformación se puede implementar de manera eficiente como una transformación de las medias y del vector de características:

$$N(o;\mu,H\Sigma H) = \frac{1}{|H|^2} N(H^{-1}o;H^{-1}\mu,\Sigma) = |A|^2 N(Ao;A\mu,\Sigma)$$

Siendo $A = H^{-1}$. Utilizando esta forma es posible estimar y aplicar transformaciones completas eficientemente.

5.4.2 Árbol de clases de regresión

Con el fin de aumentar la flexibilidad del proceso de adaptación es posible determinar un conjunto apropiado de clases principales dependiendo de la cantidad de datos de adaptación que tengamos. Si sólo disponemos de una pequeña cantidad de datos, se generaría entonces únicamente una transformación global. Ésta se aplica a todas las Gaussianas que componen el modelo. Sin embargo, según aumenta la cantidad de datos se puede mejorar la adaptación incrementando el número de transformaciones a realizar. Cada una de estas transformaciones es más específica y se aplica a un determinado agrupamiento de Gaussianas. Por ejemplo, las Gaussianas se podrían agrupar según las clases de fonemas: silencio, vocales, nasales, fricativas, etc. Los datos de adaptación se utilizarían ahora para construir transformaciones más específicas para aplicarlas a esos grupos.

MLLR hace uso de un árbol de clases de regresión para agrupar las Gaussianas en el modelo, de manera que el conjunto de transformaciones a estimar se puede elegir de manera dinámica de acuerdo con la cantidad de datos de adaptación disponibles. Este árbol se construye para agrupar componentes que se encuentran próximas entre sí en el espacio acústico y se construye partiendo del modelo original independiente de locutor.

Los nodos terminales del árbol especifican las agrupaciones finales y se les denomina clases de regresión principales. Cada Gaussiana presente en el modelo pertenece a una de estas clases. La figura muestra un ejemplo de arbol de regresión con 4 clases terminales $\{C_4,C_5,C_6,C_7\}$.

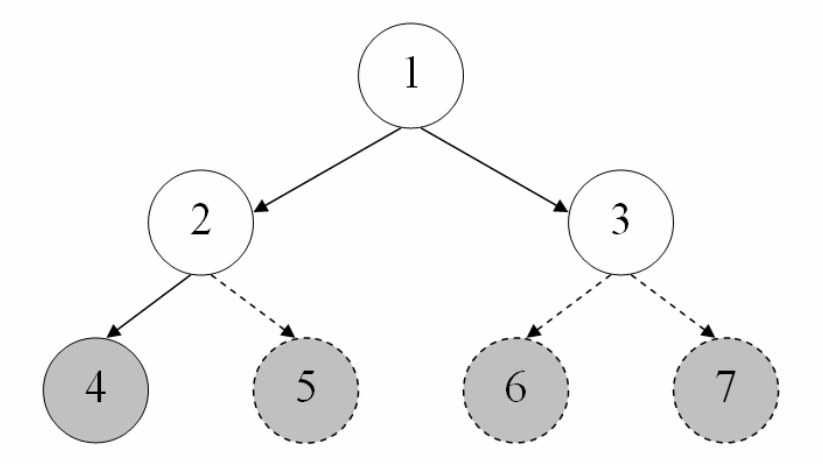


Figura 22: Ejemplo de árbol de regresión binario

Las flechas y nodos continuos indican que hay suficientes datos para que se genere una matriz de transformación utilizando los datos asociados a esa clase, mientras que las flechas y nodos discontinuos significan que no hay suficientes datos. Para determinar en que nodos se va a realizar la transformación, se recorre el árbol desde la raíz y se genera una transformación para aquellos nodos que cumplen que tienen datos suficientes y son nodos terminales o tienen algún hijo sin datos suficientes. En el ejemplo anterior se generan entonces transformaciones para los nodos 2, 3 y 4 que llamamos W_2, W_3, W_4. Por lo tanto, a las componentes Gaussianas de cada clase principal de regresión se le aplican las matrices de transformación (medias y varianzas) de la siguiente manera:

$$\begin{Bmatrix} W_2 & \rightarrow & \{C_5\} \\ W_3 & \rightarrow & \{C_6, C_7\} \\ W_4 & \rightarrow & \{C_4\} \end{Bmatrix}$$

Es importante destacar por último que, una adaptación global corresponde al caso en que se tiene un árbol sólo con el nodo raíz.

5.5 Adaptación de modelos MAP (Maximum a Posteriori)

La adaptación de modelos ocultos de Markov también puede hacerse con adaptación MAP (Maximum A Posteriori). La adaptación MAP considera que el parámetro del modelo independiente del locutor es la información a priori sobre dicho parámetro. Con la voz de entrenamiento (con la información nueva observada) del locutor estima dicho parámetro. Finalmente combina ambos de acuerdo con 2 variables, un factor de adaptación que es necesario fijar para ponderar más o menos la información nueva y la cantidad total de información nueva empleada para estimar el parámetro. De forma gráfica la adaptación MAP adapta cada una de las Gaussianas de forma independiente.

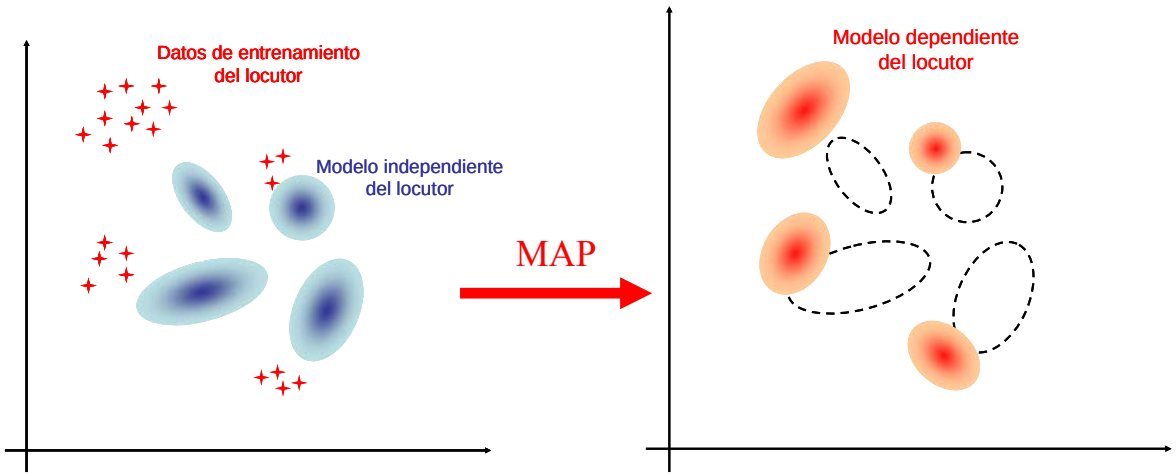


Figura 23: comportamiento de las Gaussianas en la adaptación MAP

La fórmula de actualización de las medias para el estado *j* y la mezcla *m* es:

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

Donde τ es el peso de la información a priori, N la probabilidad de ocupación de los datos de adaptación, μ_{jm} la media del modelo independiente de locutor y $\overline{\mu}_{jm}$ la media de los datos de adaptación. Como se puede ver por la fórmula, si la probabilidad de ocupación de una componente Gaussiana N_{jm} es pequeña, entonces la estimación de la media se mantendrá cercana a la media de la componente independiente de locutor.

Una desventaja de la adaptación MAP es que requiere más datos que la adaptación MLLR para ser efectiva. Esto se debe a que la adaptación MAP se realiza a nivel de las Gaussianas componentes del modelo. Sin embargo, cuando disponemos de cantidades mayores de datos de adaptación, la adaptación MAP empieza a funcionar mejor que MLLR, debido a esta actualización detallada de cada componente.

6 Bases de datos utilizadas

Para poder evaluar la calidad del funcionamiento del sistema de reconocimiento de locutor dependiente de texto desarrollado se necesitan bases de datos con las que obtener resultados. En este proyecto se han utilizado dos bases de datos para este propósito, la base de datos BioSec Baseline y la base de datos YOHO.

Una base de datos biométricos consiste en un conjunto de adquisiciones de rasgos físicos de un conjunto de personas. Los rasgos adquiridos pueden ser varios o uno sólo y pueden ser adquiridos de maneras diferentes o de una única forma, además los rasgos pueden haber sido adquiridos en días distintos e incluso en entornos diferentes. Todo esto sirve para evaluar el comportamiento de los sistemas de reconocimiento biométrico en situaciones diversas, intentando simular condiciones reales que puedan dar lugar a una merma de la efectividad del sistema de reconocimiento biométrico. De esta forma se puede estudiar el comportamiento del sistema ante tales situaciones adversas y de este modo investigar posibles soluciones para paliar los efectos que éstas producen sobre el sistema de reconocimiento biométrico.

Para estudiar el problema de una forma más amplia surgen las bases de datos multibiométricas que combinan varias fuentes de información biométrica. Pueden ser de varios tipos:

- Multimodal: una base de datos multimodal es aquella que combina varios rasgos del sujeto, por ejemplo iris y firma o voz, huella dactilar y geometría de la mano.
- Multisesión: este tipo de bases de datos tienen adquisiciones separadas en el tiempo del rasgo bajo estudio. Esto es muy útil para rasgos que presentan variabilidad temporal, como es el caso de la voz (que puede variar por el estado emocional del sujeto o por circunstancias como un catarro) o la huella dactilar (que puede variar en el tiempo debido a cortes o quemaduras). De hecho el estudio de la variabilidad intersesión es un área de investigación por si misma [2] muy importante.
- Múltiples sensores: también es útil saber como se va a comportar nuestro algoritmo de reconocimiento biométrico enfrentado a sensores distintos. Por ejemplo no es lo mismo realizar un reconocimiento con voz grabada en un estudio con un micrófono de alta calidad que con voz grabada por el micrófono integrado de una webcam.
- Múltiples codificaciones: aun habiendo sido adquirido el rasgo con sensores de un mismo tipo, la conversión de los datos de analógico a digital y el tratamiento posterior de los mismos puede haberse hecho de maneras muy distintas. En el caso de la voz, por poner un ejemplo, se pueden observar variaciones en el comportamiento de un sistema de reconocimiento si la voz ha sido transmitida a través de un móvil usando tecnología GSM o tecnología UMTS.
- Múltiples realizaciones de lo mismo: dentro de una misma sesión de adquisición de un rasgo dicho rasgo puede ser adquirido varias veces de la misma forma. Esto puede servir para tener más datos con los que entrenar los modelos del rasgo del sujeto bajo condiciones idénticas pero con una pequeña variabilidad natural que es imposible evitar. Por ejemplo nunca colocamos el

dedo sobre un sensor de huella dactilar de una forma exactamente igual al igual que tampoco lo hacemos al pronunciar una misma frase.

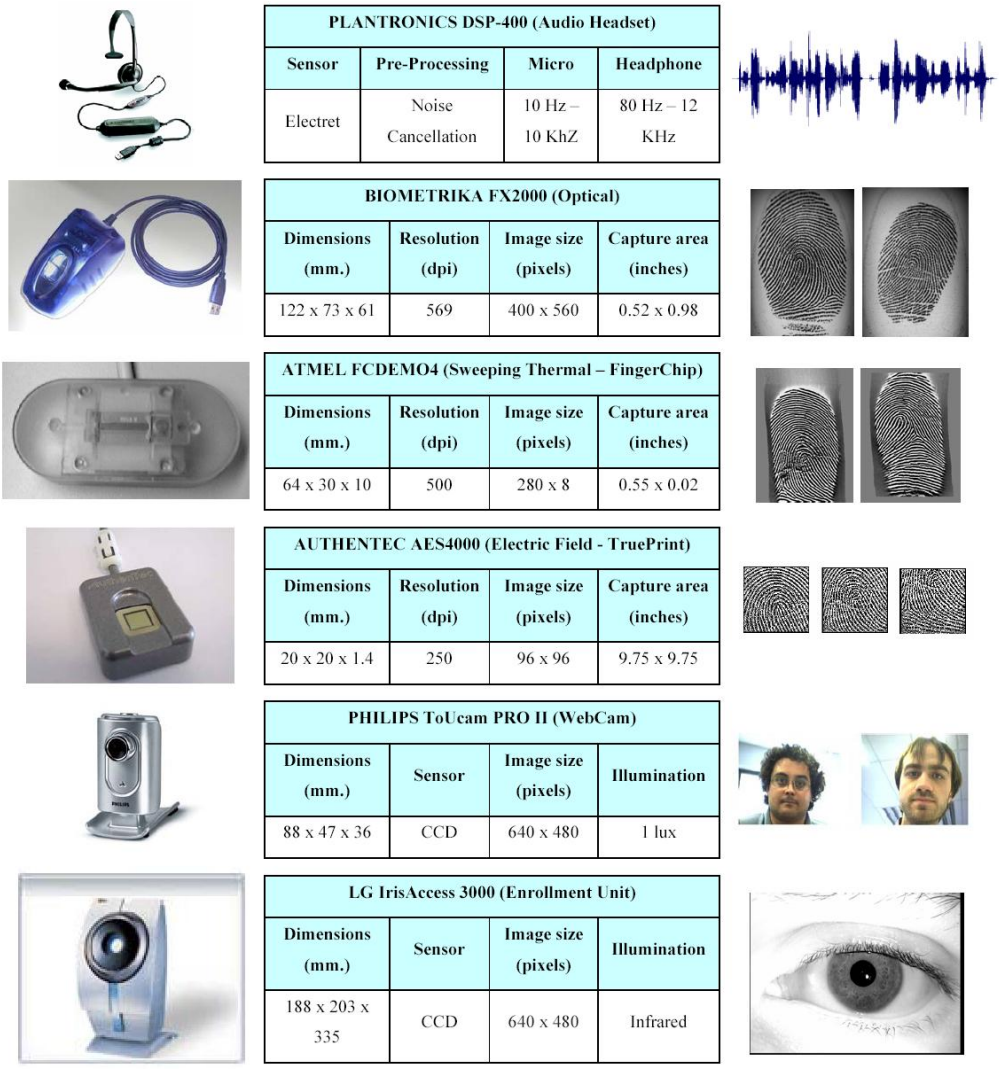
- Múltiples instancias: esto se refiere a la adquisición de distintas instancias de un mismo rasgo. Por ejemplo en huella se puede adquirir dedo índice y pulgar, en iris del ojo izquierdo y del derecho y en voz, hablando distintas frases o en distintos idiomas.
- Quiméricas: Las bases de datos quiméricas surgen de la necesidad de bases de datos multimodales. Las bases de datos quiméricas son bases de datos multimodales en las que cada individuo es un ser virtual compuesto por distintos rasgos de distintos sujetos (quimera). Por ejemplo un sujeto virtual puede estar compuesto por la voz de una persona, las huellas de otra y la firma de un tercero.

Dada la enorme cantidad de datos que nos podemos encontrar en una base de datos debemos encontrar una manera de ordenar los datos para ser utilizados por nuestro sistema de reconocimiento biométrico de forma coherente, es decir debemos encontrar un protocolo de evaluación. En el protocolo de evaluación definimos qué instancias de qué rasgos de qué individuos utilizamos para entrenar los modelos, realizar la evaluación, construir el modelo universal o formar la cohorte de normalización.

La particularidad que tiene que cumplir una base de datos destinada a reconocimiento de locutor dependiente de texto es que cada archivo de audio debe ir acompañado de un archivo con la trascripción textual del habla pronunciada en el audio.

6.1 BioSec Baseline

En esta base de datos [52] se encuentran los rasgos biométricos de 150 personas. Se trata de una base de datos multimodal en la que están recogidas imágenes frontales de cara, imágenes de iris, huella dactilar y voz. Es a la vez una base de datos multisesión, ya que dichos rasgos fueron adquiridos en 2 sesiones separadas entre 1 y 4 semanas. Además se utilizaron 3 tipos de sensores diferentes para la adquisición de huellas dactilares y dos tipos de micrófonos para la adquisición de voz (micrófono lejano integrado en una cámara web y micrófono cercano integrado en unos auriculares). Por último las condiciones ambientales no son controladas (iluminación y ruido de ambiente) para conseguir un entorno real.



PLANTRONICS DSP-400 (Audio Headset)			
Sensor	**Pre-Processing**	**Micro**	**Headphone**
Electret	Noise Cancellation	10 Hz – 10 KhZ	80 Hz – 12 KHz

BIOMETRIKA FX2000 (Optical)			
Dimensions (mm.)	**Resolution (dpi)**	**Image size (pixels)**	**Capture area (inches)**
122 x 73 x 61	569	400 x 560	0.52 x 0.98

ATMEL FCDEMO4 (Sweeping Thermal – FingerChip)			
Dimensions (mm.)	**Resolution (dpi)**	**Image size (pixels)**	**Capture area (inches)**
64 x 30 x 10	500	280 x 8	0.55 x 0.02

AUTHENTEC AES4000 (Electric Field - TruePrint)			
Dimensions (mm.)	**Resolution (dpi)**	**Image size (pixels)**	**Capture area (inches)**
20 x 20 x 1.4	250	96 x 96	9.75 x 9.75

PHILIPS ToUcam PRO II (WebCam)			
Dimensions (mm.)	**Sensor**	**Image size (pixels)**	**Illumination**
88 x 47 x 36	CCD	640 x 480	1 lux

LG IrisAccess 3000 (Enrollment Unit)			
Dimensions (mm.)	**Sensor**	**Image size (pixels)**	**Illumination**
188 x 203 x 335	CCD	640 x 480	Infrared

Figura 24: Rasgos adquiridos y sensores empleados en la adquisición de la base de datos BioSec Baseline

Cada una de las frases que debía pronunciar cada sujeto fue grabada en idioma inglés y español (castellano), siendo el idioma nativo de la mayoría de los locutores el español. Las frases pronunciadas por cada locutor consisten en 8 dígitos pronunciados por separado seguidos uno detrás de otro, por ejemplo "4-4-4-7-3-5-7-8" debe ser pronunciado "cuatro, cuatro, cuatro, siete, tres, cinco, siete, ocho". Cada locutor tiene asignada una combinación de 8 dígitos como la descrita anteriormente que debe repetir 4 veces y otras 3 veces la combinación de dígitos de otro usuario para simular un asalto al sistema en el que un impostor conoce la clave del usuario. Estas 7 locuciones deben ser repetidas en cada sesión en ambos idiomas para cada uno de los micrófonos, resultando un total de 56 locuciones por usuario.

6.2 YOHO

YOHO es la base de datos más utilizada en reconocimiento de locutor dependiente de texto, es por ello por lo que se ha usado para gran parte de los experimentos. En

reconocimiento de locutor independiente de texto hay bases de datos para comparar resultados gracias a que todos los años el NIST (National Institute of Standards and Technology) organiza evaluaciones competitivas en esta materia. De esta forma los desarrolladores pueden comparar sus sistemas con los de otros laboratorios en igualdad de condiciones en el sentido de que a todos los laboratorios participantes se les entrega el mismo audio para evaluar su sistema. Sin embargo no existen competiciones de este tipo para reconocimiento de locutor dependiente de texto, de manera que no hay forma de comparar los sistemas de unos laboratorios con otros para esta disciplina. YOHO surgió como respuesta a una creciente demanda de una base de datos con la que poder comparar resultados entre laboratorios para reconocimiento de locutor dependiente de texto. En [53] Campbell describe un protocolo de utilización de la base de datos YOHO.

YOHO consta de 138 locutores, de ellos 106 son hombres y 32 son mujeres, todos ellos realizan las locuciones en idioma inglés americano el cual es idioma nativo de la mayoría de ellos (la mayor parte de ellos son procedentes de Nueva York). Las grabaciones están realizadas con un micrófono, en entorno de oficina y con muy poco ruido. Los datos de habla de cada locutor están divididos en entrenamiento y test. Los datos de entrenamiento están formados por 4 sesiones de 24 frases por sesión y los datos de test están divididos en 10 sesiones (separadas una media de 3 días) con 4 frases por sesión. Cada frase consiste en 3 pares de dígitos pronunciados de manera que el par de dígitos sea una única cantidad, por ejemplo “24-67-89” debe ser pronunciado “twenty four, sixty seven, eighty nine”.

6.3 TIMIT

La base de datos TIMIT contiene habla de 630 locutores (438 hombres y 192 mujeres) representantes de varios dialectos del inglés americano. La voz está grabada con micrófono. Cada locutor tiene grabadas 10 frases fonéticamente balanceadas.

El hecho de que las frases grabadas sean fonéticamente balanceadas significa que tienen una gran riqueza fonética, con ello conseguimos con unas pocas frases estén representados la mayor parte de los fonemas de una lengua. Este hecho, junto con la propiedad de que los locutores son hablantes de varios dialectos distintos del inglés americano, nos proporciona una amplia variedad de fonemas pronunciados de distintas formas, lo cual es idóneo para construir modelos fonéticos independientes del locutor.

6.4 ALBAYZIN

Esta base de datos está formada enteramente por locutores hablantes de idioma español. La base de datos está formada por 304 locutores, de los cuales la mitad son hombres y la otra mitad mujeres. Esta base de datos está dividida en 3 partes, también llamadas subcorpus. La primera de ellas es de carácter principalmente fonético, formado por frases fonéticamente balanceadas, teniendo un total de 200 frases distintas, de forma que generan un total de 6800 locuciones. La segunda parte de la base de datos es de aplicación y contiene 3900 locuciones correspondientes a una tarea de consulta a una base de datos. La tercera parte está formada por habla grabada en condiciones adversas, se compone de partes del corpus fonético y del corpus de aplicación grabadas bajo efecto Lombard.

7 Sistema desarrollado

A continuación se describen el sistema de partida así como las sucesivas evoluciones del mismo en el mismo orden en el que posteriormente serán presentados los resultados, para así poder observar la mejora introducida por cada una de las modificaciones.

7.1 Sistema de partida

Partimos de un sistema de reconocimiento de locutor dependiente de texto basado en HMMs fonéticos (tipo Left-to-Right), donde cada fonema se modela con 3 estados. El modelo independiente de locutor se genera con la base de datos TIMIT para el idioma inglés y con la base de datos ALBAYZIN para idioma castellano. El esquema general de funcionamiento es el siguiente:

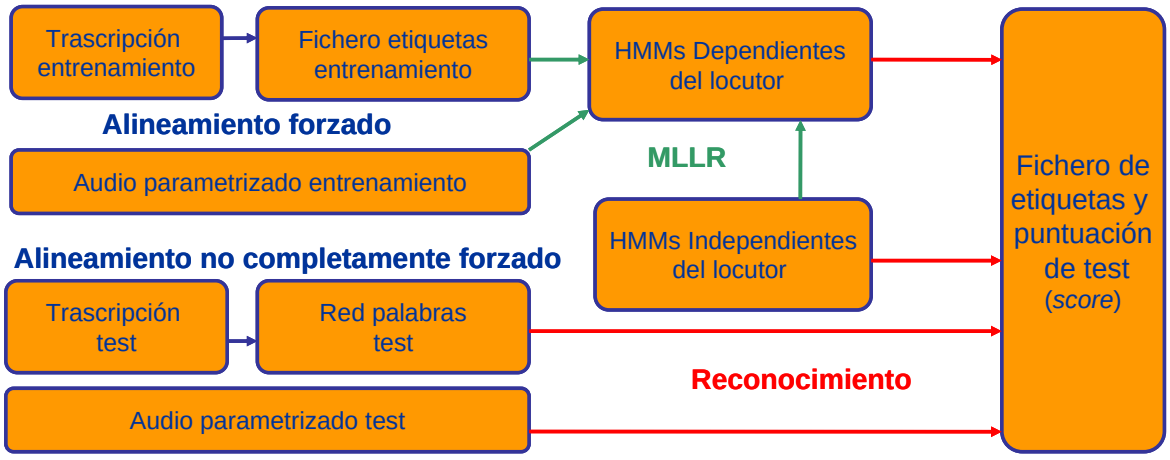


Figura 25: Esquema del sistema de partida

El primer paso, como en todo sistema de reconocimiento de voz es realizar la parametrización del audio, para ello se ha empleado una ventana temporal de 25ms con un solape entre ventanas de 10ms. Cada una de las ventanas de audio se pasa por un banco de filtros de Mel, extrayendo del resultado 13 coeficientes cepstrales (13 MFCC), además se extrae la diferencia entre la ventana anterior y posterior, dando lugar a otros 13 coeficientes (Δ) y por último se vuelve a extraer la diferencia entre los coeficientes Δ de la ventana anterior y posterior, dando lugar a otros 13 coeficientes ($\Delta\Delta$). El resultado final presenta un total de 39 coeficientes por ventana ($13 MFCCs + \Delta + \Delta\Delta$) a los que se les aplica el algoritmo CMN (en inglés Cepstral Mean Normalization) [Furui 81] consistente en restarle a cada vector el vector media de la locución para así realizar una pequeña compensación de canal.

Por otra parte, dado que se trata de reconocimiento de locutor dependiente de texto necesitamos tener la trascripción textual de lo que el locutor pronuncia o debería pronunciar en el archivo de audio (ya parametrizado). Con la trascripción y un diccionario fonético se realiza una descomposición de las palabras en fonemas. De esta forma obtenemos un fichero de etiquetas del audio de entrenamiento que nos indica el alineamiento temporal de los fonemas.

Posteriormente se toma el fichero de etiquetas del audio de entrenamiento y el propio audio de entrenamiento parametrizado y se adaptan al locutor los modelos fonéticos independientes del locutor mediante el algoritmo MLLR. Para la adaptación los modelos fonéticos al locutor primero se realiza una adaptación MLLR global, resultando de ésta un conjunto de matrices de transformación global, de modo que modificando el modelo independiente del locutor con estas matrices de transformación obtenemos el modelo adaptado al locutor. Posteriormente, una vez tenemos la transformación global hecha, se procede a dividir el conjunto de Gaussianas en un número concreto de conjuntos de las mismas o clases de regresión, para posteriormente adaptar mediante MLLR cada una de esas clases de regresión. Como resultado de esta segunda adaptación volvemos a obtener una serie de matrices de transformación, que aplicada al modelo adaptado con las matrices de transformación global nos dará el modelo final adaptado al locutor.

El siguiente paso que realiza el sistema de partida es preparar los archivos de audio destinados a la etapa de test, parametrizándolos de igual forma que los archivos de audio de entrenamiento. Además realiza un tratamiento distinto de las transcripciones consistente en elaborar, a partir de las trascripciones, una red de palabras donde se reflejan todas las posibles transiciones forzadas u opcionales entre palabras. Esto se hizo para poder contemplar la posibilidad de silencios opcionales entre palabras. Por ejemplo, para la base de datos YOHO una red de palabras con silencios opcionales entre pares de dígitos tendría la siguiente estructura:

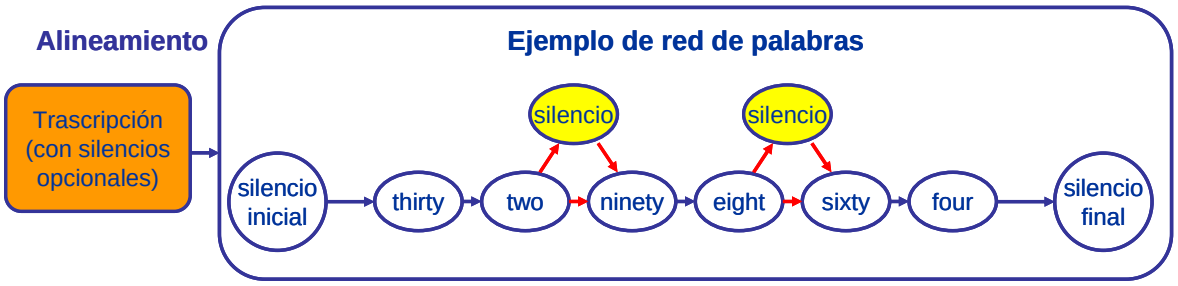


Figura 26: Red de palabras empleada en el sistema de partida para las trascripciones de test

Aquí podemos observar la red de palabras con las transiciones forzadas entre palabras (en azul) así como las transiciones opcionales entre palabras (en rojo).

Con la red de palabras, el audio de test parametrizado y el modelo fonético independiente del locutor se obtiene un fichero de etiquetas en el que se indican, los instantes de principio y final de cada una de las palabras, así como los instantes de principio y final de sus fonemas y de los estados de los fonemas. En ese archivo de etiquetas también se muestra la puntuación obtenida (score) al enfrentar el audio con el modelo fonético independiente del locutor. El audio enfrentado estará comprendido entre dos instantes marcados por el alineamiento, como el alineamiento nos dice qué fonema debe ser pronunciado en ese lapso de tiempo, el fragmento de audio parametrizado se enfrentará con el modelo acústico correspondiente a dicho fonema. Se procederá de igual forma con los modelos fonéticos dependientes del locutor, obteniendo otro fichero de etiquetas con sus marcas temporales y puntuaciones. Los enfrentamientos consisten en un reconocimiento mediante el algoritmo de Viterbi.

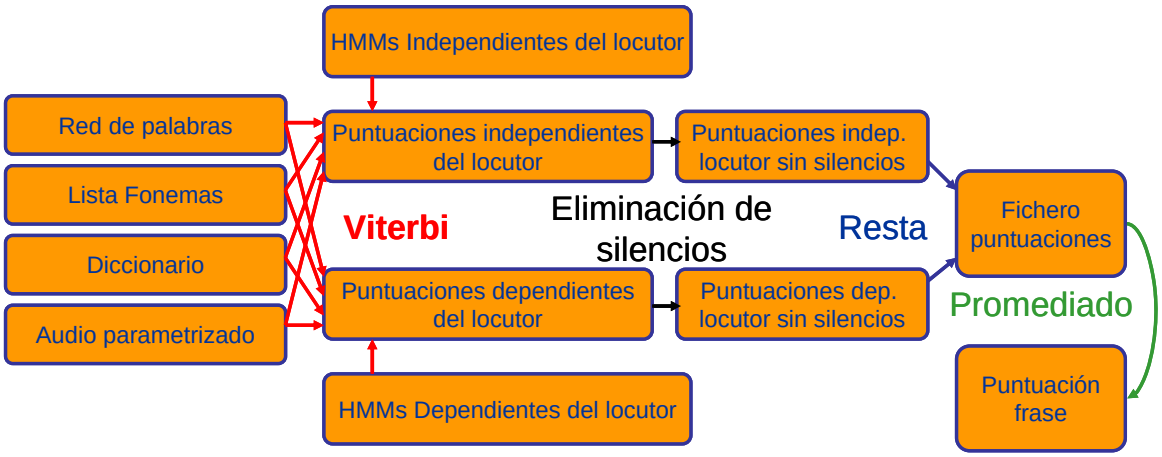


Figura 27: Esquema de la fase de reconocimiento

Posteriormente se eliminan las etiquetas y puntuaciones referidas a silencios, ya que no nos interesan para el reconocimiento. Con las puntuaciones de los fonemas que quedan, se restan las obtenidas por el modelo independiente del locutor ($score_{Indep}$) de las obtenidas por el modelo dependiente del locutor ($score_{Dep}$), obteniendo así un fichero de etiquetas y puntuaciones definitivas por estado, fonema y palabra.

$$score_{Dep} - score_{Indep}$$

Las puntuaciones de los estados de los fonemas con las que estamos trabajando están ponderadas por su duración, de modo que finalmente las multiplicaremos por la duración del estado, sumaremos todas ellas y dividiremos por la duración total (sin silencios), obteniendo así la puntuación por frase.

7.2 Combinación de adaptación MLLR y MAP

La primera mejora realizada consiste en añadir una etapa más a la adaptación de los modelos fonéticos al locutor. Observemos varios experimentos con el sistema de partida [22] con adaptación MLLR global y con clases de regresión:

		Gaussianas por estado				
		5	**10**	**20**	**40**	**80**
clases de regresión	**1**	6.5	6.0	5.9	5.8	5.6
	2	5.3	4.8	4.7	4.6	4.3
	4	9.1	5.6	4.8	4.5	4.2
	8	9.1	5.4	5.1	4.6	4.2
	16	9.1	5.4	4.9	4.7	4.2
	32	9.1	5.4	4.9	4.7	4.2

Tabla 8: EERs obtenidas con diferentes configuraciones del sistema de partida

Dados estos resultados se ha decidido continuar trabajando con 40 Gaussianas por estado y 2 clases de regresión, buscando un equilibrio entre la eficiencia del sistema y el coste computacional (éste se hace mayor cuantas más clases de regresión y más Gaussianas se

empleen). En este punto se pretende mejorar los resultados añadiendo una etapa de adaptación MAP después de la adaptación MLLR. De esta forma la adaptación del sistema resultante queda así:

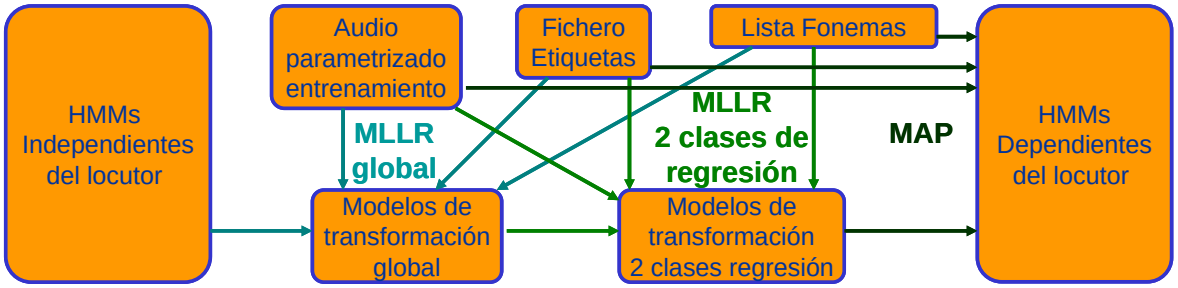


Figura 28: Esquema del proceso de entrenamiento de modelos dependientes del locutor

Se realiza una primera adaptación MLLR global de la que resultará una matriz de transformación, que aplicada al modelo independiente del locutor nos dará una primera adaptación al locutor. En la siguiente etapa se toma el modelo adaptado en la etapa anterior y se realiza una nueva adaptación MLLR, pero esta vez con 2 clases de regresión, resultando una nueva matriz de transformación que aplicada al modelo adaptado al locutor en la etapa anterior nos da un nuevo modelo adaptado al locutor. Por último, tomando como modelo de partida el obtenido como combinación de las dos etapas anteriores, se vuelve a adaptar dicho modelo al locutor mediante adaptación MAP, resultado el modelo adaptado al locutor definitivo.

7.3 Alineamiento no completamente forzado

Esta mejora consiste en modificar las trascripciones tanto para entrenamiento como para test. Para ello cambiamos las trascripciones para que presenten silencios opcionales entre cada palabra, de esta forma también cambia la red de palabras. Esto ya se hacía (aunque no con silencios opcionales entre todas las palabras) en el sistema de partida para las trascripciones de test, pero no para las trascripciones de entrenamiento.

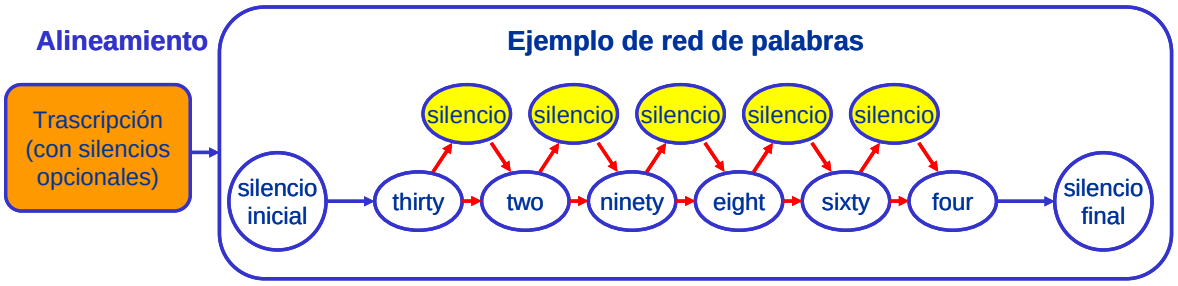


Figura 29: Ejemplo de red de palabras empleada para el reconocimiento previo a la adaptación

Posteriormente se realiza una etapa de reconocimiento previo a la adaptación mediante el algoritmo de Viterbi. Así obtenemos un fichero de etiquetas de entrenamiento con el alineamiento real de palabras, que será el alineamiento reconocido en el audio por el modelo independiente del locutor de todos los probables alineamientos representados en la red de palabras.

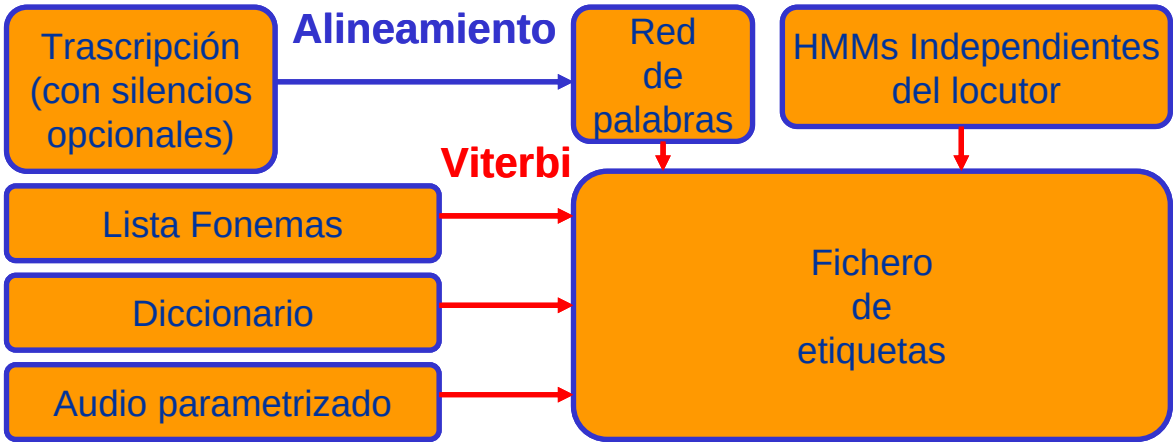


Figura 30: Esquema del reconocimiento previo a la adaptación

Realizar este tipo de trascripción en entrenamiento hace que el fichero de etiquetas de entrenamiento sea más completo y detallado, de esta forma cuando se realice la adaptación se minimiza la posibilidad de que se esté adaptando un fonema con un fragmento de audio de silencio. Con esta mejora, en definitiva conseguimos que la adaptación de los modelos al locutor se haga con el audio correspondiente a los fonemas más acotado y por tanto de una forma más precisa. De esta manera el sistema resultante queda como muestra la siguiente figura:

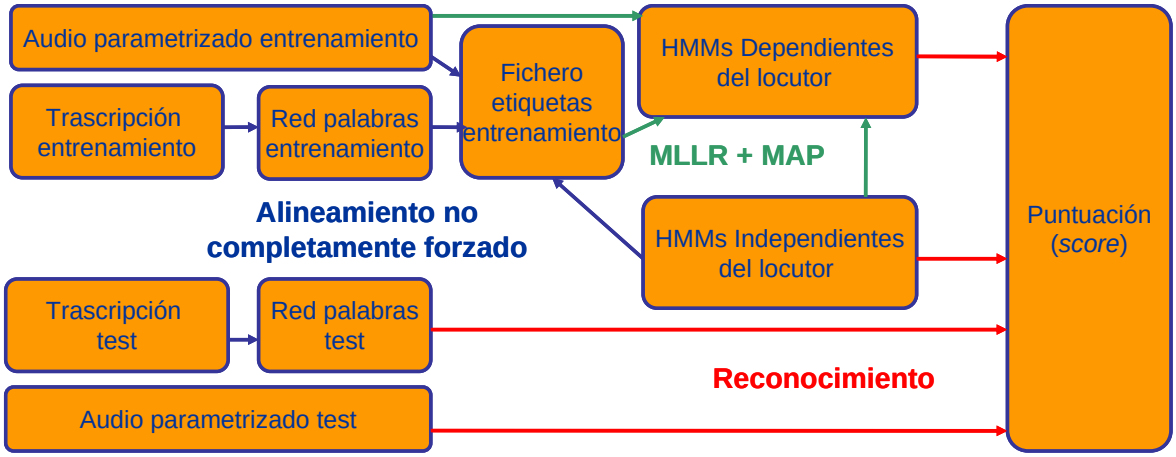


Figura 31: Esquema del sistema resultante después de las mejoras en entrenamiento

7.4 T-norm a nivel de fonema y estado

Una vez se han obtenido las puntuaciones se puede ir más allá y dar un paso más para mejorar los resultados normalizando las puntuaciones. Se ha escogido la normalización de test o T-norm, que consiste en seleccionar una cohorte de impostores cuyos modelos fonéticos se enfrenten a la locución bajo estudio, se calcula la media y la desviación típica de las puntuaciones obtenidas por la cohorte de impostores y se calcula la nueva puntuación. La puntuación definitiva se calcula como la puntuación obtenida por el modelo que enfrentamos a la locución menos la media de las puntuaciones de la cohorte de impostores entre la desviación típica de dichas puntuaciones.

$$L_{T-norm}(X \mid \lambda, T) = \frac{L(X \mid \lambda) - \mu_T(X)}{\sigma_T(X)}$$

Tradicionalmente se ha seguido un esquema de normalización en el que las puntuaciones con las que se trabajaba eran las puntuaciones promedio de la frase o locución completa, siguiendo el siguiente esquema:

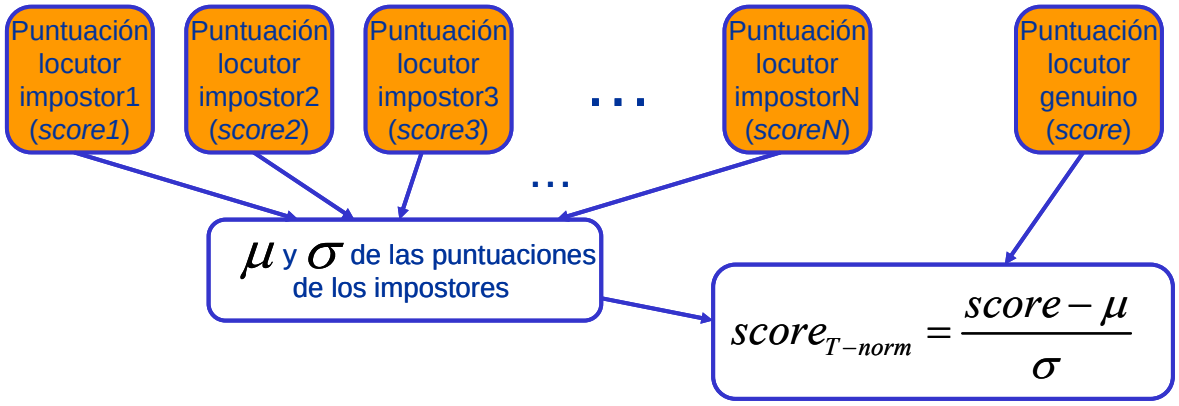


Figura 32: Forma de puntuación con T-norm a nivel de frase

En este proyecto se ha ido más allá y se han implementado dos estructuras de normalización similares pero a distinto nivel, a nivel de fonema y a nivel de estado. Básicamente consisten en realizar T-norm pero en vez de hacerlo con las puntuaciones promedio de la frase, se hace con la puntuación de cada fonema o de cada estado de cada fonema de la frase. Una vez tenemos cada fonema o cada estado con su puntuación normalizada mediante T-norm se calcula la puntuación promedio de la frase de la misma manera que sin T-norm. El esquema de funcionamiento es el siguiente:

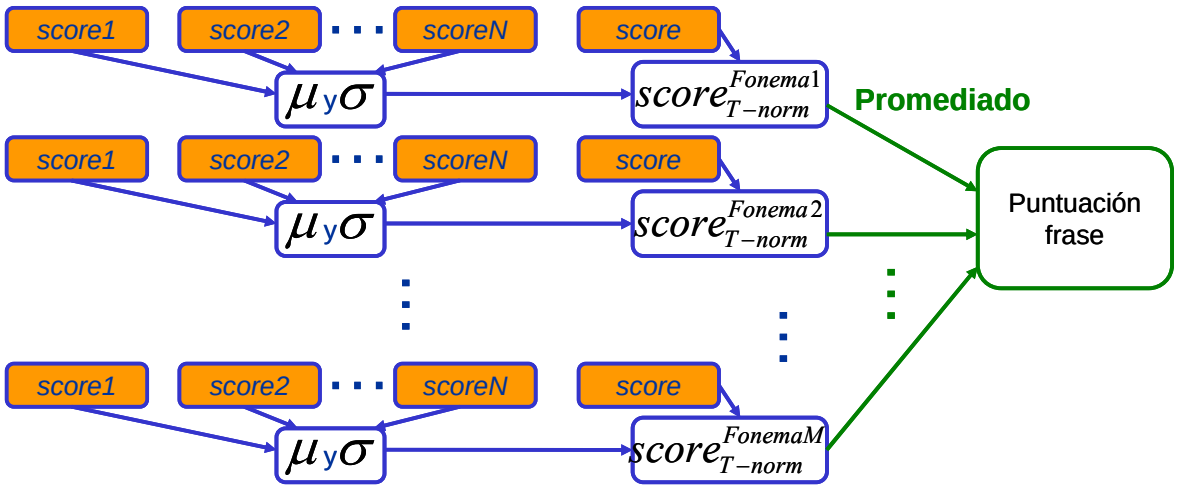


Figura 33: Esquema de puntuación con T-norm a nivel de fonema

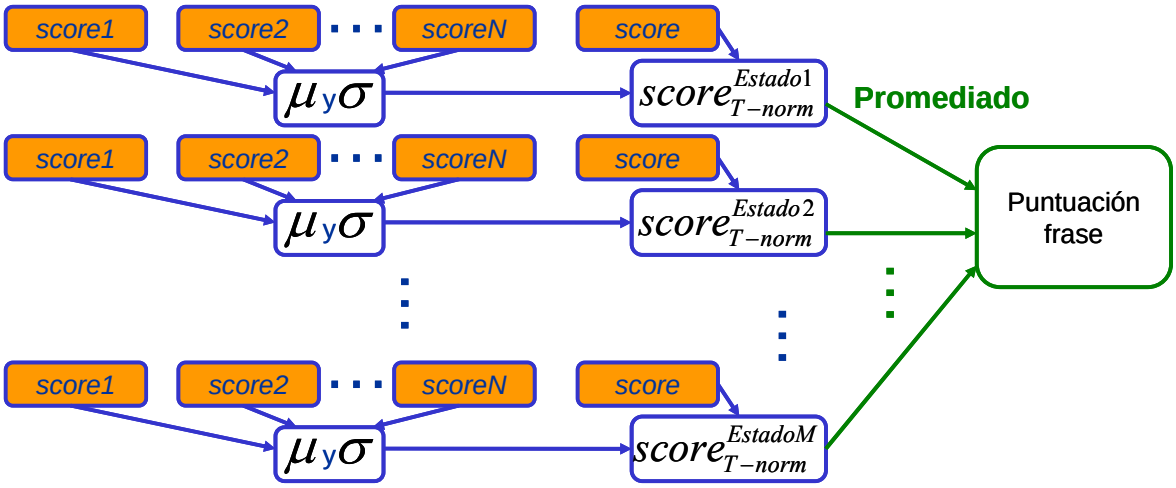


Figura 34: Esquema de puntuación con T-norm a nivel de estado

8 Experimentos y resultados

En este capítulo veremos los experimentos realizados con cada base de datos, el protocolo de pruebas seguido y los resultados obtenidos al aplicar las mejoras.

8.1 Experimentos con la base de datos BioSec Baseline

Con la base de datos BioSec Baseline [52] se ha utilizado el sistema de partida con la configuración de 40 Gaussianas por estado y adaptación MLLR global junto con MLLR con 2 clases de regresión. El motivo de usar esta base de datos es observar la influencia del idioma y la calidad del audio sobre el sistema, complementando así otro estudio [22] realizado sobre esta misma base de datos. Un hecho que debemos tener en cuenta es que los locutores de esta base de datos son en su mayoría hablantes nativos de lengua española, de modo que tenemos habla en español y en inglés pero en ambos casos el locutor tendrá como idioma nativo el español.

Para realizar los experimentos que a continuación se describen se han utilizado las 4 locuciones con el código del locutor de la primera sesión para entrenar 4 modelos dependientes del locutor (uno con cada frase). Para la etapa de test se han utilizado como locuciones target las 4 frases con el código del locutor de la segunda sesión, usando como locuciones non-target la primera frase de la primera sesión del resto de locutores. Hay que fijarse en que hemos utilizado el mismo léxico para las locuciones de entrenamiento y test, lo cual influirá positivamente sobre el resultado [45].

Se han realizado 4 experimentos con el sistema de partida, inglés grabado con micrófono de habla cercana (integrado en unos auriculares), inglés grabado con micrófono de habla lejana (integrado en una webcam), español grabado con un micrófono de habla cercana y español grabado con un micrófono de habla lejana. Las EER (expresados en %) resultantes son las que siguen:

Canal\Idioma	Castellano	Inglés
Mic. cercano	1.68	2.17
Mic. lejano	17.24	12.72

Tabla 9: Resultados obtenidos con la base de datos BioSec Baseline

Las curvas DET son las que se presentan a continuación:

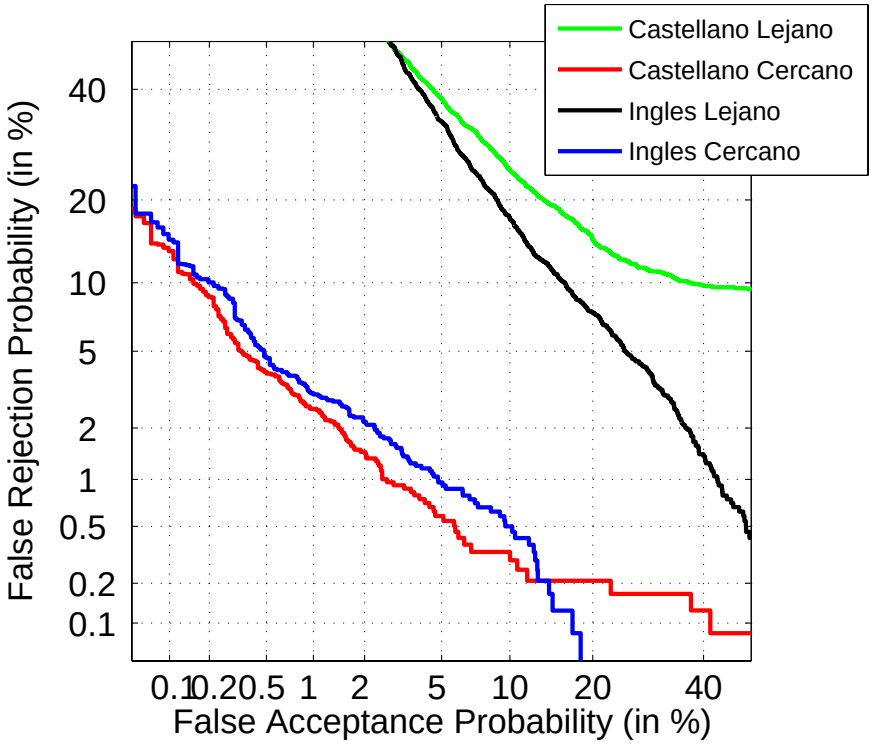


Figura 35: Comparación de los resultados obtenidos con la base de datos BioSec Baseline

De donde podemos concluir que el micrófono de habla cercana presenta una calidad muy superior al micrófono de habla lejana y esto influye notablemente en el resultado del reconocimiento. Además los resultados con el micrófono de habla cercana en esta base de datos son considerablemente mejores que los obtenidos con el mismo sistema y configuración en la base de datos YOHO, exceptuando que para la base de datos YOHO utilizamos 6 frases de entrenamiento en vez de 1. La explicación más probable es que en este caso no existe desajuste léxico, es decir entrenamos los modelos con los mismos fonemas que luego reconoceremos, por lo tanto el sistema reconoce muy bien el patrón que ya tenía entrenado.

8.2 Experimentos con la base de datos YOHO

Esta es la base de datos sobre la que se han evaluado las mejoras hechas al sistema de partida. Además al ser la base de datos más usada en reconocimiento de locutor dependiente de texto permite comparar los resultados obtenidos con los obtenidos en otros laboratorios [54]. El protocolo de experimentos que se ha seguido ha sido emplear 6 locuciones de la primera sesión para entrenar los modelos dependientes del locutor, para locuciones de test target se han utilizado las 40 locuciones de test (repartidas en 10 sesiones) de la base de datos, usando una locución de test aleatoria de cada uno de los demás locutores como intentos non-target.

Todos los experimentos que a continuación se describen se han realizado con una configuración de 40 Gaussianas por estado y 2 clases de regresión en la adaptación MLLR

con clases de regresión. Con el sistema de partida el resultado obtenido con esta base de datos, usando el protocolo anteriormente descrito es el siguiente:

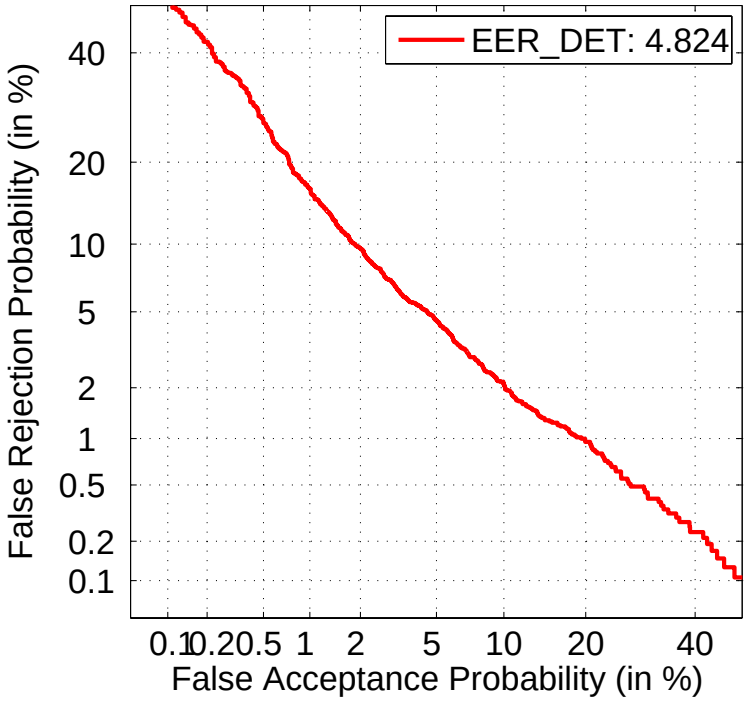


Figura 36: Curva DET obtenida con el sistema de partida

8.2.1 Combinación de adaptación MLLR y MAP

Los primeros experimentos realizados tienen la finalidad de comprobar si realizar adaptación MAP posteriormente a adaptación MLLR es beneficioso. De este modo se realizan 2 sencillos experimentos. En el primero de ellos se realizará adaptación al locutor MAP después de realizar adaptación MLLR global, resultando lo siguiente:

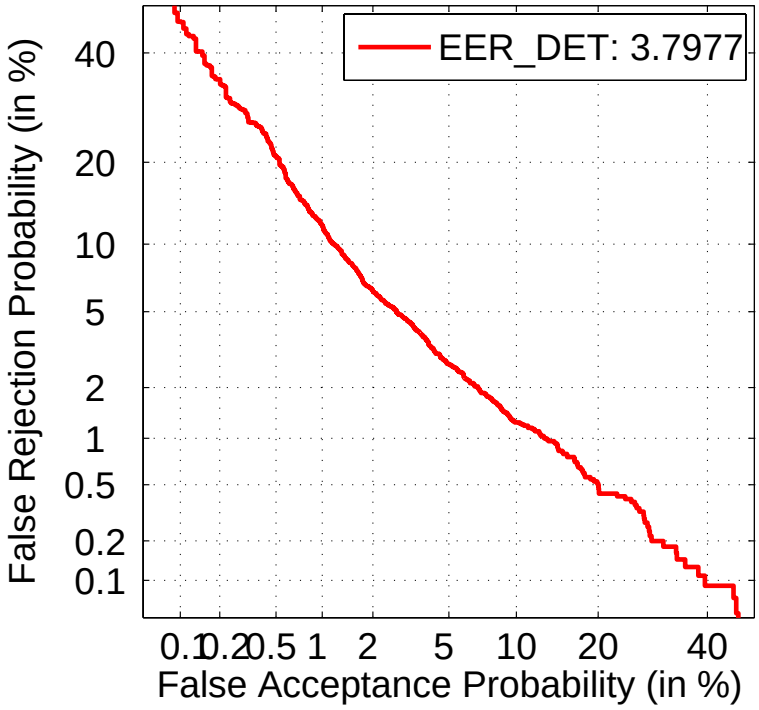


Figura 37: Resultado obtenido combinando MLLR global y MAP

A continuación se repite el experimento anterior pero realizando adaptación MAP después de realizar adaptación MLLR global y posteriormente con 2 clases de regresión. Así es como queda el resultado:

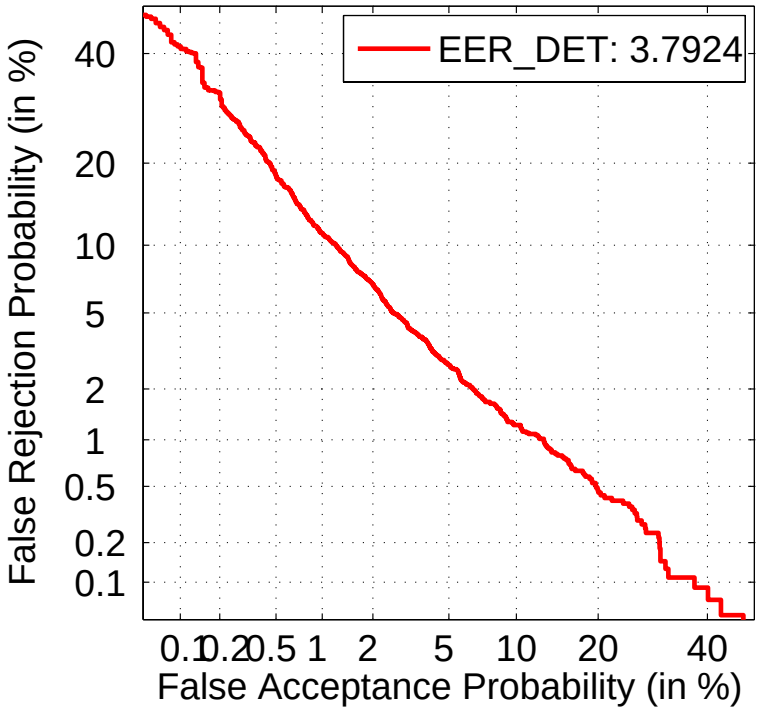


Figura 38: Resultado obtenido combinando MLLR global, MLLR con 2 clases de regresión y MAP

Vemos que aunque la EER mejora, la mejora del sistema es más sustancial para valores de baja probabilidad de falsa aceptación.

De este modo concluimos que introducir adaptación MAP después de realizar adaptación MLLR global y posteriormente MLLR con 2 clases de regresión es beneficioso para el sistema, introduciendo una mejora relativa en la EER del 21.32%

8.2.2 Alineamiento no completamente forzado

Partiendo de las mejoras introducidas descritas en el apartado anterior, se ha realizado una nueva mejora consistente en introducir silencios opcionales en las trascripciones tanto de entrenamiento como de test. Además se ha realizado un reconocimiento previo del audio de entrenamiento para mejorar la trascripción final que se proporciona para realizar la adaptación. Con todo ello el nuevo resultado es el siguiente:

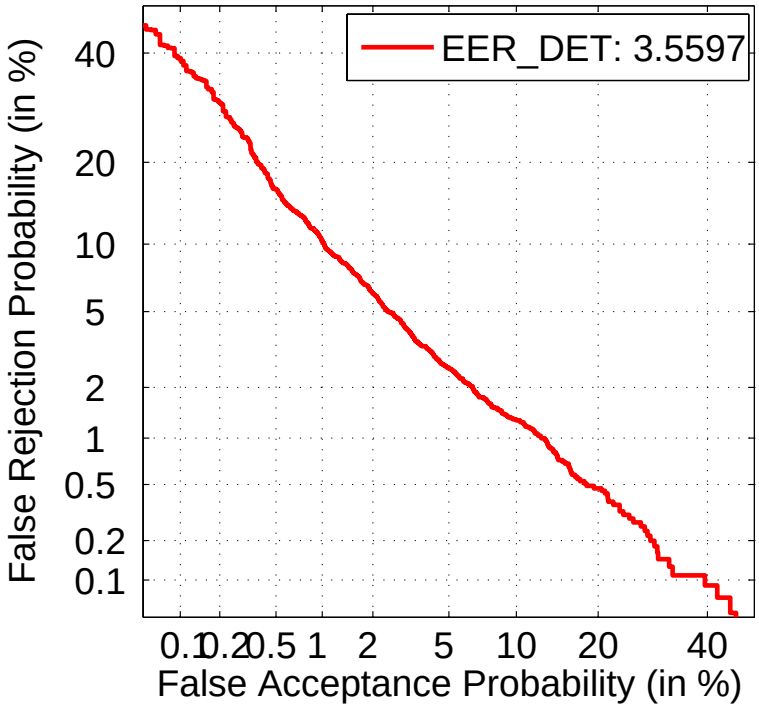


Figura 39: Resultado obtenido realizando alineamiento no completamente forzado y reconocimiento previo al entrenamiento

Con lo cual podemos concluir que realizar un reconocimiento previo contemplando la posibilidad de silencios opcionales entre palabras mejora el rendimiento del sistema, concretamente se produce una mejora relativa de la EER del 6.14%

8.2.3 T-norm a nivel de fonema y estado

Finalmente, y manteniendo las mejoras anteriormente descritas se han realizado una nueva serie de experimentos normalizando las puntuaciones mediante T-norm. Para realizar este tipo de experimentos debemos seleccionar previamente una cohorte de impostores, de este

modo con sus modelos se realizarán los reconocimientos necesarios para tener puntuaciones con las que implementar T-norm.

En esta serie de experimentos entran en juego 2 variables que se intentarán relacionar con los resultados, la selección de impostores de la cohorte de T-norm y el nivel al que se implementará. La selección de impostores de la cohorte de T-norm se ha basado a su vez en 2 parámetros, el género y el número de impostores. Los niveles a los que se ha realizado T-norm han sido nivel de frase, nivel de fonema y nivel de estado. De este modo se han ido realizando los siguientes experimentos:

Primero se ha seleccionado una única cohorte de impostores para T-norm con 10 hombres y 10 mujeres (TN10) extraídos de la base de datos YOHO. Los 20 locutores que forman la cohorte de T-norm han sido empleados únicamente para generar puntuaciones con las que calcular la media y desviación típica para implementar T-norm. Esto quiere decir que no habrá enfrentamientos de ningún modelo con audio proveniente de estos locutores y además serán siempre los mismos. De esta forma en la base de datos quedan 118 locutores de los cuales 96 son hombres y 22 mujeres, siendo de éstos de donde se saquen las locuciones que generarán los enfrentamientos. Con este protocolo hemos implementado T-norm a nivel de frase, a nivel de fonema y a nivel de estado. En la Figura 40 que se muestra a continuación podemos observar las curvas DET obtenidas y compararlas con el resultado obtenido sin implementar T-norm:

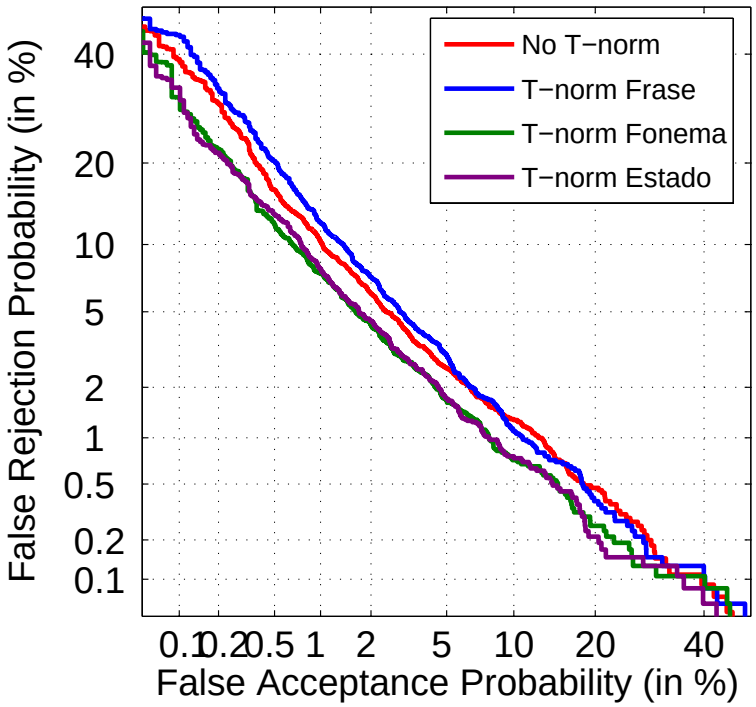


Figura 40: Comparación de resultados con T-norm independiente de género a distintos niveles

Los porcentajes de EER obtenidos para los experimentos comentados son los que se pueden observar en la tabla 10. Podemos ver que el resultado mejora considerablemente implementado T-norm a nivel de fonema y estado, incluso podemos observar que en este caso es preferible no realizar T-norm a implementar T-norm a nivel de frase. Si nos

fijamos un poco más vemos también que parece resultar sensiblemente mejor implementar T-norm a nivel de fonema que a nivel de estado.

T-norm\Nivel	Frase	Fonema	Estado
No	3.56		
TN10	3.91	2.98	3.04

Tabla 10: Comparación de resultados con T-norm independiente de género a distintos niveles

El siguiente paso ha sido dividir la cohorte de T-norm anteriormente descrita en 2 cohortes de impostores, una de hombres y otra de mujeres, cada una de ellas con 10 impostores. De este modo si la locución pertenecía a un hombre se implementaba T-norm con la cohorte de impostores masculinos, utilizando la cohorte femenina si la locución pertenecía a una mujer. Con este planteamiento hemos vuelto a realizar los experimentos anteriormente descritos analizando el comportamiento del sistema ante un intento de acceso de un hombre, de una mujer y el comportamiento del sistema en general (ambos). Los resultados obtenidos son los siguientes:

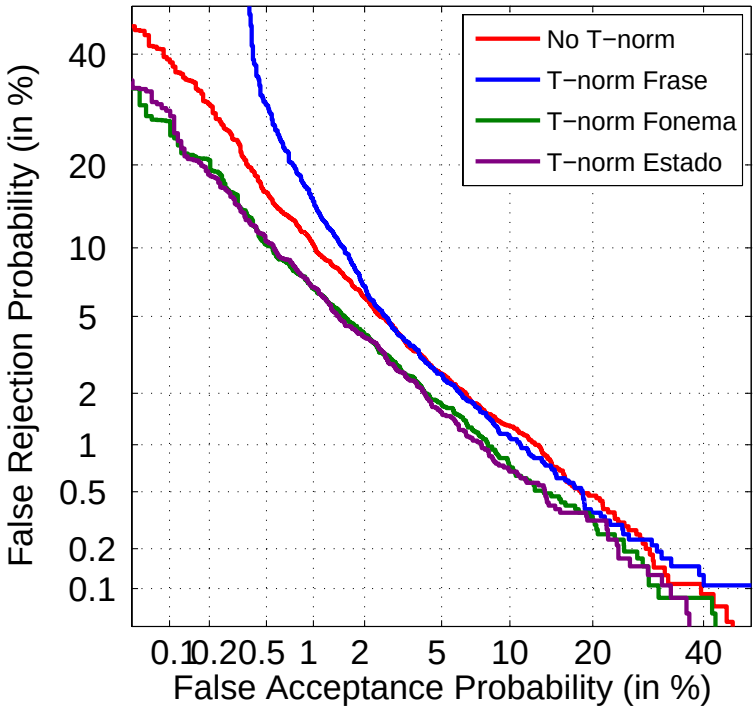


Figura 41: Comparación de resultados con T-norm (TN10) dependiente de género a distintos niveles (ambos géneros)

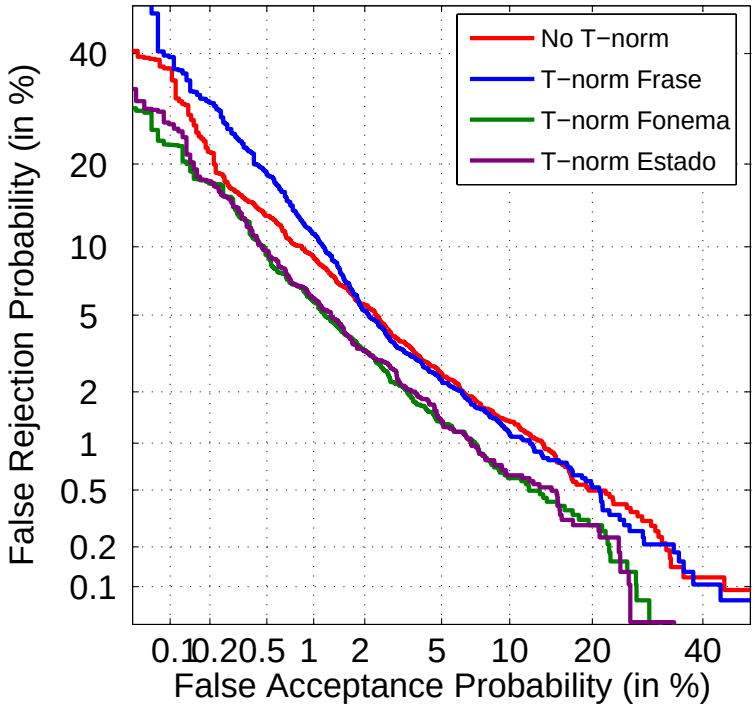


Figura 42: Comparación de resultados con T-norm (TN10) dependiente de género a distintos niveles (género masculino)

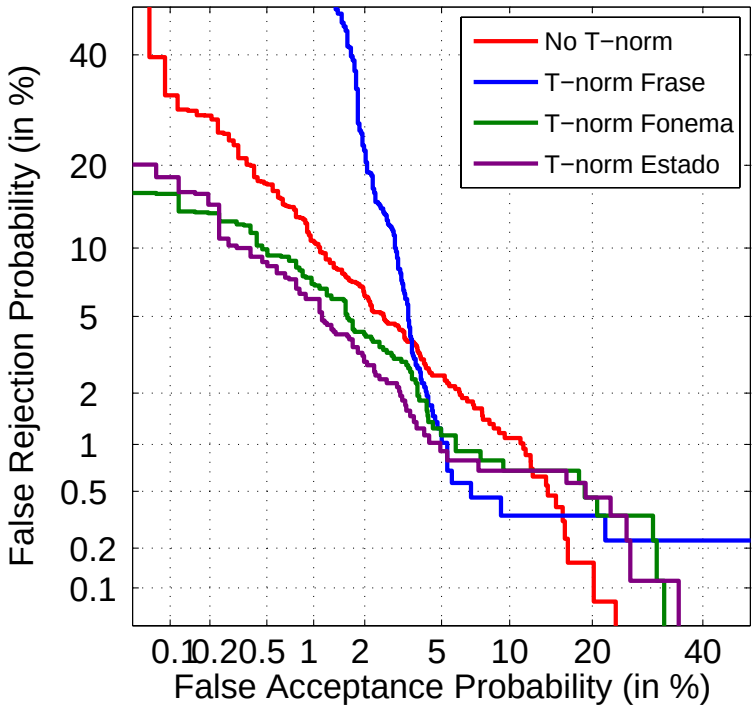


Figura 43: Comparación de resultados con T-norm (TN10) dependiente de género a distintos niveles (genero femenino)

A continuación hemos aumentado el número de impostores de cada género a 30 (TN30), pero para ello, y debido a las limitaciones en número de locutores de la base de datos YOHO, hemos tenido que generar cohortes de impostores variables. Lo que viene a significar es que para cada intento de acceso generado se seleccionan los modelos de una cohorte consistente en 30 locutores del mismo género que el que intenta acceder al sistema. Para los intentos target se excluye de dicha cohorte el modelo del locutor que intenta acceder al sistema (que coincide con el modelo del locutor real del audio), excluyendo además el modelo del locutor real en los intentos non-target (en estos intentos el modelo del locutor por el que se pretende hacer pasar es distinto del locutor real).

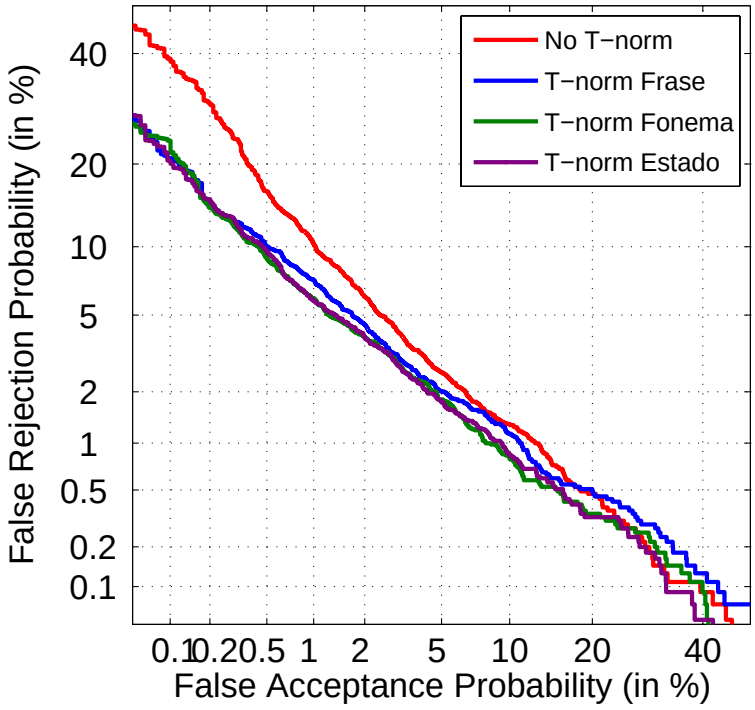


Figura 44: Comparación de resultados con T-norm (TN30) dependiente de género a distintos niveles (ambos géneros)

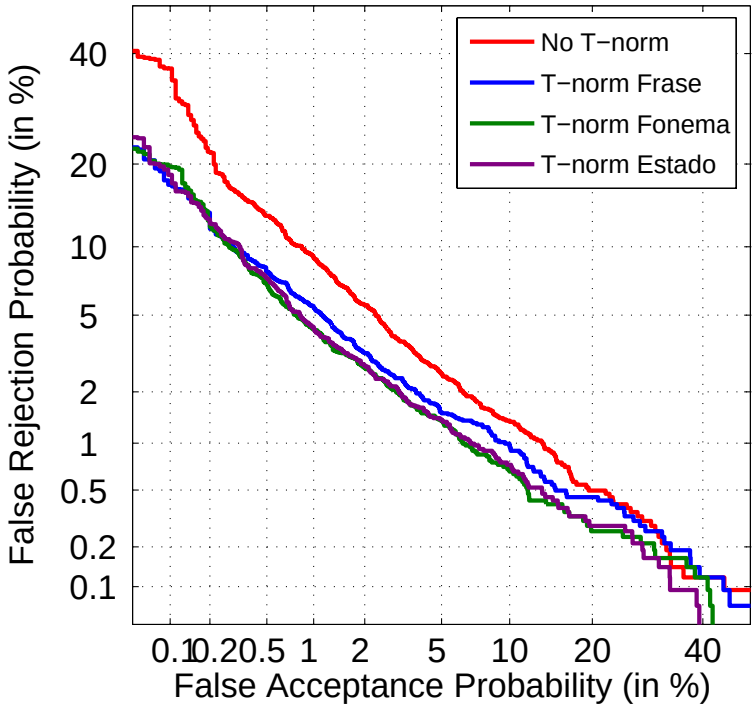


Figura 45: Comparación de resultados con T-norm (TN30) dependiente de género a distintos niveles (género masculino)

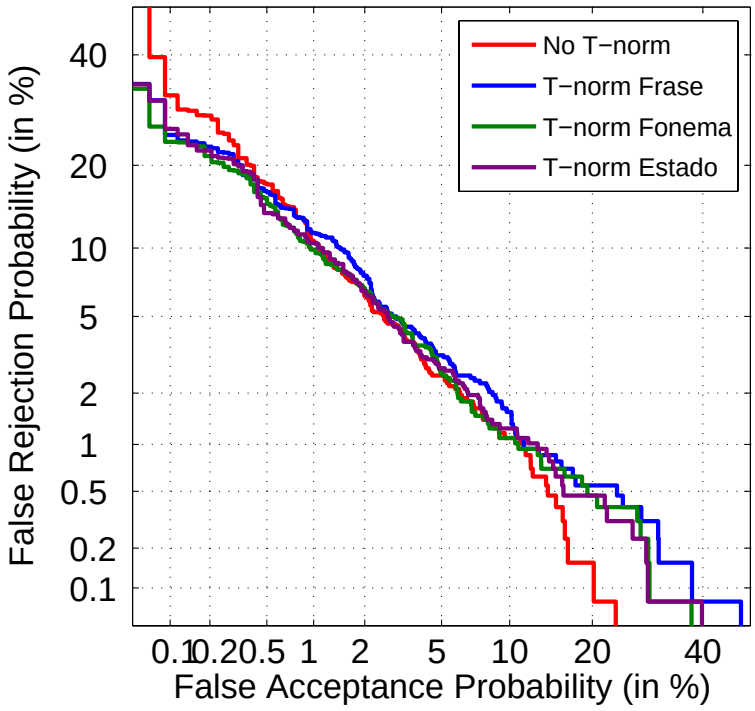


Figura 46: Comparación de resultados con T-norm (TN30) dependiente de género a distintos niveles (género femenino)

Como con la base de datos YOHO no podíamos obtener una cohorte de impostores mayor para el género femenino, se ha decidido realizar una prueba con una cohorte mayor pero únicamente con los locutores de género masculino. En este caso la cohorte de impostores es también variable y está compuesta por los modelos de todos los hombres exceptuando aquel a quien pertenece la voz realmente en el caso de los intentos target (105 modelos). En el caso de los intentos non-target la cohorte es variable y está compuesta por los modelos de todos los hombres, exceptuando tanto el modelo del locutor real como el modelo del locutor por el que se intenta hacer pasar el impostor (104 modelos). Los resultados se muestran a continuación:

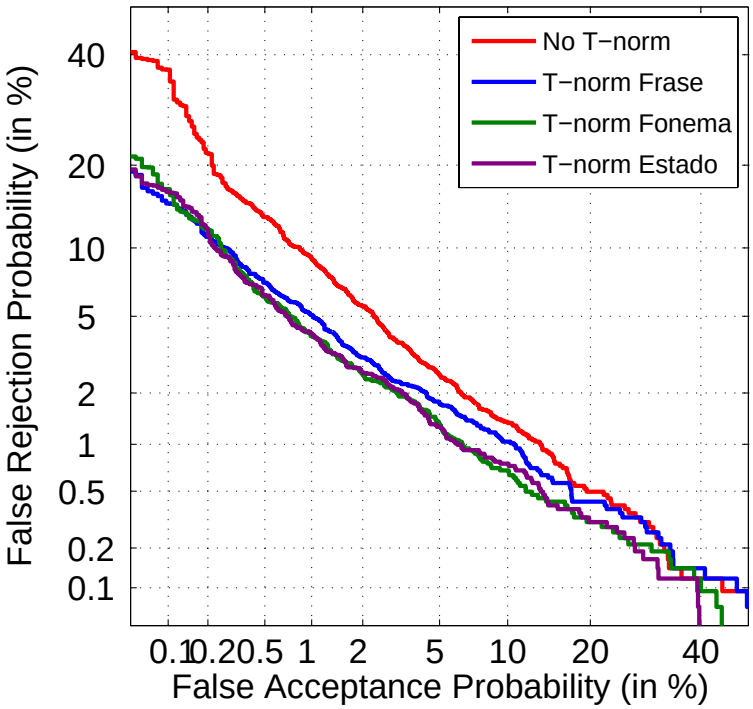


Figura 47: Comparación de resultados con T-norm (TNMasc) dependiente de género a distintos niveles

Para dar una visión más genérica de los resultados se ha confeccionado la siguiente tabla:

T-norm	Género	Frase	Fonema	Estado
No	Masc	3.54		
	Fem	3.72		
	Ambos	3.56		
TN10	Masc	3.32	2.64	2.80
	Fem	3.57	3.15	2.45
	Ambos	3.64	2.97	2.91
TN30	Masc	2.69	2.48	2.46
	Fem	3.99	3.67	3.67
	Ambos	3.10	2.98	2.96
TNMasc		2.66	2.41	2.53

Tabla 11: Comparación de resultados de distintas cohortes de T-norm dependiente de género

De la tabla podemos sacar varias conclusiones. En primer lugar podemos observar que exceptuando los resultados para locutores de género femenino, los resultados son mejores cuanto mayor es la cohorte de locutores. También podemos ver que es mejor implementar T-norm a nivel de fonema o estado que a nivel de frase, hasta el punto de que hay casos en los que es mejor no realizar T-norm que realizarlo a nivel de frase. Esto se presume que es debido a que realizando T-norm a nivel de estado y fonema se compensa de alguna manera el desajuste léxico. El desajuste léxico es la diferencia existente entre las palabras empleadas para entrenar los modelos y las empleadas para realizar el reconocimiento. Por eso si realizamos T-norm a nivel de fonema y estado mejoramos de una forma tan abultada los resultados, porque así aprovechamos la ventaja de poder trabajar a nivel de fonema y estado no sólo en las etapas de entrenamiento y reconocimiento, sino también en la etapa de normalización de puntuaciones. Si comparamos esta tabla con la anterior además se puede observar que los resultados globales del sistema mejoran

Sin embargo a la vista de los resultados, aunque podemos asegurar que realizar T-norm a nivel de fonema o estado es mejor que no hacerlo o hacerlo a nivel de frase, no podemos determinar de forma concluyente si es mejor realizar T-norm a nivel de fonema o a nivel de estado. Ante esta situación se ha realizado un estudio analítico de las puntuaciones obtenidas para los distintos fonemas y estados por una cohorte de impostores de T-norm, en concreto TN10. Con el fin de entender este comportamiento se han calculado varios parámetros estadísticos de las puntuaciones obtenidas por la cohorte de impostores TN10. Para cada una de las puntuaciones a las que aplicamos T-norm se calcula la media y la desviación típica de las puntuaciones obtenidas por la cohorte de impostores. Aprovechando esta circunstancia se han guardado todas las medias y desviaciones típicas generadas en el proceso de T-norm para cada estado de todos los enfrentamientos realizados para el experimento de T-norm a nivel de estado. Posteriormente se han calculado la media y desviación típica de todas las medias y desviaciones típicas generadas para cada estado, obteniendo el siguiente resultado.

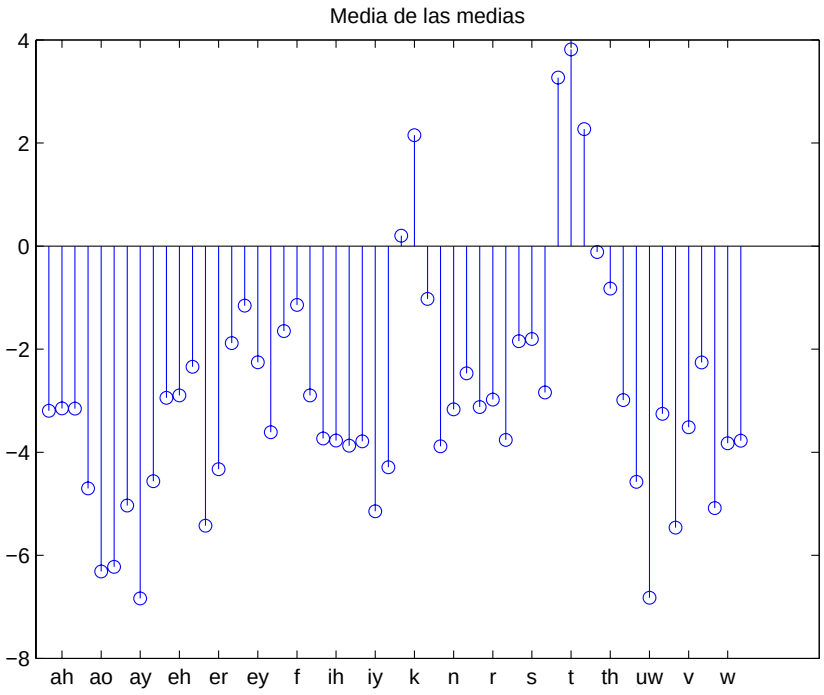


Figura 48: Medias de las medias de las puntuaciones para cada estado

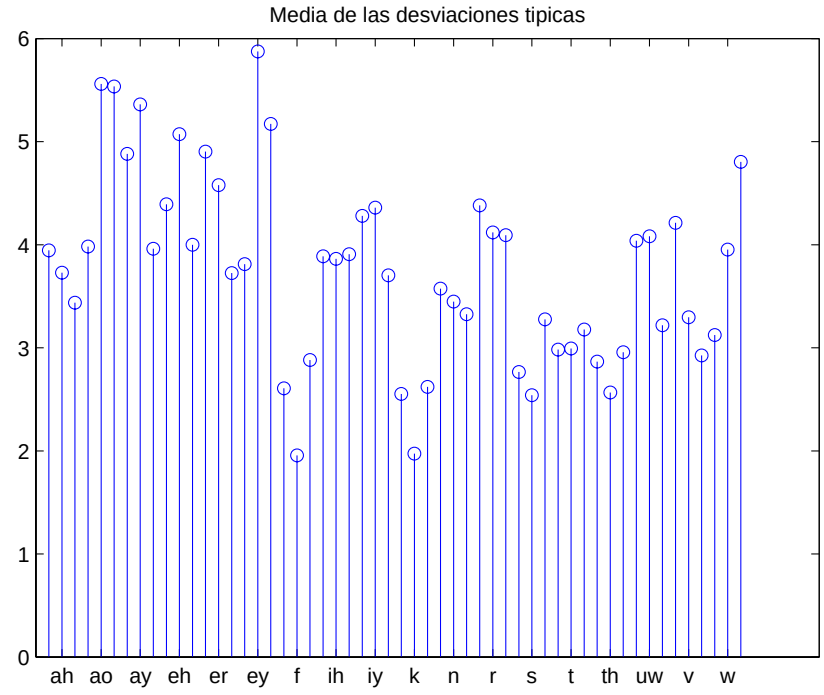


Figura 49: Medias de las desviaciones típicas de las puntuaciones para cada estado

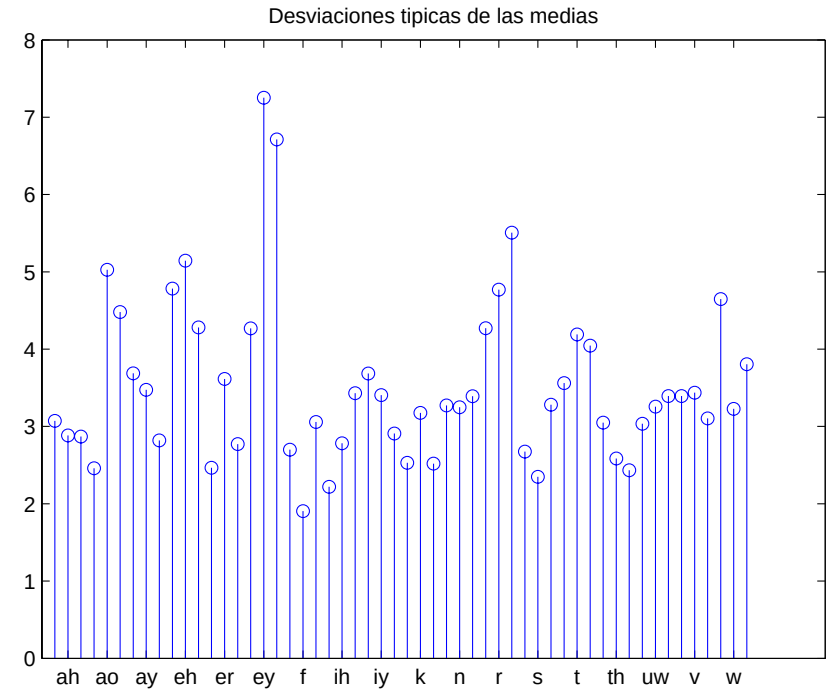


Figura 50: Desviaciones típicas de las medias de las puntuaciones para cada estado

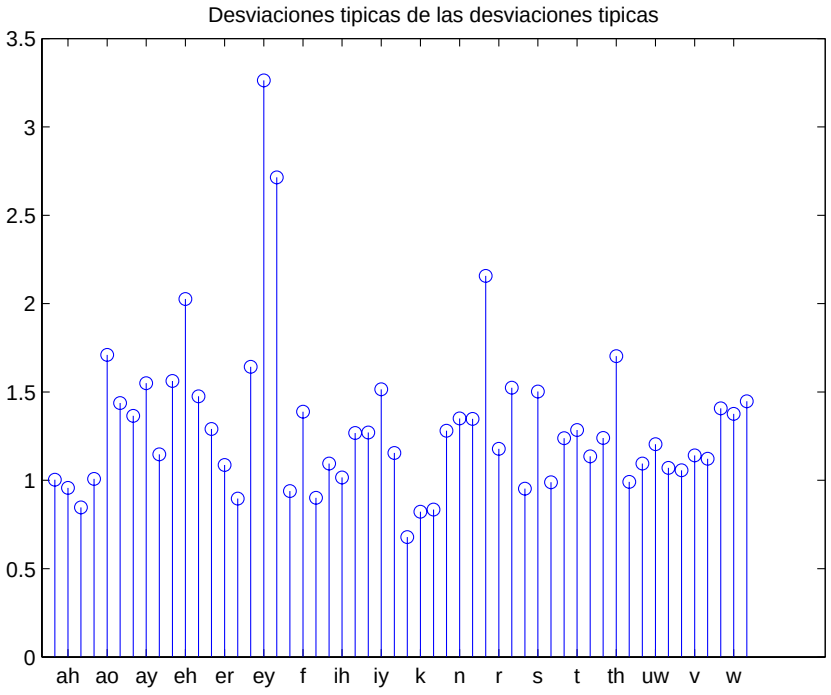


Figura 51: Desviaciones típicas de las desviaciones típicas de las puntuaciones para cada estado

En el eje vertical de estas gráficas (Figuras de la 48 a la 51) se representa el valor del parámetro estadístico bajo análisis y en el eje horizontal podemos ver el estado y el fonema al que se asocia. Cada fonema tiene 3 estados, de modo que se ha colocado el nombre del fonema en lo que sería su segundo estado.

En estas gráficas se puede observar que hay cierta correlación entre las puntuaciones de los estados de un mismo fonema, sin embargo no podemos decir lo mismo entre fonemas distintos. Esto explica el hecho de que no exista una gran diferencia entre los resultados obtenidos realizando T-norm a nivel de fonema y T-norm a nivel de estado.

Si nos fijamos en la formulación de T-norm y nos paramos a reflexionar sobre cómo esta técnica modifica los resultados, podemos intuir que cuanto más se parezca la cohorte de impostores al locutor que pronunció la locución mejores resultados tendremos. Esto se traduce en que si la cohorte de impostores se parece al locutor original puntuará más alto, lo cual hará que la media de las puntuaciones sea mayor. Esto produce que la puntuación final obtenida por el locutor original (próxima a la media y previsiblemente por encima) esté más separada de la puntuación obtenida por un impostor (lejana a la media y previsiblemente bastante por debajo) después de realizar T-norm.

Si comparamos la Tabla 11 con la Tabla 10 podemos observar que para el caso TN10 los resultados globales (ambos géneros) del sistema mejoran, sin embargo la realidad es que para este caso tenemos menos locutores en la cohorte de impostores que en caso de T-norm independiente de género. Lo que ocurre es que aunque para el caso independiente de género tengamos una cohorte de 20 impostores en total, teniendo sólo 10 para el caso dependiente de género, lo que hemos hecho ha sido seleccionar la cohorte que se parezca más al locutor. Por lo tanto parece ser más importante la selección de la cohorte de impostores que el número de éstos.

9 Conclusiones y trabajo futuro

9.1 Conclusiones

Dados los resultados obtenidos en los diferentes experimentos con la base de datos YOHO se puede concluir que realizar adaptación MAP después de adaptación MLLR global y con clases de regresión en el entrenamiento de los modelos acústicos dependientes del locutor mejora los resultados obtenidos introduciendo una mejora relativa en la EER del 21.32%.

También podemos concluir que realizar un reconocimiento previo al entrenamiento contemplando la posibilidad de silencios opcionales entre palabras, obteniendo así un alineamiento no completamente forzado, mejora el rendimiento del sistema, concretamente se produce una mejora relativa de la EER del 6.14%

Por otra parte se ha visto que el desajuste léxico tiene una gran influencia en la eficiencia de un sistema de reconocimiento de locutor dependiente de texto. La principal razón es que en este campo se entrenan modelos de unidades léxicas por debajo de la locución completa (palabra, fonema, estado…). Y el hecho de que se intente reconocer al locutor con modelos de unidades léxicas que se han podido no entrenar previamente (desajuste léxico), o que se han entrenado en contextos léxicos distintos, hace que los resultados empeoren de forma muy abultada. De este modo realizando T-norm a nivel de fonema y estado se han mejorado los resultados en un 16.29% de mejora relativa, llegando a una mejora global del sistema del 18.26% de mejora relativa realizando T-norm a nivel de estado dependiente de género. Por ello puede intuirse que realizar T-norm a nivel de fonema y estado ayuda a reducir el efecto del desajuste léxico. También se ha visto como aumentando el número de locutores de la cohorte de impostores de T-norm los resultados mejoran aunque no de forma abultada. Por esto se ha llegado a la conclusión de que lo que influye con más peso en la mejora de T-norm es tener una cohorte parecida al locutor original más que el número de impostores de la cohorte.

Aunque los resultados obtenidos con la normalización a nivel de estado y fonema son positivos, superando los resultados a nivel de frase, el problema del desajuste léxico sigue sin estar resuelto completamente y sigue teniendo una influencia importante en los resultados.

Por otra parte con los experimentos realizados sobre la base de datos BioSec hemos podido comprobar cómo influye la calidad del micrófono en los resultados obtenidos. Además hemos podido intuir que los buenos resultados obtenidos en esta base de datos para micrófono de calidad, son en gran parte debidos a que en esta base de datos y según el protocolo que hemos empleado no presentan desajuste léxico.

Los resultados obtenidos en este proyecto han dado lugar a dos publicaciones en congresos internacionales:

- BioSec Multimodal Biometric Database and its use in Text-Dependent Speaker Recognition Research, publicado en LREC 2008.
- MAP and Sub-Word Level T-Norm for Text-Dependent Speaker Recognition, publicado en Interspeech 2008.

Además un tercer artículo ha sido enviado a congreso y se encuentra a la espera de ser aceptado:

- T-Norm y desajuste Léxico y Acústico en Reconocimiento de Locutor Dependiente de Texto, enviado a las Jornadas de las Tecnologías del Habla 08.

9.2 Trabajo futuro

La mejora más inmediata que se podría introducir sería la probada en [54], donde usan un modelo generativo para realizar el entrenamiento de los modelos de cada locutor. Posteriormente se entrenan los pesos que deberá tener cada unidad del nivel de trabajo en que nos encontremos (palabras, fonemas, estados…), creando así, de una forma parecida al Boosting, un modelo discriminativo para verificación. Con este método consiguen una mejora relativa del 36.41% frente a los métodos tradicionales de verificación LRT (en inglés, Likelihood Ratio Test).

Otra posible mejora a introducir sería realizar el entrenamiento de forma transparente, esto es sin que el locutor tenga que realizar ningún tipo de actividad específica para ello. En un sistema de entrenamiento transparente el locutor no tendría que leer frases de entrenamiento, aunque sí debería identificarse por algún otro medio. Por ejemplo, un cliente de un banco llama para darse de alta en el servicio de banca telefónica, se identifica mediante algún tipo de clave y da sus datos. Con el habla que se ha producido en la conversación se realiza el entrenamiento de los modelos. El problema de no tener trascripción puede ser resuelto realizando un reconocimiento de habla espontánea previo, utilizando el resultado del reconocimiento como trascripción, o bien una búsqueda de palabras concretas (Word Spotting) y entrenar los modelos de esas palabras, para posteriormente realizar el reconocimiento basado en las mismas.

Referencias

[1] D. Maltoni, D. Maio, A. K. Jain y S. Prabhakar. "Handbook of Fingerprint Recognition", Springer 2003.

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

[3] R. Clarke. "Human identification for information systems: Management challengers and public policy iusses", Info. Technol. People, vol.7, nº4 pp6-37, 1994.

[4] J.D. Woodward. "Biometrics: Privaciy´s Foe Privaciy´s Friend?", Proc. IEEE Special Issue on Automatic biometrics, vol 8, nº 9 pp. 1480-1492, 1997.

[5] A. K. Jain, A. Ross y S. Prabhakar. "An introduction to biometric recognition", IEEE Transactions on Circuits Systems for Video Technology, Vol. 14, N.1, pp. 4-20, 2004.

[6] NIST SRE. Descripciones de las distintas evaluaciones NIST de locutor. http://www.nist.gov/speech/tests/spk

[7] FVC, Fingerprint Verification Competition. Web oficial de la evaluación de 2006. http://bias.csr.unibo.it/fvc2006/

[8] L. Flom y A. Safir. "Iris recognition system", United States Patent 4.641.349, 1987.

[9] NIST ICE. Descripciones de las distintas evaluaciones NIST de iris. http://iris.nist.gov/ICE/

[10] A. Martin, M. Przybocky, G. Doddington, D.A. Reynolds. "The NIST speaker recognition evaluation", Overview, methodology, systems, results, perspectives, Speech Commun. 31, 225-254, 2000.

[11] T. Ganchev, N. Fakotakis, G. Kokkinakis. "Comparative evaluation of various MFCC implementations on the speaker verification task", 10th International Conference on Speech and Computer (SPECOM 2005), vol. 1, 2005, pp. 191–194.

[12] Y. Liu, M. Russell, M. Carey. "The role of dynamic features in text-dependent and -independent speaker verification", Proc. IEE ICASSP 2006(1), 669-672, 2006.

[13] D. Reynolds. "Channel Robust Speaker Verification via Feature Mapping", Proc. IEEE ICASSP 2003(2), 53-56, 2003.

[14] R. Teunen, B. Shahshahani, L.P. Heck. "A model-based transformational approach to robust speaker recognition", Proc. ICSLP 2000(2), 495-498, 2000.

[15] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel. "A study of inter-speaker variability in speaker verification", IEEE Trans. Audio Speech and Language Processing Vol 16, No 5, 2008.

[16] L. R. Rabiner. "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE Vol 77, No 2, 1989.

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

[18] M. Schmidt, H. Gish. "Speaker identification via support vector classifiers", Proc. ICASSP 1996, Vol 1, 1996.

[19] W. M. Campbell. "A SVM/HMM system for speaker recognition", Proc. IEEE ICASSP 2001 Vol 1, 2001.

[20] S. Furui. "Cepstral analysis techniques for automatic speaker verification", IEEE Trans. Acoust. Speech, No 29, 1981.

[21] A. Sankar, R. J. Mammone. “Growing and pruning neural tree networks”, IEEE Trans. Comput., No 42, 1993.
[22] C. Esteve-Elizalde. “Reconocimiento de locutor dependiente de texto mediante adaptación de modelos ocultos de Markov fonéticos”, Proyecto de Fin de Carrera, Universidad Auntónoma de Madrid, 2007.
[23] D. A. Reynolds. “Comparison of background normalization methods for text-independent speaker verification”, Proc. EuroSpeech 1997, Vol 2, 1997.
[24] M. Hébert, D. Boies. “T-norm for text-dependentcommercial speaker verification applications: effect of lexical mismatch”, Proc. IEEE ICASSP 2005, Vol 1, 2005.
[25] K. R. Farrell. “Speaker verification with data fusion and model adaptation”, Proc. ICSLP 2002, Vol 2, 2002.
[26] C. Barras, S. Meignier, J.-L. Gauvain. “Unsupervised online adaptation for speaker verification over the telephone – Proc. Odyssey Speaker Recognition Workshop (2004)
[27] K. Wadhwa: Voice verification: technology overview and accuracy testing results”, Proc. Biometrics Conference, 2004.
[28] L. P. Heck. “On the deployment of speaker recognition for commercial applications”, Proc. Odyssey Speaker Recognition Workshop, 2004.
[29] N. Mirghafori, M. Hébert. “Parametrization of the score threshold for a text-dependent adaptive speaker verification system”, Proc. IEEE ICASSP 2004, Vol 1, 2004.
[30] T. F. Quatieri, E. Singer, R. B. Dunn, D. A. Reynolds, J. P. Campbell. “Speaker and language recognition using speech codec parameters”, Proc. EuroSpeech, 1999.
[31] L. P. Heck, Y. Konig, M. K. Sönmez, M. Weintraub. “Robustness to telephone handset distortion in speaker recognition by discriminative feature design”, Speech Commun. No 31, 2000.
[32] M. Siafarikas, T. Ganchev, N. Fakotakis, G. Kokkinakis: Overlapping wavelet packet features for speaker verification – Proc. EuroSpeech (2005)
[33] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, P. Laface. “Channel Factors compensation in model and feature domain for speaker recognition”, Proc. IEEE Odyssey, The Speaker and Language Recognition Workshop, 2006.
[34] R. Vogt, S. Sridharan. “Explicit modelling of session variability for speaker verification”, Computer Speech and Language, No 22, 2008.
[35] T. Kato, T. Shimizu. “Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence-preserving connected digit patterns”, Proc. IEEE ICASSP 2003, Vol 2, 2003.
[36] D. Boies, M. Hébert, L. P. Heck. “Study of the effect of lexical mismaths in text-dependent speaker verification”, Proc. Odyssey Speaker Recognition Workshop, 2004.
[37] M. Hébert, L. P. Heck. “Phonetic class-based speaker verification”, Proc. EuroSpeech, 2003.
[38] D. Reynolds. “Speaker identification and verification using Gaussian mixture speaker models”, Speech Commun., No 17, 1995.
[39] O.Siohan, C. –H. Lee, A. C. Surendran, Q. Li. “Background model design for flexible and portable speake verification systems”, Proc. IEEE ICASSP 1999, Vol 2, 1999.
[40] M. Hébert, N. Mirghafori. “Desperately seeking impostors: data-mining for competitive impostor testing in a text-dependent speaker verification system”, Proc. IEEE ICASP 2004, Vol 2, 2004.
[41] R. Teunen, B. Shahshahani, L. P. Heck. “A model-based transformational approach to robust speaker recognition”, Proc. ICSLP 2000, Vol 2, 2000.

[42] D. Charlet, D. Jouvet, O. Collin. “An alternative normalization scheme in HMM-based text-dependent speaker verification”, Speech Commun. No 31, 2000.
[43] L. P. Heck, N. Mirghafori. “Online unsupervised adaptation in speaker verification”, Proc. ICSLP, 2000.
[44] L. P. Heck, N. Mirghafori. “Unsupervised on-line adaptation in speaker verification: confidence-based updates and improved parameter estimation”, Proc. Adaptation in Speech Recognition, 2001.
[45] M. Hébert. “Text-Dependent Speaker Recognition”, en “Handbook of Speech Processing”, Benesty, Sondhi y Huang (Eds.), Springer, Capítulo 37, pp 743-762, 2008.
[46] N. Mirghafori, M. Hébert. “Parametrization of the score threshold for a text-dependent adaptive speaker verification system”, Proc. IEEE ICASSP 2004, Vol 1, 2004.)
[47] D. Hernando, J. R. Saeta, J. Hernando. “Threshold estimation with cotinuously trained models in speaker verification”, Proc. Odyssey Speaker Recognition Workshop, 2006.
[48] T. Matsui, T. Nishitani, S. Furui. “Robust methods for updating model and a priori threshold in speaker verification”, Proc. IEEE ICASSP, 1996.
[49] D. Genoud, G. Chollet. “Deliberate imposture: a challenge for automatic speaker verification systems”, Proc. EuroSpeech, 1999.
[50] D. Matrouf, J. –F. Bonastre, C. Fredouille. “Effect of speech transformation on impostor acceptance”, Proc. IEEE ICASSP, 2006.
[51] Huang, X, A. Acero, H-W Hon, “Spoken Language Processing - A Guide to Theory, Algorithm and System Development”, Prentice Hall, 2001.
[52] J. Fierrez, J. Ortega-Garcia, D. T. Toledano, J. Gonzalez-Rodriguez, “Biosec baseline corpus: A multimodal biometric database”, Pattern Recognition, vol 40, no 4, Abril 2007.
[53] J. P. Campbell, “Testing with the YOHO CD-ROM voice verification corpus”, Proc. ICASSP 1995, vol 1, pp 341 –344, 1995.
[54] A. Subramanya, Z. Zhang, A. C. Surendran, P. Nguyen, M. Narasimhan, A. Acero. “A generative-discriminative framework using ensemble methods for text-dependent speaker verification”, Proc. IEEE ICASSP 2007, Vol 4, 2007.

Glosario

AFIS	Automatic Fingerprint Identification System
ASIS	Automatic Speech Identification System
DCT	Discrete Cosine Transform
DET	Detection Error Tradeoff
DTW	Dynamic Time Warping
EER	Equal Error Rate
EM	Expectation-Maximization
FA	Falsa Aceptación
FFT	Fast Fourier Transform
FR	Falso Rechazo
FVC	Fingerprint Verification Competition
GMM	Gaussian Mixture Model
H-norm	Handset Normalization
HMM	Hidden Markov Model
ICE	Iris Challenge Evaluation
LPCC	Linear Predictive Cepstral Coefficients
LRT	Likelihood Ratio Test
MAP	Maximum A Posteriori
MFCC	Mel-Frequency Cepstral Coefficients
MLLR	Maximum Likelihood Linear Regression
NIST	National Institute of Standards and Technology
ROC	Receiver Operating Curve
SRE	Speaker Recognition Evaluation
SVM	Support Vector Machine
T-norm	Test Normalization
TIC	Tecnologías de la Información y Comunicaciones
TTS	Text To Speech
UBM	Universal Background Model
VRS	Variable Rate Smoothing
Z-norm	Zero Normalization

Anexos

A Publicaciones en congresos internacionales

BioSec Multimodal Biometric Database and its use in Text-Dependent Speaker Recognition Research, publicado en LREC 2008.

MAP and Sub-Word Level T-Norm for Text-Dependent Speaker Recognition, publicado en Interspeech 2008.

BioSec Multimodal Biometric Database in Text-Dependent Speaker Recognition

Doroteo T. Toledano, D. Hernandez-Lopez, C. Esteve-Elizalde, J. Fierrez, J. Ortega-Garcia, D. Ramos and J. Gonzalez-Rodriguez

ATVS Biometric Recognition Group, Universidad Autónoma de Madrid.
Escuela Politécnica Superior, C/ Francisco Tomás y Valiente, 11, 28049 Madrid, SPAIN.
doroteo.torre@uam.es, d.hernandezlopez@uam.es, cristina.esteve@uam.es, julian.fierrez@uam.es, javier.ortega@uam.es, daniel.ramos@uam.es, joaquin.gonzalez@uam.es

Abstract

In this paper we briefly describe the BioSec multimodal biometric database and analyze its use in automatic text-dependent speaker recognition research. The paper is structured into four parts: a short introduction to the problem of text-dependent speaker recognition; a brief review of other existing databases, including monomodal text-dependent speaker recognition databases and multimodal biometric recognition databases; a description of the BioSec database; and, finally, an experimental section in which speaker recognition results on BioSec and other database widely used in speaker recognition are presented and compared, using the same underlying speaker recognition technique in all cases.

1. Introduction to text-dependent speaker recognition

Automatic speaker recognition tries to recognize the speaker that produces a particular speech utterance. Depending on the constraints imposed on the linguistic content of the utterance there are two types of speaker recognition: text-independent speaker recognition in which the linguistic content of the speech recording is unknown by the system and text-dependent speaker recognition where the linguistic content of the speech is known.

In recent years the National Institute of Standards and Technology (NIST) has promoted research in the context of text-independent speaker recognition with the organization of yearly international competitive evaluations (NIST, 2008; Przybocki, Martin & Le, 2006) which have fostered the definition of challenging tasks through a strong effort in the development of publicly available speech databases. Despite its potential applications in interactive voice response systems, the absence of similar competitive evaluations has kept text-dependent speaker recognition at a slower pace of development and the number and extent of the databases for research in this field is more limited. For that reason BioSec is an important contribution in this area.

In the field of text-dependent speaker recognition there are two methods that have been used for years: Dynamic Time Warping (DTW) and Hidden Markov Models (HMMs). DTW is simpler, but less flexible (Ramasubramanian, Das & Kumar, 2006). HMMs on the other hand are more complex, provide more flexibility and at least comparable results, and are the most commonly used technique in text-dependent speaker recognition (Hébert, 2008; Matsui & Furui, 1993; Che, Lin & Yuk, 1996; Bimbot et al., 1997).

Most of the works previously reported for text-dependent speaker recognition using HMMs tend to use a speaker independent set of HMMs and retrain the parameters of these HMMs using Baum-Welch reestimation to produce a speaker-dependent set of HMMs. After these models have been trained, an utterance is verified by performing speech recognition with the speaker independent and the speaker-dependent HMMs and comparing the acoustic scores obtained. Recently other works in the literature (Subramanya et al., 2007; Toledano et al., 2008) have started to modify this method by substituting Baum-Welch retraining by Maximum Likelihood Linear Regression (MLLR) adaptation (Leggetter & Woodland, 1995) of the speaker independent HMMs. This allows to use more complex (and, if properly trained, more reliable) HMMs while keeping the speaker models small (since only the MLLR transformation matrices need to be stored). This is the basic methodology that we have used for the comparison of text-dependent recognition results in this paper. We have avoided using here recent improvements in text-dependent speaker recognition, such as the use of discriminative methods after the MLLR adaptation (Subramanya et al., 2007) or phoneme or state-based T-Normalization (Toledano et al., 2008) because our main interest in this paper is the comparison of different databases for speaker recognition research. Therefore, we preferred to keep our speaker recognition system simple, yet still in line with the current state of the art in speaker recognition research.

2. Other databases for text-dependent speaker recognition

In this section we present several other databases for text-dependent speaker recognition research grouped into two broad categories: unimodal and multimodal databases.

2.1 Other unimodal databases for text-dependent speaker recognition

For years YOHO (Campbell & Higgins, 1994; Campbell, 1995) has been the best known database for evaluation of text-dependent speaker recognition. It consists of 96 utterances for enrolment collected in 4 different sessions

and 40 utterances for test (10 sessions) for each of 138 speakers. Each utterance consists of different combinations of three pairs of digits (e.g. “12-34-56”) in English. However, YOHO has several limitations that more modern corpora try to address. For instance, the MIT Mobile Device Speaker Verification Corpus (Woo, Park and Hazen, 2006) has been specifically designed for research on text-dependent speaker verification on realistic noisy conditions.

2.2 Other multimodal databases for biometric recognition

Due to the increasing interest in multimodal biometric recognition (of which text-dependent speaker recognition is just a particular modality), and given that one of the main difficulties in capturing a biometric database is recruiting donors, many of the newly developed biometric databases are multimodal and cover several biometric traits. Some of these databases include speech as a particular modality and can potentially be used for text-dependent speaker recognition research.

Some of the most veteran and widely used biometric databases are XM2VTS (Messer et al., 1999) containing microphone speech and face images of 295 people captured in 4 different sessions, and MCYT (Ortega-Garcia et al., 2003) database including fingerprints and signature of 330 subjects. More recent databases include BIOMET (Garcia-Salicetti et al., 2003), BANCA (Bailly-Bailliere et al., 2003), MYIDEA (Dumas et al., 2005), MBioID (Dessimoz et al., 2007), and M3 (Meng et al., 2006). Other current initiatives in multimodal database collection closely related to the BioSec database are the following (Faundez-Zanuy et al. 2006; Flynn, 2007):

- BiosecurID. This database includes 7 unimodal biometric traits, namely: speech, iris, face, handwriting, fingerprints, hand and keystroking. The database comprises 400 subjects and was acquired in a realistic office-like scenario.

- BioSecure (BioSecure, 2007). This database considers three acquisition scenarios, namely: unsupervised Internet acquisition, including voice, and face; supervised office-like scenario, including voice, finger prints, face, iris, signature and hand; and acquisition in a mobile device, including signature, fingerprints, voice, and face. The database comprises over 1000 subjects for the Internet scenario, and about 700 users the other two.

3. The BioSec database

The BioSec database was acquired under FP6 EU BioSec Integrated Project (Fierrez-Aguilar et al., 2007), and comprises fingerprint images acquired with three different sensors, frontal face images from a webcam, iris images, and voice utterances of 250 subjects.
The speech part of the corpus (the most interesting part for this paper) was recorded at 44 KHz stereo with 16 bits (PCM with no compression) using both a headset and a distant webcam microphone. Each subject utters 4 repetitions of a user-specific keyword consisting of 8 digits both in English and Spanish. Speakers are mainly native Spanish speakers. In addition, every subject says 3 keywords corresponding to other users to simulate informed forgeries in which an impostor has access to the number of a client. The 8 digits were always pronounced digit-by-digit in a single continuous and fluent utterance.

In addition to the increased number of subjects and a more balanced distribution of donors, the BioSec database has several advantages with respect to other well known databases such as YOHO. For instance it allows the simulation of informed forgeries. The BioSec database also allows studies based on age and the combination of BioSec, BiosecurID and BioSecure allows long term (2 year) temporal variability studies, because they have some subjects in common.

4. Experimental results

Text-Dependent speaker recognition experiments have been performed on YOHO and BioSec Baseline using exactly the same techniques to compare the two databases for experimentation in Text-Dependent speaker recognition. One particularity of these experiments is that in all trials the text spoken coincides with the text expected by the system. In this sense, the experiments are more representative of text-prompted systems in which the system asks the user to utter a specific phrase. In all cases the technique used for speaker recognition has been the following: we start with a set of speaker-independent phonetic HMMs that were trained on TIMIT (for English) or ALBAYZIN (for Spanish). Using the enrolment data we adapt (with MLLR) these models to produce speaker-adapted HMMs. We have also tried reestimation

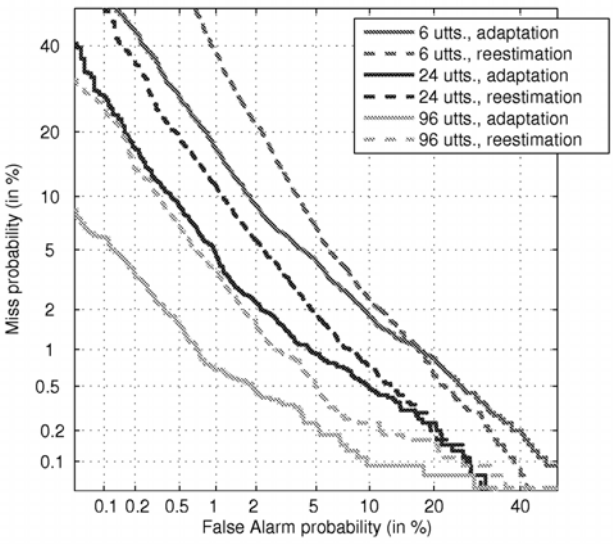


Figure 1: Results (DET curves) obtained on YOHO using MLLR adaptation and Baum-Welch re-estimation using as enrolment material 6, 24 or 96 utterances.

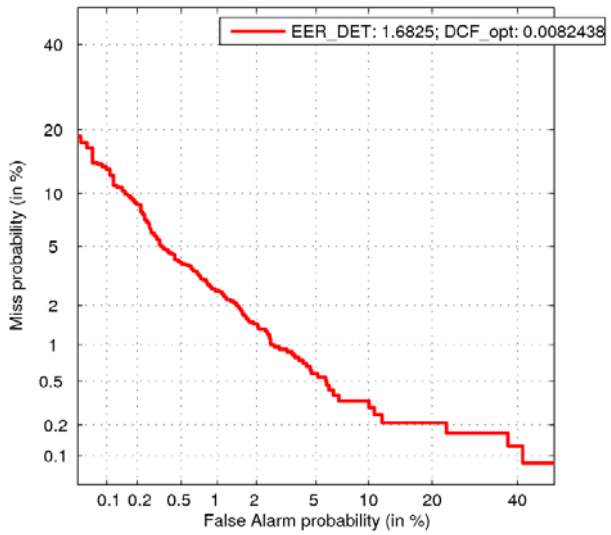


Figure 2: Results (DET curves) obtained on BioSec using MLLR adaptation for Spanish and close-talking microphone.

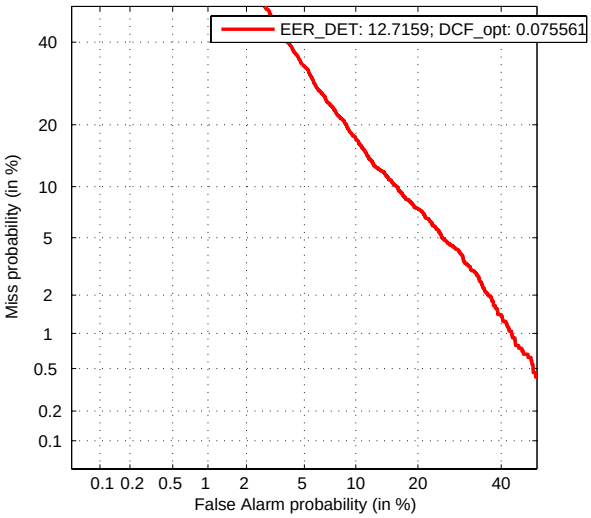


Figure 3: Results (DET curves) obtained on BioSec using MLLR adaptation for English and webcam distant microphone.

with Baum-Welch instead of MLLR adaptation in YOHO, but results were worse, as can be seen in Fig. 1. In the speaker-verification phase we subtract the (log) acoustic scores obtained by the speaker-adapted and the speaker-independent HMMs to obtain a verification score that is more positive to indicate a closer match.

4.1 Results with YOHO

The results presented on YOHO are based on the following experimental protocol: three sets of speaker models are trained using 6 utterances from session 1, the 24 utterances from session 1 or the 96 utterances from the 4 sessions. Speaker verification is performed using a single utterance from the test subset. The target scores are generated by matching each speaker-dependent phone HMM with all the test utterances from that user, leading to a total of 138 x 40 = 5520 scores. The impostor scores are computed by comparing each speaker model with a single utterance randomly selected from those of all other users, which yields 138 x 137 = 18906 trials. For all impostor trials speech is aligned against the actual phonetic content spoken to simulate a text-prompted system in which the impostors know what they have to say. Results obtained with this experimental protocol are presented in Figure 1.

As commented earlier, Fig. 1 shows that for all conditions tested MLLR adaptation in superior to Baum-Welch reestimation. The other important observation is the influence of the amount of enrolment material in performance. It can be seen that using the 96 available utterances for enrolment gives an EER under 1%, while for 6 utterances (which would be much more user-friendly) the EER increases to close to 5%. This result, however, can be lowered to about 3% using score normalization techniques not used in this paper (Toledano et al., 2008).

4.2 Results with BioSec

For these experiments we have considered two subsets of BioSec Baseline (a subset of the BioSec database comprising 2 acquisition sessions from 200 subjects), employing only those utterances that were spoken in Spanish and were captured by the headset microphone and the sentences spoken in English and captured by the webcam microphone. The experimental protocol we have followed is based on the BioSec Baseline core protocol over the specified 150 test subjects so that our results can be easily compared to other results on this corpus (Fierrez-Aguilar et al., 2005). The genuine matchings in this database were performed comparing each of the 4 samples in the first session with the 4 samples from the same user in the second session. This makes a total of 150 x 4 x 4 = 2400 target scores. To generate the impostor matchings, the first sample from the first session was tested against the same sample from the rest of the users, without performing symmetric matches. This leads to a total number of 150 x 149 / 2 = 11175 impostor scores. Results obtained with this experimental protocol are presented in Figures 2-3.

The first surprising fact is that EER in Fig. 2, where we use a single utterance for enrolment, is below 2% while for YOHO using 6 utterances for enrolment the EER is close to 5%. The reason for this surprising performance is the lexical content of the enrolment and test materials: in BioSec the lexical content of the enrolment and target trials is the same (a fixed password assigned to each user), while in YOHO the lexical content differs. Other interesting observation is the huge difference between the curves in Figures 2 and 3. There are two possible causes (which we are currently investigating) for this difference: the channel mismatch (close talking vs. distant webcam microphone) and the non-nativeness of most subjects in English in BioSec.

5. Conclusions

It is usual in research articles to use a database as test bed and compare different algorithms on that database. In text-dependent speaker recognition it has been mainly YOHO the corpus that has served for this purpose. However, YOHO has several limitations that more modern databases overcome. In this sense, researchers willing to use more modern and ample databases can be retracted from using them in order to be able to compare their results to those of other researchers. In this context, it is necessary to have a way of comparing results across different databases. This paper is an attempt to facilitate the use of the BioSec corpora by providing a comparison of text-dependent speaker recognition results across YOHO and BioSec, using exactly the same algorithms and analyizing some of the differences observed in performance on the two databases.

6. Acknowledgements

This work was funded by the Spanish Ministry of Science and Technology under project TEC2006-13170-C02-01.

7. References

Bailly-Bailliere, E., Bengio, S., et al. (2003). The BANCA database and evaluation protocol. In: Proc. of IAPR AVBPA, Springer LNCS-2688, 625-638.

Bimbot F., Hutter H. P., et al. (1997). “Speaker verification in the telephone network: research activities in the CAVE project”, in Proc. Eurospeech 1997, pp. 971-974.

BioSecure (2007). Biometrics for Secure Authentication, FP6 Network of Excellence (NoE), IST-2002-507634. (http://www.biosecure.info/).

Campbell J. and Higgins A. (1994). YOHO speaker verification corpus LDC94s16). Available at the LDC website: http://www.ldc.upenn.edu.

Campbell J. P. (1995). “Testing with the YOHO CD-ROM voice verification corpus”, in Proc. ICASSP 1995, vol. 1, pp. 341-344.

Che C.-W., Lin Q. and Yuk D.-S. (1996). “An HMM approach to text-prompted speaker verification”, in Proc. ICASSP 1996, vol. 2, pp. 673-676.

Dessimoz, D., Richiardi, J., et al. (2007). Multimodal biometrics for identity documents (MBioID). Forensic Science International 167, 154-159.

Dumas, B., Hennebert, J., et al. (2005). MyIdea - Sensors speci¯cations and acquisition protocol. Computer Science Department Research Report DIUF-RR 2005.01, University de Fribourg in Switzerland.

Faundez-Zanuy M., Fierrez-Aguilar J., Ortega-Garcia J. and Gonzalez-Rodriguez J. (2006). "Multimodal biometric databases: An overview", IEEE Aerospace and Electronic Systems Magazine, Vol. 21, n. 8, pp. 29-37, August 2006.

Fierrez-Aguilar J. and Ortega-Garcia J. (2005). “Extended Multimodal Database and Testing Protocol”, Deliverable D5.7, BioSec, FP6 IP IST-2002-001766, December 2005.

Fierrez, J., Ortega-Garcia, J., et al. (2007). Biosec baseline corpus: a multimodal biometric database. Pattern Recognition 40, 1389-1392.

Flynn P. J. (2007). “Biometric databases”, chapter in A. K. Jain, P. Flynn, A. A. Ross (Eds.), Handbook of Biometrics, Springer, 2007.

Garcia-Salicetti, S., Beumier, C., et al. (2003). BIOMET: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. In: Proc. of IAPR AVBPA, Springer LNCS-2688 845-853.

Hébert, M. (2008), “Text-Dependent Speaker Recognition”, chapter 37 in Benesty, Sondhi and Huang (Eds.) “Handbook of Speech Processing”, Springer.

Leggetter C. J. and Woodland P. C. (1995). “Flexible speaker adaptation using maximum likelihood linear regression”, in Proc. Eurospeech 1995, pp. 1155-1158.

Matsui T. and Furui S. (1993). “Concatenated phoneme models for text-variable speaker recognition”, in Proc. ICASSP 1993, vol. 2, pp. 391-394.

Meng, H., Ching, P.C., et al. (2006). The multi-biometric, multi-device and multilingual (M3) corpus. In: Proc. MMUA Workshop.

Messer, K., Matas, J., et al. (1999). XM2VTSDB: The extended M2VTS database. In: Proc. of IAPR AVBPA.

NIST (2008). National Institute of Standards and Technology. Speaker Recognition Evaluation Home Page, http://www.nist.gov/speech/tests/spk/index.htm, (accessed Feb. 2008).

Ortega-Garcia, J., Fierrez-Aguilar, J., et al. (2003): MCYT baseline corpus: a bimodal biometric database. IEE Proc. VISP 150, 391-401.

Przybocki M. A., Martin A. F., and Le A. N. (2006). “NIST speaker recognition evaluation chronicles part 2”, in Proc. IEEE Odyssey 2006: The speaker and language recognition workshop.

Ramasubramanian V., Das A. and Kumar V. P. (2006). “Text-dependent speaker recognition using one-pass dynamic programming algorithm”, in Proc. ICASSP 2006, vol. 1, pp. 901-904.

Subramanya, A.; Zhengyou Zhang; Surendran, A.C.; Nguyen, P.; Narasimhan, M.; Acero, A. (2007). “A Generative-Discriminative Framework using Ensemble Methods for Text-Dependent Speaker Verification” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. Volume 4, 15-20 April 2007 Page(s): IV-225 - IV-228.

Toledano D. T., Esteve-Elizande C., Gonzalez-Rodriguez J., Fernandez-Pozo R. and Hernandez-Gomez L. (2008). “Phoneme and Sub-Phoneme T-Normalization for Text-Dependent Speaker Recognition”, in Proc. IEEE Speaker and Language Recognition Workshop (Odyssey) 2008.

Woo R. H., Park A. and Hazen T. J. (2006). “The MIT mobile device speaker verification corpus: data collection and preliminary experiments”, in Proc. IEEE Odyssey 2006: The speaker and language recognition workshop.

MAP and Sub-Word Level T-Norm for Text-Dependent Speaker Recognition

Doroteo T. Toledano[1], Daniel Hernandez-Lopez[1], Cristina Esteve-Elizalde[1], Joaquin Gonzalez-Rodriguez[1], Ruben Fernandez Pozo[2] and Luis Hernandez Gomez[2]

[1] ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
[2] GAPS, SSR, Universidad Politecnica de Madrid, Spain

doroteo.torre@uam.es

Abstract

This paper presents improvements in text-dependent speaker recognition based on the use of Maximum A Posteriori (MAP) adaptation of Hidden Markov Models and the use of new sub-word level T-Normalization procedures. Results on the YOHO corpus show that the use of MAP adaptation provides a relative improvement of 22.6% in Equal Error Rate (EER) in comparison with Baum-Welch retraining and Maximum Likelihood Linear Regression (MLLR) adaptation. The newly proposed sub-word level T-Normalization procedures provide additional relative improvements, particularly for small cohorts, of up to 20% in EER in comparison with the normal utterance-level T-Normalization.
Index Terms: speaker recognition, text-dependent.

1. Introduction

Automatic Speaker Recognition (SR) aims to recognize the speaker that produces a particular speech utterance. It can be either text-independent or text-dependent depending on whether the linguistic content of the test speech utterance is unknown or known by the system. In the latter case the text can be a password set by the user or a random text prompted to the user (text-prompted). Despite its potential applications in interactive voice response systems, text-dependent SR has developed at a slower pace than text-independent SR, probably due to the lack of competitive evaluation campaigns such as NIST text-independent SR evaluations [1].

The most widely used modeling technique in text-dependent SR is Hidden Markov Models (HMMs) [2, 3, 4]. This paper also focuses on text-dependent SR using HMMs. Our previous work [5] compared Baum-Welch retraining versus Maximum Likelihood Linear Regression (MLLR) adaptation [6] for training the speaker models. In this paper we extend this comparison to the use of Maximum A Posteriori (MAP) [7] adaptation of the HMMs as a better way for obtaining the speaker models.

Besides this comparison, the other novelties in this paper are two new T-Norm procedures particularly designed for its use in text-dependent SR and an extensive experimentation with them. The main idea behind these new T-Norm procedures is to perform T-Norm on scores computed on smaller segments of speech (such as phonemes or HMM states) so that the averaging of the scores over the full utterance is performed on already normalized scores. This idea contrasts with the normal way of applying T-Norm in which first scores are averaged over the whole utterance and T-Norm is applied afterwards to these utterance-level scores. We call this normal way of T-Norm *Utterance-Level T-Norm* to distinguish it from the newly proposed schemes operating at the sub-word level, which we call *Phoneme-Level T-Norm* and *State-Level T-Norm*. We introduced these T-Norm schemes in [5], where we showed that using a single cohort composed of 10 male and 10 female speakers *Utterance-Level T-Norm* actually decreased performance, while *Phoneme-Level* and *State-Level T-Norm* yielded important improvements. Although the results were quite clear, some concerns could be raised about the generality of the conclusions given that the cohort included both genders (and therefore included gender-related variance), was small, and results included same-gender and cross-gender tests. This paper tries to give answer to these concerns by extending the experimentation to the cases of using two gender-dependent cohorts of 10 and 30 speakers and a male only test using a cohort of over 100 speakers. We also try to analyze the data in more detail to get insights into the reasons for the behavior observed. For the moment, all the experiments with T-Norm shown in this paper are performed on the well-known YOHO database [8]. We are currently working on extending these experiments to other databases [9].

The use of T-Norm for text-dependent SR has received little attention until very recently [4, 5, 10]. Of particular interest for this paper is the work in [10], where the authors propose the effect of the lexical mismatch as one of the reasons for the modest performance of T-Norm in text-dependent SR. In [10] the authors propose a technique for smoothing the normalization that yields improvements. Here we present an alternative way of improving the performance of normalization, by performing T-Norm at the phoneme or sub-phoneme levels instead of at the utterance level. This method, does not solve the problem of the lexical mismatch in the speech used in the enrollment of the models and in the utterance to verify, but we consider that by reducing the amount of the lexical content of the test segment used to compute the score before applying T-Norm to one phoneme or sub-phoneme the problem could be somewhat alleviated.

The rest of the paper is organized as follows: section 2 describes briefly the baseline algorithm used for text-dependent SR with HMMs. Section 3 describes the three different alternatives considered for performing T-Norm, section 4 describes the experimental protocol, section 5 presents experimental results, section 6 presents a discussion on the reasons for the behavior observed in the experiments, and finally, section 7 presents conclusions and future work.

2. General framework for text-dependent SR based on phonetic HMMs

The general framework used in this paper for text-dependent SR is defined by a common parameterization; a speaker-dependent *sentence* model of the utterance to be verified, a speaker-independent *sentence* model and a common way of scoring. This general framework is described in detail in [5], so we refer the interested reader to this article and will give here just a brief summary.

The front end starts with a pre-emphasis filter, after which the signal is windowed using 25 ms. Hamming windows with a window shift of 10 ms. From each window 13 Mel Frequency Cepstral Coefficients (MFCCs) are extracted (including C0), and their first and second-order differences are calculated, for a total of 39 features per frame.

A speaker-independent sentence model is built for each utterance to verify from a set of speaker-independent phonetic HMMs, a phonetic lexicon and the orthographic transcription of the sentence. The HMMs are 39 context-independent English phonetic HMM models previously trained on TIMIT. The phonetic models have 3 states, with a Bakis (left-to-right) topology with no skips.

This model will compete against a speaker-dependent sentence model built exactly in the same way but using speaker-dependent phonetic HMMs obtained from a small amount of speech (enrollment data) from that speaker. These speaker-dependent phonetic HMMs have exactly the same structure as the speaker-independent HMMs and can be obtained in different ways. We have explored three of them: performing Baum-Welch reestimation [11] of the speaker-independent phonetic HMMs on the enrollment data, adapting the speaker-independent HMMs using MLLR [6], and finally performing MLLR followed by MAP adaptation [7].

After the speaker-independent and the speaker-dependent models of the utterance have been built the utterance to verify is aligned to each of these two models using a Viterbi algorithm which produces the acoustic scores for each frame given the speaker-dependent and the speaker-independent models of the utterance. The final score is the ratio between the average score per frame obtained with the speaker-dependent model and the average score per frame obtained with the speaker-independent model. Assuming that the textual content of the utterance is the correct, the larger the score the larger the confidence the system has in verifying the speaker. This set-up models a text-prompted system where the text uttered normally coincides with the expected text.

In spite of the score normalization provided by the use of speaker-independent scores, which can be viewed as similar to a UBM (Universal Background Model), the speaker-dependent score variations and the need for speaker-independent decision thresholds usually requires the inclusion of further score normalization techniques (Z-norm, T-norm, …). In this sense we will consider that the scores obtained as described in this section are unnormalized scores. In next section we will describe three different ways to perform T-norm in this context.

3. T-Norm for text-dependent SR at the utterance, phoneme and state levels

In text-independent SR it is very common to use T-Normalization by comparing the score obtained with a test segment, not only to the model of the speaker in the test segment, but also against the models of other speakers (i.e. against a cohort of impostors).

The direct translation of this approach to text-dependent SR is what we call *Utterance-Level T-Norm*, to distinguish it from the novel T-Normalization schemes proposed in following sections. As with any T-Normalization scheme, we need to define a cohort of M speakers and compute the unnormalized scores (as described in Section 2) not only using the model of the speaker to verify but also the models for the M speakers in the cohort. After we have done this we T-Normalize the score in the usual way:

$$sc^{TNorm} = \frac{sc - \mu}{\sigma}, \qquad (11)$$

Where sc is the unnormalized score, μ and σ are the mean and the standard deviation of the scores obtained against the cohort of M speakers and sc^{TNorm} is the T-Normalized score.

With this T-Normalization scheme we T-Normalize the final scores after averaging over the whole utterance. In this sense, we are combining scores computed on very different parts of the test utterance (i.e. on different phonemes or different parts of the phonemes) which may produce scores with very different distributions. For that reason it seems to be a good idea to try to normalize the scores for similar segments before averaging the scores. We propose the use of sub-word level T-Normalization schemes in which we perform T-Normalization on averages of the acoustic scores over segments corresponding to phonemes or even HMM states within the phoneme before averaging the already T-Normalized scores over the whole utterance. We call these methods *Phoneme-Level T-Normalization* and *State-Level T-Normalization*. The idea behind these new T-Normalization schemes is relatively simple and we consider that a detailed description here is unnecessary. However, the interested reader can find a detailed description of these methods in [5].

4. YOHO experimental protocol

For the experiments we have used YOHO [3], probably the most widely used and well known benchmark for text-dependent SR system comparison and assessment. It consists of 96 utterances for enrollment collected in 4 different sessions and 40 utterances for testing collected in 10 sessions for each of a total of 138 speakers, 106 male and 32 female. Each utterance is a different set of three digit pairs (e.g. “12-34-56”). The results presented on YOHO are based on the following experimental protocol. Speaker models are trained using 6 utterances from session 1, the 24 utterances from session 1 or the 96 utterances from the 4 sessions. Our main focus was on the single session, 6 utterances, since it is the closest to what we expect to find in realistic operational conditions. Most experiments are referred to this condition. Speaker verification is performed using a single utterance from the test subset. The target scores are generated by matching each speaker model with all the test utterances from that user, leading to a total of 138 x 40 = 5,520 scores. The impostor scores are computed by comparing each speaker model with a single utterance randomly selected from those of all other users, which yields 138 x 137 = 18,906 trials. For all impostor trials the sentence models are produced using the actual text spoken to simulate a text-prompted system in which the impostors know what they have to say.

For experiments using T-Norm the experimental protocol has been slightly modified. We have considered 3 different cohort sizes for T-norm: 10 male and female speakers, 30 male and female speakers (this is the maximum we can reach with the 32 female speakers in YOHO) and all male speakers. For the 10 male and 10 female cohorts we have removed these speakers from the test. This way the number of target scores is reduced to 118 x 40 = 4,720, and the number of impostor scores to 118 x 117 = 13,806. For the 30 male and 30 female speakers and for the all male cohorts we cannot remove so many speakers from the test, so we have used Jackknife to use all trials and large (trial-dependent) cohorts with speakers not included in each trial.

5. Results

We have organized this section into three subsections. The first one compares results without score normalization using Baum-Welch and MLLR. The second presents results without normalization and with MAP. Finally, the third one focuses on the three proposed ways of performing T-Normalization, comparing them using several set-ups for the cohort.

5.1. Results with Baum-Welch and MLLR

In this section we compare MLLR adaptation and Baum-Welch re-estimation for different amounts of enrollment speech. In particular, we have compared the best results achieved by MLLR adaptation and Baum-Welch retraining for the condition of 6 utterances from the first training session, 24 utterances from the first training session, and of all 96 utterances in the 4 training sessions. Table 1 and Figure 1 show the best results obtained after an optimization performed on the number of Gaussians per state, the number of iterations of Baum-Welch re-estimation and the number of regression classes in MLLR adaptation. For Baum-Welch re-estimation the number of Gaussians per state was varied between 1 and 5 and the number of re-estimation iterations was either 1 or 4. For MLLR adaptation the number of Gaussians per state was varied between 5 and 80 in steps of 5 and the number of regression classes between 1 and 32 in power-of-2 steps. Our best results show that, even in the cases with the largest amount of data, MLLR adaptation outperforms Baum-Welch re-estimation in text-dependent speaker recognition. In fact, the difference in favour of MLLR tends to increase as the amount of enrollment material increases. The reason for this may be that the amount of enrollment material, even using the 96 utterances for training, is still very limited for Baum-Welch re-estimation. MLLR adaptation seems to be more adequate for the whole range of enrollment speech considered.

5.2. Results with MLLR plus MAP

After these experiments we tried to get more accurately speaker-adapted HMMs by performing MAP [7] adaptation after the MLLR adaptation. This yields increased speaker recognition performance (Fig. 1 and Table 1). The EER decreased by 1.04% absolute (22.6% relative improvement). This improvement comes at increased computational and storage costs (we need to store a whole new set of phonetic HMMs for each speaker, not only the transformation matrices) but in some applications we can take advantage of it. We have only performed experiments with MLLR followed by MAP for the 6 utterances enrollment condition because this is the most interesting condition for the applications we are considering currently.

5.3. Results with Utterance-Level, Phoneme-Level and State-Level T-Norm

In this section we make use of the method that produced the best results in the former sections, adaptation with MLLR followed by MAP, and focus on user enrollment with 6 utterances, which we consider the case most close to the applications we envisage. With these settings we have tested the three different schemes for T-Normalization described in section 3 with different set-ups of the cohort. Results from this extensive testing are summarized in terms of Equal Error Rate (EER) in percentage in Table 2.

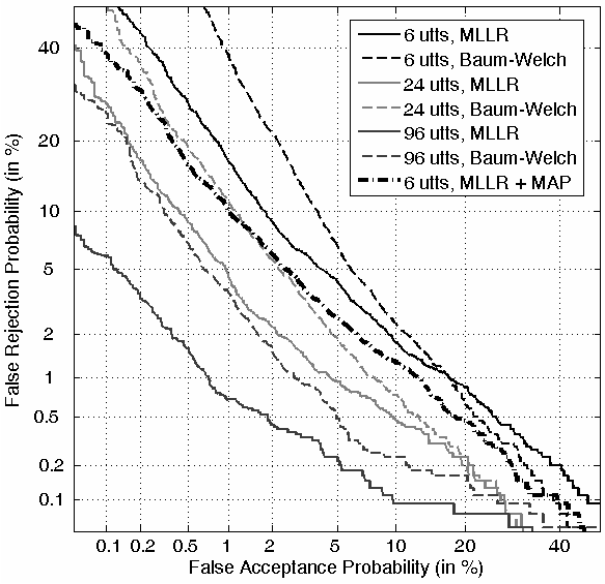


Figure 1: *DET curves with Baum-Welch re-estimation, MLLR adaptation and MLLR adaptation followed by MAP with 6, 24 and 96 utterances for enrollment.*

Table 1. *EERs (%) with Baum-Welch re-estimation, MLLR adaptation and MLLR adaptation followed by MAP with 6, 24 and 96 utterances for enrollment.*

Enrollment utterances (and sessions)	Baum-Welch	MLLR	MLLR + MAP
6 (1 session)	5,6	4,6	3,56
24 (1 session)	3,2	2,1	--
96 (4 sessions)	1,9	0,9	--

The first line of Table 2 presents results obtained with MLLR plus MAP adaptation without normalization, and serves as the baseline results. These correspond to Figure 1 but have been further detailed according to the gender in the trials. The last column of the table presents global results obtained by considering all trials, including same gender and cross gender trials.

The rest of the table is organized in blocks of three lines which represent results obtained with *Utterance-Level*, *Phoneme-Level* and *State-Level T-Norm* for the following cohorts of impostors:

- *G.I. 10m+10f:* A gender independent cohort including 10 male speakers and 10 female speakers.
- *G.D. 10m – 10f:* Two gender dependent cohorts obtained by dividing the previous cohort into two gender-dependent cohorts.
- *G.D. 30m – 30f:* Two gender-dependent cohorts with 30 speakers for each gender.
- *G.D. All male:* A male cohort including all speakers in YOHO except those involved in the trial.

For the two first cases we removed the speakers in the cohort from the test, while for the two last we used Jackknife and trial-dependent cohorts excluding speakers in the trial.

From the table we observe that *Phoneme-Level* and *State-Level T-Norm* clearly outperform *Utterance-Level T-Norm* for the smaller cohorts (10 male and 10 female), irrespective of whether the cohorts are gender-dependent or independent. In

Table 2. *T-Norm results (EERs in %) obtained on YOHO (with only 6 utterances from a single session as enrollment material) using MLLR and MAP adaptation. The table compares results obtained without normalization and with Utterance-Level, Phoneme-Level and State-Level T-Norm for different set-ups for the cohort.*

Cohort	Type of T-Norm	Gender Condition		
		Male	Female	All
NO	NO	3.90	7.26	3.56
G.I. 10m + 10f	Utterance	4.13	5.84	3.91
	Phoneme	3.21	4.76	2.98
	State	3.34	4.55	3.04
G.D. 10m – 10f	Utterance	3.53	13.85	3.64
	Phoneme	2.92	5.19	2.97
	State	3.02	4.55	2.91
G.D. 30m – 30f	Utterance	2.74	4.07	3.10
	Phoneme	2.52	4.13	2.98
	State	2.47	4.03	2.96
G.D. All male	Utterance	2.55	--	--
	Phoneme	2.43	--	--
	State	2.52	--	--

these cases, *Utterance-Level T-Norm* actually worsens the results obtained without normalization, while *Phoneme* and *State-Level T-Norm* produce important improvements. In the case of two gender-dependent cohorts with 10 male and 10 female speakers the relative improvement achieved by *State-Level T-Norm* over *Utterance-Level T-Norm* reaches 20.1% (0.73% absolute) in the all gender condition.

When we move to larger cohorts we observe that *Phoneme* and *State-Level T-Norm* still tend to perform better than *Utterance-Level T-Norm*. However, the increase of the cohort has a larger improvement effect on *Utterance-Level T-Norm* than on sub-word levels T-Norm. This reduces the difference between utterance and sub-word levels T-Norm.

6. Discussion

It is reasonable to consider that the different phonemes have different discrimination capabilities. In fact, this is the hypothesis of a recent work [12] in which the scores produced by different phonemes are combined with different weights using boosting for improved performance. In the context of T-Norm this will mean that the scores produced by different phonemes should be normalized in different ways. In fact, we have studied the impostor score distributions for different phonemes (not presented here due to space limitations) and have noticed important differences among them, which again suggest the convenience of sub-word level normalizations. Our experiments in this paper, however, have made that advantages clear particularly for small cohorts, pointing out other important advantage of sub-word score normalization schemes: their robustness to small cohorts.

7. Conclusions

In this paper we have experimented with three different ways of obtaining the speaker models from the enrollment material for a text-dependent SR system based on HMMs: Baum-Welch reestimation, MLLR adaptation and MLLR followed by MAP adaptation. Among them, we have found that MLLR followed by MAP tends to produce the best results, which are over 22.6% relatively better in terms of EER than those achieved by the second best, MLLR. We have also performed an extensive experimentation with T-Normalization methods, comparing the normal method, *Utterance-Level T-Norm*, with two novel methods, *Phoneme-Level T-Norm* and *State-Level T-Norm*. Experiments have been performed with different cohort set-ups, showing that *Phoneme-Level T-Norm* and *State-Level T-Norm* tend to perform better than *Utterance-Level T-Norm*. These differences are particularly noticeable (up to 20.1% relative improvements in EER) when small cohorts are used for T-Norm, probably due to the higher robustness to small cohorts of these new sub-word T-Norm methods compared to the normal, utterance-based T-Norm.

8. Acknowledgements

This work was funded by the Spanish Ministry of Science and Technology under project TEC2006-13170-C02-01.

9. References

[1] "National institute of standard and technology. Speaker Recognition Evaluation Home Page", http://www.nist.gov/speech/tests/spk/index.htm.

[2] T. Matsui and S. Furui, "Speaker Recognition Using Concatenated Phoneme HMMs," Proc. ICSLP, Banfl, Th.sA M.4.3 (1992).

[3] F. Bimbot, H. P. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg and J. B. Pierrot, "Speaker verification in the telephone network: research activities in the CAVE project", in *Proc. Eurospeech 1997, pp. 971-974.*

[4] Hébert, M., "Text-Dependent Speaker Recognition", chapter 37 in Benesty, Sondhi and Huang (Eds.) "Handbook of Speech Processing", Springer, 2008.

[5] Toledano D. T., Esteve-Elizande C., Gonzalez-Rodriguez J., Fernandez-Pozo R. and Hernandez-Gomez L. "Phoneme and Sub-Phoneme T-Normalization for Text-Dependent Speaker Recognition", in Proc. IEEE Odyssey 2008.

[6] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression", in *Proc. Eurospeech 1995, pp. 1155-1158.*

[7] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[8] J. Campbell and A. Higgins. Yoho speaker verification (ldc94s16). http://www.ldc.upenn.edu.

[9] Toledano D. T., Hernandez-Lopez D., Esteve-Elizalde C., Fierrez J., Ortega-Garcia J., Ramos D. and Gonzalez-Rodriguez J., "BioSec Multimodal Biometric Database in Text-Dependent Speaker Recognition", in Proc. LREC 2008 (to appear).

[10] M. Hébert and D. Boies, "T-Norm for text-dependent commercial speaker verification applications: effect of lexical mismatch", in *Proc. ICASSP 2005*, pp. 729-732.

[11] L. R. Rabiner, "A Tutorial on Hidden Markov Models", In *Proceedings of the IEEE*, vol. 77, n. 2, February 1989, pp. 257-286.

[12] Subramanya, A.; Zhengyou Zhang; Surendran, A.C.; Nguyen, P.; Narasimhan, M.; Acero, A.; "A Generative-Discriminative Framework using Ensemble Methods for Text-Dependent Speaker Verification" in Proc. ICASSP, 2007. Volume 4, 15-20 April 2007 pp. IV-225 - IV-228.

B Publicaciones enviadas a congresos (a la espera de ser aceptadas)

T-Norm y desajuste Léxico y Acústico en Reconocimiento de Locutor Dependiente de Texto, enviado a las Jornadas de las Tecnologías del Habla 08.

T-NORM Y DESAJUSTE LÉXICO Y ACÚSTICO EN RECONOCIMIENTO DE LOCUTOR DEPENDIENTE DE TEXTO

Daniel Hernández López[1], Doroteo Torre Toledano[1], Cristina Esteve Elizalde[1], Joaquín González Rodríguez[1], Rubén Fernández Pozo[2] y Luis Hernández Gómez[2]

[1]ATVS Biometric Recognition Group, Universidad Autónoma de Madrid, España
[2]GAPS, SSR, Universidad Politécnica de Madrid, España

RESUMEN

Este trabajo presenta un estudio extenso sobre T-norm aplicado a Reconocimiento de Locutor Dependiente de Texto, analizando también los problemas del desajuste léxico y acústico. Veremos cómo varían los resultados teniendo en cuenta la dependencia de género y realizando T-norm a nivel de frase, fonema y estado con cohortes de impostores de distintos tamaños. El estudio demuestra que implementar T-norm por fonema o estado puede llegar a conseguir mejoras relativas de hasta un 16% y que realizar una selección de cohorte basada en el género puede mejorar más aún los resultados con respecto al caso independiente de género.

1. INTRODUCCIÓN

El Reconocimiento Automático de Locutor es una disciplina de la biometría que consiste reconocer la identidad de una persona (locutor) a través de la voz. Dentro de ésta hay dos grandes vertientes, el Reconocimiento de Locutor Independiente de Texto y el Reconocimiento de Locutor Dependiente de Texto. La segunda de ellas parece haber quedado en segundo plano comparada con la primera, muy probablemente debido a la ausencia de evaluaciones competitivas como las hay para Reconocimiento de Locutor Independiente de Texto [1].

El Reconocimiento de Locutor Dependiente de Texto tiene la particularidad de que el sistema dispone, tanto para entrenamiento como para test, de las trascripciones de la locución. Esto significa que mediante un diccionario fonético podemos disponer de la trascripción fonética de lo que se dice en la locución, lo que hace que se consigan buenos resultados con menor cantidad de habla que en Reconocimiento de Locutor Independiente de Texto. Como es habitual en este tipo de sistemas, en el nuestro utilizamos Modelos Ocultos de Harkov (HMMs) [2] para modelar las características fonéticas de los locutores. Utilizar HMMs permite tener modelos independientes de cada fonema para cada locutor, donde cada uno de los fonemas estará modelado como una serie de probabilidades de transición entre estados, y cada estado estará representado mediante un Modelo de Mezclas de Gaussianas (GMM) [3]. Con estas herramientas y disponiendo de la trascripción fonética, se puede realizar un reconocimiento fonético, utilizando el algoritmo de Viterbi, que proporcione una trascripción fonética con los instantes de comienzo y fin de cada uno de los fonemas y de sus correspondientes estados.

Esta serie de características suponen varias ventajas al Reconocimiento de Locutor Dependiente de Texto frente al Independiente de Texto, pero sin duda la mayor de ellas es poder trabajar, tanto en entrenamiento como en reconocimiento, con niveles por debajo de la frase (palabra, fonema y estado). Dicha ventaja ha sido utilizada múltiples veces en esta disciplina tanto en entrenamiento como en reconocimiento. Entonces ¿porqué no utilizarla en T-norm?

En este trabajo veremos cómo se pueden mejorar los resultados finales del sistema mediante la conocida técnica de T-norm. Hasta ahora esta técnica ha sido muy utilizada en Reconocimiento de Locutor Independiente de Texto y, aunque en menor medida, también en Reconocimiento de Locutor Dependiente de Texto. Sin embargo en todos los casos en que se ha utilizado, T-norm ha sido aplicado a la puntuación global de la locución de test. En el caso de Reconocimiento de Locutor Independiente de Texto parece lógico que se haga así, pero en el caso de Reconocimiento de Locutor Dependiente de Texto parece mejor aprovecharse de la ventaja de poder trabajar con niveles inferiores. Además estudiaremos cómo influye el género y el tamaño de la cohorte de impostores de T-norm.

El resto del artículo está organizado de la siguiente manera: en la Sección 2 describiremos el sistema del que se parte y que ha evolucionado a lo largo de los experimentos, en la Sección 3 explicamos como se implementa T-norm para los experimentos realizados, en la Sección 4 se muestran las bases de datos utilizadas para los experimentos de las Secciones 5 y 6 y por último se presentan las conclusiones en la Sección 7.

2. DESCRIPCIÓN DEL SISTEMA DE PARTIDA

Se parte de un sistema de Reconocimiento de Locutor Dependiente de Texto basado en HMMs. La parametrización que se ha aplicado al audio usado en

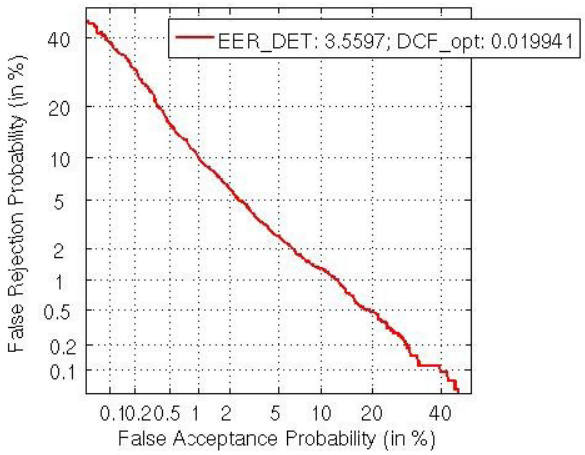


Figura 1. *Curva DET para el sistema de partida.*

entrenamiento y test se basa en la extracción de coeficientes cepstrales mediante filtros de Mel (MFCC, *Mel Frequency Cepstral Coefficients*), tomados en formato $(13+\Delta+\Delta\Delta)$. El sistema de partida realiza alineamiento no completamente forzado de la trascripción tanto para entrenamiento como para verificación (no es totalmente forzado porque se incluyen silencios opcionales entre palabras). Esto es porque aunque se tenga una trascripción textual de lo que se ha pronunciado en la locución no se sabe si hay silencio entre palabras ni de qué duración es éste. De esta forma, realizado un reconocimiento fonético con un modelo de silencio opcional se mejora el alineamiento temporal, lo cual favorece tanto a la etapa de entrenamiento como posteriormente la de reconocimiento.

Dada la trascripción fonética correctamente alineada en el tiempo y el audio parametrizado, el sistema realiza la adaptación al locutor del modelo acústico independiente del locutor. Para ello la adaptación se realiza en tres fases.

En una primera fase se adaptan las medias de las Gaussianas de los Modelos de Mezclas de Gaussianas de cada uno de los estados de cada fonema de forma global. Esto quiere decir que se adaptan todas las medias de forma conjunta. Esta adaptación se hace según el algoritmo MLLR [4] (*Maximum Likelihood Linear Regression*) de forma global, sin clases de regresión. De esta adaptación se obtiene el modelo de transformación lineal, que consiste en una matriz para transformar los modelos fonéticos independientes del locutor en modelos adaptados al locutor. De esta forma no es necesario guardar un modelo de cada locutor, sino simplemente el modelo de transformación. Posteriormente se adaptan los modelos resultantes empleando MLLR, ahora con 2 clases de regresión, obteniendo un nuevo modelo de transformación lineal. Finalmente se aplica adaptación MAP (*Maximum A Posteriori*) [4] a los modelos después de haber sido transformados con el modelo de adaptación MLLR global y posteriormente con 2 clases de regresión.

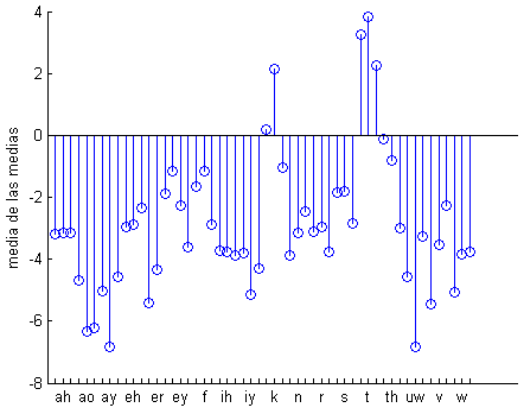


Figura 2. *Medias de las puntuaciones medias de los fonemas (compuestos por 3 estados) y estados en tests de impostores.*

Una vez adaptados los modelos se procede a realizar la fase de evaluación con intentos tanto *target* (donde la locución a verificar es del locutor representado por el modelo) como *non-target* (donde la locución a verificar es de un locutor distinto al representado por el modelo). Se obtienen puntuaciones para cada estado de cada fonema enfrentando la locución de test con el modelo adaptado al locutor y restándola de la puntuación obtenida de enfrentarla al modelo independiente del locutor.

Por último se eliminan las puntuaciones obtenidas por los silencios y se promedia la puntuación general de la locución de test. El resultado obtenido de esta forma para la base de datos YOHO (descrita en la Sección 4), se representa en forma de curva DET en la Figura 1.

3. T-NORM A DISTINTOS NIVELES

Lo que se propone en este artículo es un estudio sobre T-norm, que básicamente consiste en tomar las puntuaciones obtenidas por una cohorte de impostores y calcular la media (μ) y la desviación típica (σ) de dichas puntuaciones. De esta forma se calcula la nueva puntuación ($score_{T-norm}$) como la puntuación obtenida por el locutor que realiza el intento de acceso ($score$) menos la media, dividido entre la desviación típica, como se muestra en la Fórmula 1.

$$score_{T-norm} = \frac{score - \mu}{\sigma} \qquad (1)$$

La clave de este estudio consiste en que se realizará este proceso no sólo a nivel de locución (como suele ser habitual) sino también a nivel de fonema y de estado. Por otra parte se implementará también T-norm dependiente de género. Esto significa que la cohorte de impostores estará compuesta por locutores del mismo género que el locutor *target* para experimentos de este tipo.

La idea de realizar este tipo de T-norm surge de un estudio analítico de las puntuaciones. Como podemos observar en la Figura 2 las puntuaciones de impostor de los estados de un mismo fonema tienen una cierta correlación mientras que entre fonemas las puntuaciones son muy dispares. Esto nos induce a pensar que realizar T-norm a nivel de fonema o estado puede reportarnos buenos resultados debido a que de este modo alinearemos las puntuaciones obtenidas por cada fonema y estado, que parecen desalineadas (Fig. 2).

4. DESCRIPCIÓN DE LAS BASES DE DATOS

4.1. YOHO

Se ha usado la base de datos YOHO [5] para este experimento. Esta base de datos tiene 138 locutores, de los cuales 106 son hombres y 32 son mujeres. Cada locutor presenta 96 locuciones de entrenamiento repartidas en 4 sesiones y 40 locuciones de test repartidas en 10 sesiones. Se han utilizado 6 locuciones de entrenamiento de la primera sesión para realizar el entrenamiento y todas las locuciones de test del locutor como intentos *target* para la etapa de verificación, tomando una locución al azar de cada uno de los demás locutores como intentos *non-target*. Cabe destacar que el léxico de esta base de datos consiste en frases de pares de dígitos (p.e. 32-98-64) y que no hay ninguna relación entre los dígitos pronunciados en entrenamiento y test, con lo cual tenemos un importante desajuste léxico.

4.2. BioSec

Para otro experimento realizado se ha usado la base de datos BioSec Baseline [6]. En esta base de datos hay 150 locutores cuyo idioma nativo es el castellano. Cada locutor ha grabado 2 sesiones con 4 locuciones cada una de un número aleatorio asignado al usuario (el mismo para todas las locuciones de las 2 sesiones). Se han utilizado las 4 frases de la primera sesión para entrenar los modelos acústicos del locutor (se entrena un modelo con cada frase) y las 4 de la segunda sesión como intentos *target* para la fase de test, siendo los intentos *non-target* la primera frase de la primera sesión del resto de impostores (sin enfrentamientos simétricos). Todas las locuciones descritas anteriormente se han realizado de forma idéntica para 4 escenarios. Castellano grabado con un micrófono de unos auriculares (cercano), castellano grabado con un micrófono integrado en una webcam (lejano) y otros 2 escenarios equivalentes en inglés.

5. EXPERIMENTOS CON T-NORM EN FUNCIÓN DEL NIVEL, COHORTE Y GÉNERO

Para poder implementar T-norm, se ha realizado un reconocimiento de cada locución de test con los modelos de locutores de la cohorte de impostores por cada enfrentamiento tanto *target* como *non-target*. De esta forma se han realizado 3 tipos de T-norm en función de la cohorte de impostores: una con una cohorte fija de 20 locutores, 10 hombres y 10 mujeres, a la que llamaremos TN10; otra con una cohorte variable de 60 locutores, 30 hombres y 30 mujeres, a la que llamaremos TN30; una última con una cohorte masculina variable que incluye como impostores todos aquellos locutores que no sean ni el *target* ni el *non-target* (en el caso de que se trate de una prueba *non-target*), a la que llamaremos TNMale.

Para el caso sin T-norm y TN10 hemos realizado experimentos tanto dependientes de género como independientes de género, mientras que para TN30 y TNMale únicamente hemos realizado experimentos dependientes de género. Para los experimentos independientes de género anteriormente descritos se han obtenido los resultados expresados en EER (*Equal Error Rate*) mostrados en la Tabla 1.

T-norm\Nivel	Frase	Fonema	Estado
No	3.56		
TN10	3.91	2.98	3.04

Tabla 1. *EERs (%) obtenidas para distintos tipos de T-norm independiente de género en función del nivel.*

Como podemos ver en la Tabla 1 resulta mucho mejor realizar T-norm a nivel de estado o fonema que a nivel de frase, de hecho podemos ver que es incluso mejor no implementar T-norm que hacerlo a nivel de frase para este experimento en concreto. A continuación vemos en la Tabla 2 como evolucionan los resultados al incrementar el número de impostores de la cohorte y realizar una selección por género de la cohorte de impostores a utilizar.

T-norm	Género	Frase	Fonema	Estado
No	Masc	3.54		
	Fem	3.72		
TN10	Masc	3.32	2.64	2.80
	Fem	3.57	3.15	2.45
	Ambos	3.64	2.97	2.91
TN30	Masc	2.69	2.48	2.46
	Fem	3.99	3.67	3.67
	Ambos	3.10	2.98	2.96
TNMale		2.57	2.41	2.53

Tabla 2. *EERs (%) obtenidas para distintos tipos de T-norm en función del nivel y género.*

En la Tabla 2 vemos cómo varían las tasas de error obtenidas en función del género y el número de impostores de la cohorte. En líneas generales podemos observar que parece ser que cuanto mayor es la cohorte de impostores mejor funciona el sistema, debido probablemente a que al tener un número mayor de impostores tenemos más probabilidades de encontrarnos con modelos próximos al del locutor *target*. Sin

embargo esto no se cumple para todos los casos y es debido, muy probablemente, a que también nos encontraremos más modelos que se alejen del modelo del locutor *target*. Por otra parte también vemos que se generaliza la suposición de que es mejor realizar T-norm a nivel de fonema o estado que a nivel de frase. Además vemos que no hay mucha diferencia entre realizarlo a nivel de estado o fonema, ya que hay casos en los que resulta mejor uno que otro y viceversa.

6. OTROS EXPERIMENTOS

A fin de extender nuestra experimentación al idioma castellano realizamos también experimentos con la base de datos BioSec Baseline [7]. Esta base de datos, aparte de permitirnos comparar resultados en inglés y castellano con el mismo entorno experimental nos permite comparar la influencia del canal de grabación (micrófono de habla cercana frente a micrófono de habla lejana) y la influencia de la coincidencia léxica entre entrenamiento y test (cosa que ocurre en BioSec pero no en YOHO). Otra diferencia con los resultados presentados anteriormente es que en los resultados con BioSec se ha empleado únicamente MLLR y no MAP.

Canal\Idioma	Castellano	Inglés
Mic. cercano	1.68	2.17
Mic. lejano	17.24	12.72

Tabla 3. *EER (%) obtenidas para distintos tipos de micrófono e idioma.*

Como podemos observar en la Tabla 3 la calidad del micrófono supone una gran contribución a la eficiencia del sistema de Reconocimiento de Locutor Dependiente de Texto. Vemos que los resultados obtenidos con el micrófono de la webcam son mucho peores que con el micrófono cercano integrado en los auriculares. Se ha de indicar que en estos experimentos no se han utilizado técnicas de compensación de canal de ningún tipo (salvo CMN).

Sin desajuste	7.02
SNR	7.47
Canal	9.76
Léxico (2 dígitos en común)	8.23
Léxico (1 dígito en común)	13.4
Léxico (0 dígitos en común)	36.3

Tabla 4. *EER (%) obtenida para distintos tipos de desajuste para el estudio realizado en [8].*

Por otra parte si nos fijamos en los resultados para micrófono cercano observamos que son mucho mejores para esta base de datos que con YOHO (utilizando la misma técnica sólo con MLLR en YOHO el EER resultante es de 4.82%). La principal diferencia entre ambas bases de datos es el desajuste léxico existente en YOHO e inexistente en BioSec. El problema del desajuste léxico ya se ha analizado con anterioridad [8] y se ha comprobado (ver Tabla 4 con resultados publicados en [8]) que el desajuste léxico puede ser el tipo más perjudicial de desajuste, incluso peor que el de canal.

7. CONCLUSIONES

Dados los resultados obtenidos en los diferentes experimentos se puede concluir que el desajuste léxico tiene una gran influencia en la eficiencia de un sistema de Reconocimiento de Locutor Dependiente de Texto. La principal razón es que en este campo se entrenan modelos de unidades léxicas por debajo de la locución completa (palabra, fonema, tri-fonema, estado…). Y el hecho de que se intente reconocer al locutor con modelos de unidades léxicas que hemos podido no entrenar previamente (desajuste léxico), o que hemos entrenado en contextos léxicos distintos, hace que los resultados empeoren de forma muy abultada.

Demostrado esto, el principal objetivo de la técnica de T-Norm a nivel de fonema y estado era tratar de reducir la influencia del desajuste léxico, para así ponderar la influencia de cada fonema en el proceso de verificación. Aunque los resultados obtenidos con la normalización a nivel de estado y fonema son positivos, superando los resultados a nivel de frase, el problema del desajuste léxico sigue sin estar resuelto y sigue teniendo una influencia importante en los resultados.

8. BIBLIOGRAFÍA

[1] "National institute of standard and technology. Speaker Recognition Evaluation Home Page", http://www.nist.gov/speech/tests/sre/

[2] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", Proceedings of the IEEE, vol 77, no 2, pp. 257-286, Febrero 1989.

[3] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted gaussian mixture models", Digital Signal Processing, vol 10, no 1, pp. 19-41, Enero 2000.

[4] D. T. Toledano, D. Hernandez-Lopez, C. Esteve-Elizalde, J. Gonzalez-Rodriguez, R. Fernandez, L. Hernandez, "MAP and sub-word level T-norm for text-dependent speaker recognition", to appear in Interspeech 2008.

[5] J. P. Campbell, "Testing with the YOHO CD-ROM voice verification corpus", Proc. ICASSP, vol 1, pp. 341-344, 1995.

[6] J. Fierrez, J. Ortega-Garcia, D. T. Toledano, J. Gonzalez-Rodriguez, "Biosec baseline corpus: A multimodal biometric database", Pattern Recognition, vol 40, no 4, Abril 2007.

[7] D. T. Toledano, D. Hernandez-Lopez, C. Esteve-Elizalde, J. Fierrez, J. Ortega-Garcia, D. Ramos, J. Gonzalez-Rodriguez, "BioSec Multimodal Biometric Database in Text-Dependent Speaker Recognition", Proc. LREC, Mayo 2008.

[8] D. Boies, M. Hébert, L. P. Heck, "Study of the effect of lexical mismatch in text-dependent speaker verification", Proc. Odyssey Speaker Recognition Workshop, vol 1, pp. 135-140, Junio 2004.

PRESUPUESTO

1) Ejecución Material

- Compra de ordenador personal (Software incluido)..... ..... 2.000 €
- Alquiler de impresora láser durante 12 meses ..... 120 €
- Material de oficina ..... 150 €
- Total de ejecución material ..... 2.270 €

2) Gastos generales

- 16 % sobre Ejecución Material ..... 363 €

3) Beneficio Industrial

- 6 % sobre Ejecución Material ..... 136 €

4) Honorarios Proyecto

- 960 horas a 15 € / hora ..... 14400 €

5) Material fungible

- Gastos de impresión ..... 180 €
- Encuadernación ..... 30 €

6) Subtotal del presupuesto

- Subtotal Presupuesto ..... 17379 €

7) I.V.A. aplicable

- 16% Subtotal Presupuesto ..... 2780 €

8) Total presupuesto

- Total Presupuesto ..... 20159 €

Madrid, Septiembre de 2008

El Ingeniero Jefe de Proyecto

Fdo.: Daniel Hernández López
Ingeniero de Telecomunicación

PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, Mejoras en modelado acústico para reconocimiento del locutor dependiente de texto. 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.