eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Aportaciones al reconocimiento de locutores para su integración en la inteligencia ambiental

MEMORIA
que para optar al grado de Doctor por la UPV/EHU presenta:

Maider Zamalloa Aquizu

bajo la dirección de:

Germán Bordel Garcia
Luis Javier Rodríguez Fuentes

Leioa, Junio de 2010

Departamento de Electricidad y Electrónica
Facultad de Ciencia y Tecnología
Universidad del País Vasco

Resumen

Esta tesis aborda aspectos científico-tecnológicos del reconocimiento de locutores mediante la voz y su aplicación en tareas de verificación, identificación y seguimiento. El estudio está motivado principalmente por el interés en la adecuación de este tipo de sistemas a entornos de inteligencia ambiental (AmI), entornos que disponen de dispositivos electrónicos que permiten detectar la presencia de usuarios y adaptarse a sus necesidades de forma transparente. Debido a que gran parte de estos dispositivos son embebidos, tan sólo se dispone de una cantidad limitada de recursos computacionales. Los sistemas que se construyan deberán ser capaces de funcionar ininterrumpidamente, ser auto-configurables y presentar baja latencia, de forma que el usuario pueda acceder a los servicios en cualquier instante.

El primer aspecto abordado se centra en el objetivo de ahorrar esfuerzo computacional con una baja degradación o incluso una mejora del rendimiento del sistema, y para ello se han estudiado diferentes metodologías de reducción de dimensionalidad de la representación acústica de las señales. La reducción se aplica sobre un conjunto de parámetros estándar empleado por la mayoría de sistemas de procesamiento del habla, compuesto por los coeficientes cepstrales, la energía del tramo y sus derivadas. Se ha propuesto y validado la selección de un subconjunto de parámetros mediante algoritmos genéticos (GA), que permite realizar búsquedas sin establecer límites a toda la combinatoria posible. De los resultados obtenidos se concluye que, si debido a limitaciones computacionales, se tuviera que emplear un conjunto reducido de parámetros, los parámetros estáticos (los coeficientes cepstrales) constituirían la mejor elección, conjunto que aumentaríamos con la energía si el entorno de implantación proporcionase calidad de habla limpia. Como metodología de contraste se ha analizado el empleo de transformaciones lineales que permiten seleccionar parámetros así como eliminar posibles correlaciones. Para subconjuntos de dimensión alta y señales con calidad telefónica, las transformaciones lineales presentan un mejor rendimiento que los subconjuntos definidos por el GA. Pero para subconjuntos de dimensión pequeña, GA muestra un rendimiento mejor, probablemente debido a que la reducción se aplica tratando de maximizar la tasa de reconocimiento de locutores lo cual es coherente con el objetivo final y valida el empleo de GA para este tipo de búsquedas heurísticas.

En segundo lugar, se ha propuesto un sistema de seguimiento de locutores continuo, de baja latencia, para su aplicación en un hogar inteligente. En los principales escenarios de aplicación de los sistemas de seguimiento y diarización de locutores (reuniones, noticias, conversaciones telefónicas), el recurso de audio se conoce de forma íntegra desde el momento en el que se inicia el procesamiento, de modo que las aproximaciones definidas en la bibliografía son iterativas o cíclicas, con una fase de segmentación y otra de detección, que se aplican secuencialmente. Con objeto de cumplir con los requisitos de continuidad y latencia baja, el sistema propuesto sigue un algoritmo muy simple: la segmentación y la detección de la señal de entrada se realizan de forma conjunta mediante la definición de segmentos de longitud fija (concretamente, de un segundo de duración). Los locutores objetivo se modelan mediante mezclas de gaussianas estimadas mediante adaptación bayesiana de un

modelo universal. Estos modelos presentan un buen rendimiento para la detección de segmentos de duración limitada y son computacionalmente eficientes. El sistema se ha evaluado sobre la base de datos multimodal AMI (*Augmented Multi-party Interaction*) grabada en una sala de reuniones inteligente. El rendimiento del sistema propuesto se ha comparado con un sistema de referencia desacoplado clásico, con una fase de segmentación y otra de detección. Como cabía esperar, el sistema de referencia presenta un rendimiento mejor, pero la degradación relativa del sistema continuo es únicamente del 1.2% (con respecto a la medida-F). Por tanto, en situaciones en que el escenario de aplicación requiere una respuesta inmediata, resulta adecuada la aplicación del sistema propuesto, ya que cumple los requisitos de continuidad y latencia baja, a costa de una pequeña degradación del rendimiento. Por otro lado, desde un punto de vista práctico, se ha considerado el desarrollo de aplicaciones según la especificación SOA (*Service Oriented Architecture*), estándar que define un marco muy apropiado para proporcionar movilidad e interoperabilidad a las aplicaciones. Se ha desarrollado una aplicación de seguimiento de locutores, denominada *AmISpeaker*, bajo SOA que implementa el sistema continuo propuesto.

El trabajo realizado ha implicado el estudio de los distintos elementos de que consta un sistema de reconocimiento de locutores, y las metodologías de referencia, definidas en los últimos años. En esta línea, se ha propuesto una nueva metodología de modelado, denominada modelo superficial de fuente, que se estima a partir de la propia señal a evaluar y trata de representar la fuente de la señal con objeto de mejorar la robustez de los sistemas ante señales con fuentes no modeladas. Asimismo, se han analizado las infraestructuras necesarias para la evaluación de los sistemas. Entre ellas, merecen especial atención las campañas de evaluación organizadas por el NIST desde 1996, que definen un marco experimental estándar para el reconocimiento de locutores independiente del texto, reconocido y empleado a nivel mundial por los más importantes grupos de investigación dedicados a este tipo de tareas. Las metodologías de referencia para la detección y el modelado de locutores, y la normalización y la fusión de sistemas, se han evaluado sobre el conjunto experimental definido por el NIST para la evaluación del 2006. Por último, el trabajo de esta tesis conllevó la participación en la evaluación organizada por el NIST a mediados de 2008, con el desarrollo de dos sistemas cuyos resultados pueden considerarse buenos teniendo en cuenta que se trataba de un marco de evaluación completamente nuevo para el grupo de investigación.

Abstract

This thesis addresses scientific and technological issues related to speaker recognition and its application to verification, identification and tracking tasks. The study focuses on the adaptation of such systems to Ambient Intelligence (AmI) environments, which integrate electronic devices that detect the presence of users and adapt to their needs in a transparent manner. Because many of such devices are embedded in the environment, computational resources are limited. Uninterrupted, self-configurable and low-latency systems are needed, so that the user can consume the services provided at any time.

In order to decrease the computational effort with a low degradation or even an improvement in system performance, different methods of dimensionality reduction of the representation of acoustic signals are analyzed. The study aims to reduce the dimensionality of one of the most common representations used by speech processing systems: the cepstral coefficients, the energy and their first and second derivates. In this work, a Genetic Algorithm (GA) based feature selection procedure is proposed. The results obtained suggest that, if due to computational restrictions, a reduced set of feature had to be used, the static features (the cepstral coefficients) would be the best choice, augmented with the frame energy when dealing with clean laboratory speech. As a contrastive approach, the application of linear transforms has been analyzed, that can select features and remove correlations related to, for example, channel noise. In fact, when dealing with telephone speech and high dimension feature sets, linear transforms outperform GA driven selection. However, for small dimension feature sets GA shows better performance, probably due to the fact that reduction procedures try to maximize the speaker recognition rate, which is consistent with the ultimate goal, and validates the use of GA for this kind of heuristic search procedures.

On the other hand, a low-latency uninterrupted (*online*) speaker tracking system is proposed, designed for an intelligent home environment. In most speaker tracking and diarization primary application domains (meetings, broadcast news, telephone conversations) audio recordings are fully available before processing. Therefore, common approaches to these tasks are cyclical or iterative, consisting of two uncoupled steps (audio segmentation and speaker detection) and do not allow low-latency and uninterrupted execution. In order to cope with the low-latency and uninterrupted functionality requirements, a very simple speaker tracking algorithm is proposed, where audio segmentation and speaker detection are jointly accomplished by defining and processing fixed-length audio segments (one second audio segments). The target speakers are modelled by Gaussian Mixture Models (GMM) estimated by Bayesian adaptation of a Universal Background Model (UBM). This methodology has shown good speaker recognition performance when dealing with short duration speech segments and is computationally efficient. System evaluation is carried out on the AMI (*Augmented Multi-party Interaction*) corpus of meeting conversations, a multimodal data set concerned with real-time human interaction in the context of smart meeting rooms. The performance

of the proposed approach is compared to that of an *offline* system developed for reference, which follows the classical two-stage approach: audio segmentation is done over the whole input stream and speaker detection is performed on the resulting segments. As expected, the reference system actually outperforms the proposed system, which provides only a 1.2% relative degradation (with regard to F-measure). Therefore, depending on the scenario and the required latency, the proposed approach would be more feasible, as it provides low-latency uninterrupted speaker tracking with little performance degradation. From a practical point of view, it was found interesting the development of software applications that follow the SOA (*Service Oriented Architecuture*) specification, a very valuable means that supports application mobility and interoperability. A speaker tracking application, named AmISpeaker, has been developed, which implements the online speaker tracking system proposed in this work.

This thesis has involved the study of elements of a speaker recognition system (acoustic parameters, models, calibration, etc.), reference methods defined in recent years and the infrastructure necessary for the evaluation of systems. Therein, in order to improve the robustness of systems when dealing with training-test acoustic mismatch, a new modelling approach, named Shallow Source Model (SSM), has been proposed. It is a low resolution GMM estimated from the input utterance which aims to model the source of the input utterance. Regarding evaluation infrastructures the evaluation campaigns organized by NIST since 1996 deserve special attention. These evaluations define an experimental framework for text-independent speaker recognition. Research groups participate in them as a means of evaluate their technology in an international forum. In this work, the evaluation of reference methodologies for speaker modelling and detection, score normalization and system fusion has been carried out on the experimental corpus defined by NIST for the 2006 evaluation campaign. Finally, this thesis also involved participating in the NIST 2008 evaluation campaign, with the development of two speaker recognition systems, whose performance can be regarded as good, taking into account that it was an entirely new evaluation framework for the research group.

Laburpena

Honako tesia hizlari ezagutzaren funts teknologiko eta zientifikoen azterketan oinarritzen da, beti ere identifikazio, egiaztapen eta jarraipen prozesuei jarraituz. Ikerkuntza lanaren ildo nagusiak sistema horien inguru adimentsuetarako egokitzapena du helburu, erabiltzailearen presentzia detektatu eta bere beharretara egokitzen den gailu elektronikoz hornituta dauden inguruetarakoa. Gailu horietako gehienak inguruan bertan murgildu behar izateak euren baliabide konputazionalak murrizten ditu. Definitzen diren sistemek etenik gabekoak eta latentzia baxukoak izan behar dute, eta baita automatikoki konfiguratzeko gaitasundunak, erabiltzaileak ingurunean eskaintzen diren zerbitzuak edozein unetan atzi ditzan.

Sistemen koste konputazionala gutxitzeko asmoz, ahots seinaleen adierazpide akustikoaren dimentsionalitate murrizketan oinarritutako hainbat metodología aztertu dira, murrizketak eraginkortasun galera txiki bat edo, kasu onenetan, eraginkortasuna hobetzea ekar dezakeela kontuan izanda. Bilaketa hizketaren prozesamenduan oinarritutako sistema gehienek erabiltzen duten adierazpide akustikotik abiatzen da, MFCC (*Mel Frequency Cepstral Coefficients*) koefizienteak, ahots-leihoaren energa eta horien lehenengo eta bigarren deribatuak biltzen dituen adierazpidetik. Parametro multzo horretatik abiatuz, hizlarien ezagutzarako azpimultzo optimoaren hautaketa pozesu bat proposatu eta balioztatu da. Hautaketa prozesua algoritmo genetikoen (*Genetic Algorithm*, GA) bidez gauzatu da, GAek ez baitute bilaketan azter daitekeen konbinatoria posible kopurua mugatzen. Baliabide konputazionalak gutxitzeagatik hizlarien ezagutzarako parametro multzo murriztu bat erabili behar izanez gero, aukera egokiena parametro estatikoak (MFCC koefizienteak) erabiltzea dela adierazten dute lortutako emaitzek (seinaleak ahots grabaketara egokitutako laborategi batean jasoz gero, parametro multzoan ahots-leihoaren energia ere sartu beharko litzateke). AGen bidezko hautaketa prozesua transformazio linealen erabilpenarekin alderatu da; izan ere, transformazio linealek parametroak hautatzeaz gain, euren arteko korrelazioak ezabatzea ere ahalbidetzen baitute. Alderaketan, telefono kanal batetan grabatutako seinaleei dagozkien dimentsio altuko adierazpideak definitzeko transformazio linealen bidezko egokitzapen prozesua AGen bidezkoa baino egokiagoa dela ikusi da. Dimentsio txikiko parametro multzoak definitzeko garaian, ordea, AGen bidezko hautaketa prozesua eraginkorragoa da. Portaera hori AGen bidezko parametro hautaketa prozesua ezagutza tasaren maximizazioan oinarritzearen ondorio izan daitekeela uste da berau baita bilaketa prozesuaren helburu nagusia. Aldi berean, portaera horrek, era horretako bilaketa heuristikotarako GAek duten eraginkortausna balioztatzen du.

Bestalde, adimendun bizileku bateko hizlarien jarraipena egiten duen latentzia baxuko sistema jarrai bat proposatu da, hau da, berehalako erantzunak ematen dituen etenik gabeko sistema bat. Hizlarien jarraipen eta diarizazio sistemen oinarrizko aplikazio inguruetan (bilerak, berriak, teléfono hizketak), aztertu beharreko grabazioa osorik ezagutzen da prozesamendua abiatzen den unean bertan, eta ondorioz, bibliografian definitutako sistema gehienak zikliko edo iteratiboak dira, sekuentzialki

aplikatzen diren segmentazio eta detekzio fase banaz osatutakoak. Lan honetan proposatutako sistemak, aldiz, jarraitasun eta latentzia baxua izatearen baldintzak betetzen ditu, baldintza horiek bete ahal izateko algoritmo sinple bat jarraituz: jarraipena gauzatu beharreko seinalearen gaineko segmentazio eta detekzioa era bateratuan egiten da luzera finkoko seinaleak definituz, zehazki segundo bateko seinaleak definituz. Jarraipenaren helburu diren hizlariak GMM (*Gaussian Mixture Model*) modeloak erabiliz definitzen dira, modelo hauek oinarrizko modelo baten eraldaketa bidez estimatzen dira metodologia bayesiarra erabiliz. Modelo horiek oso portaera egokia erakutsi dute luzera mugatua duten segmentuen hizlaria detektatzeko garaian, eta konputazionalki ere oso eraginkorrak dira. Sistemaren ebaluazioa AMI (*Augmented Multi-party Interaction*) datu base multimodala erabiliz egin da, adimendun bilera geletan grabatutakoa. Sistema jarraia ohiko prozedura jarraitzen duen sistema batekin alderatu da, hau da segmentazio eta detekzioa fase banatuetan burutzen dituen sistema batekin. Espero bezala, ohiko prozedura jarraitzen duen sistema eraginkorragoa da, baina bien arteko aldea %1.2koa da bakarrik (F-measure metrikari jarraiki). Ondorioz, sistemaren aplikazio inguruneak berehalako erantzunak eskatzen dituen kasuetarako egokia litzateke proposatutako sistema, jarraikortasun eta latentzia baxua izateko baldintzak betetzen baititu eraginkortasun galera txiki baten trukean. Bestalde, lan ildo praktikoago bat jorratzeko asmoz, SOA (*Service Oriented Architecture*) espezifikazioa jarraitzen duten software aplikazioen garapena landu da; izan ere, SOAk abantaila ugari eskaintzen baititu aplikazioen elkarrekintza eta mobilizaziorako. Horrela, hizlarien jarraipena helburu duen software aplikazioa garatu da, *AmISpeaker* deitu zaiona eta, SOA espezifikazioa eta proposatutako sistema jarraitzen dituena.

Tesian zehar egindako lanak hizlariaren ezagutzan oinarritutako sistema bat osatzen duten elementu guztien azterketa eskatu du, eta hauetan aplikatzen diren erreferentziazko metodologiena. Ildo horri jarraiki, seinalearen jatorri akustikoa modelatzea helburu duen metodologia berri bat proposatu da lan honetan, seinalearen jatorri akustikoa modelatzea helburu duena, aurrez aztertu gabeko jatorriren batetik datozen seinaleetako hizlaria detektatzea sendotzeko. Era berean, sistemen ebaluazioa egin ahal izateko beharrezkoak diren azpiegiturak ere aztertu dira. Horien artean, aipamen berezia merezi dute NIST erakundeak 1996 urteaz geroztik antolatutako ebaluazio kanpainek. Ahoskatutako mezuarekiko idenpendenteak diren hizlari ezagutza sistemen ebaluaziorako ikerkuntza eremu komun bat definitzen dute mota horretako atazak lantzen dituzten mundu mailako ikerketa talde gehienek erabili eta onartu dutena. Era horretara, hizlariak modelatu eta detektatzeko jatorrizko metodologiak zein probabilitate tasak normalizatu eta sistemak fusionatzekoak NISTek 2006 urteko ebaluaziorako definitutako corpus experimentalaren gainean ebaluatu dira. Azkenik, ikerkuntza lan honek emandako beste fruituetako bat NISTek 2008an antolatu zuen ebaluazio kanpainan partehartzea izan da. Bi sistema garatu ziren ebaluaziorako emaitza onak lortu zituztenak kontuan izanik ikerkutza talderako eremu esperimental erabat ezezaguna zela.

Agradecimientos

Mirando a la trayectoria seguida durante estos cuatro años y medio, me resulta complicado por donde, y con quien, empezar la redacción de esta parte de la tesis. Pero lo que no da lugar a dudas es que está tesis es fruto de un trabajo realizado en equipo, en un equipo llamado GTTS (Grupo de Trabajo de Tecnologías de Software), destacado por el compañerismo y la buena disposición de sus miembros. Gracias a ellos, he conocido el área de investigación de las tecnologías del habla, un campo prácticamente desconocido cuando empecé con la tesis y resultaba cada vez más interesante. En este sentido, quisiera destacar la labor realizada por Germán Bordel y Luis Javier Rodriguez-Fuentes en dirigir esta tesis, quienes junto a Mikel Peñagarikano, han sido continuos participes del trabajo realizado. Este trabajo no hubiera visto la luz sin su colaboración. Muchas gracias a los tres, así como a todos los miembros del grupo GTTS.

Debo también destacar el apoyo recibido desde el centro tecnológico Ikerlan, y en especial, el recibido por parte de todos los miembros del area de conocimiento de las tecnologías del software. Entre ellos, merece una mención especial la colaboración de Juan Pedro Uribe, tanto por el esfuerzo realizado en impulsar esta tesis, como por las facilidades prestadas en todo momento. Asimismo, quisiera destacar la colaboración puntual de Aitor Uribarren y Jorge Parra, y recalcar el apoyo recibido por parte de todos los compañeros del área de conocimiento.

Eta azkenik, nola ez, eskerrik asko etxeko guztioi, muga gabeko babes eta laguntzagatik.

Maider Zamalloa Akizu

Aretxabaletan, 2010eko Ekainaren 23an.

Índice general

1. INTRODUCCIÓN 1
1.1 CONTEXTO Y MOTIVACIÓN 1
1.2 OBJETIVOS 7
1.3 ESTRUCTURA DE LA MEMORIA 9

2. ESTADO DEL ARTE 13
2.1 INTELIGENCIA AMBIENTAL 13
2.1.1 Hogares inteligentes 17
2.1.2 Seguimiento de usuarios 19
2.2 IDENTIFICACIÓN Y VERIFICACIÓN DE LOCUTORES 24
2.3 PARAMETRIZACIÓN ACÚSTICA 26
2.3.1 Parámetros de bajo nivel 27
2.3.2 Parámetros de alto nivel 29
2.4 RECONOCIMIENTO DE LOCUTORES 29
2.4.1 La clasificación de locutores 30
2.4.2 Modelos de locutor 32
2.4.3 Normalización de probabilidades 46
2.4.4 Fusión de Sistemas 49
2.5 SEGUIMIENTO DE LOCUTORES 51
2.5.1 Metodologías para la segmentación y la clasificación de locutores 53
2.5.2 Aproximaciones al seguimiento de locutores continuo 59
2.6 BASES DE DATOS, MÉTRICAS DE RENDIMIENTO Y CAMPAÑAS DE EVALUACIÓN 62
2.6.1 Bases de datos 62
2.6.2 Evaluaciones 67
2.6.3 Métricas de rendimiento 72

3. REDUCCIÓN DE DIMENSIONALIDAD DE LA PARAMETRIZACIÓN ACÚSTICA 79
3.1 TÉCNICAS DE REDUCCIÓN DE DIMENSIONALIDAD 80
3.1.1 Transformaciones lineales 81
3.1.2 Algoritmos genéticos 86
3.2 EXPERIMENTOS DE REDUCCIÓN DE DIMENSIONALIDAD MEDIANTE TRANSFORMACIONES LINEALES 90
3.2.1 Configuración experimental 90
3.2.2 Resultados 91
3.2.3 Conclusiones 97
3.3 EXPERIMENTOS DE REDUCCIÓN DE DIMENSIONALIDAD MEDIANTE GA 99
3.3.1 Configuración experimental 100
3.3.2 Resultados de selección mediante GA 103
3.3.3 Conclusiones 115
3.4 COMPARACIÓN DE METODOLOGÍAS 116

4. SISTEMAS DE IDENTIFICACIÓN Y VERIFICACIÓN DE LOCUTORES 121
4.1 EXPERIMENTOS DE RECONOCIMIENTO DE LOCUTORES EN ALBAYZÍN 122
4.1.1 Identificación de locutores sobre un conjunto abierto 122
4.1.2 Verificación de locutores 125
4.1.3 Aproximaciones para mejorar la identificación y verificación de locutores en señales procedentes de fuentes no modeladas 128
4.1.4 Conclusiones del apartado 4.2 145
4.2 EXPERIMENTOS DE VERIFICACIÓN DE LOCUTORES EN LA NIST SRE06 148
4.2.1 Sistemas definidos para la NIST SRE06 150
4.2.2 Resultados de los sistemas definidos para la NIST SRE06 151
4.2.3 Resultados de la normalización de probabilidades en los sistemas definidos para la NIST SRE06153
4.2.4 Resultados de la fusión de los sistemas definidos para la NIST SRE06 161
4.2.5 Conclusiones y trabajo futuro 168

5. SEGUIMIENTO DE LOCUTORES EN UN ENTORNO AMI 171
5.1 SISTEMA CONTINUO DE SEGUIMIENTO DE LOCUTORES 171
5.1.1 Descripción de los módulos del sistema 172
5.2 EXPERIMENTOS DEL SISTEMA DE SEGUIMIENTO CONTINUO 175
5.2.1 Configuración experimental 175
5.2.2 Resultados de los experimentos de seguimiento 177
5.3 AMISPEAKER: IMPLANTACIÓN DEL SISTEMA DE SEGUIMIENTO DE LOCUTORES CONTINUO 187
5.3.1 Descripción de la aplicación AmISpeaker 187
5.4 CONCLUSIONES 190

6. CONCLUSIONES Y TRABAJO FUTURO 195
6.1 RESUMEN GLOBAL DE LAS CONCLUSIONES DE LA TESIS 195
6.1.1 Reducción de dimensionalidad de la parametrización acústica 196
6.1.2 Metodologías de modelado computacionalmente eficientes, robustos y adaptables 197
6.1.3 Diseño y evaluación de un sistema de seguimiento de locutores continuo y de latencia baja198
6.1.4 Publicaciones 200
6.2 FUTURAS LÍNEAS DE TRABAJO 201

7. BIBLIOGRAFÍA 203

Lista de Acrónimos

AHS	*Arithmetic Harmonic Sphericity*
AI	Inteligencia Artificial (*Artificial Intelligence*)
AmI	Inteligencia Ambiental (*Ambient Intelligence*)
AMI	*Augmented Multi-party Interaction*
AMIDA	*Augmented Multi-party Interaction with Distance Access*
ASTS	Servicio de seguimiento de locutores de un entorno AmI *(AmI Speaker Tracking Service)*
AT-NORM	*Adaptive T-NORM*
BIC	*Bayesian Information Criterion*
CHIL	*Computers in the Human Interaction Loop*
CLR	*Cross Likelihood Ratio*
CMS	Normalización de la Media Cepstral (*Cepstral Mean Substraction*)
DCT	Transformada Discreta de Cosenos (*Discrete Cosine Transform*)
DET	*Detection Error Trade-off*
D-NORM	*Distance-NORMalization*
DSBM	*Document Speech Background Model*
EER	*Equal Error Rate*
ELRA	*European Language Resources Association*
EM	Esperanza-Maximización (*Expectation-Maximization*)
e-VQ	Distribución de Etiquetas de Cuantificación Vectorial
FFT	Transformada Rápida de Fourier (*Fast Fourier Transform*)
FOCAL	*FusiOn and CALibration Toolkit*
FW	*Feature Warping*
GA	Algoritmo Genético (*Genetic Algorithm*)
GLDS Kernel	Kernel de Secuencia de Discriminación Lineal Generalizada (*Generalized Linear Discriminant Sequence Kernel*)
GLR	*Generalized Likelihood Criterion*
GMM	Mezcla de Gaussianas (*Gaussian Mixture Model*)
GMM(EM)	GMM estimado mediante el algoritmo EM
GMM(MAP)	GMM estimado mediante adaptación bayesiana (o estimación MAP) de los parámetros de las componentes gaussianas de un GMM de referencia
GMM(SVM)	SVM con un kernel secuencial de vectores de soporte GMM
HLDA	Análisis Discriminante Lineal Heteroscedástico (*Heteroscedastic Linear Discriminant Analysis*)
HMM	Modelos Ocultos de Markov (*Hidden Markov Models*)
H-NORM	*Handset-NORMalization*

IDORM	*Intelligent DORMitory*
ISA	*Incremental Speaker Adaptation*
ISTAG	*Information Society Technologies Program Advisory Group*
JFA	*Joint Factor Analysis*
KKT	*Karush-Kuhn-Tucker*
KL	Distancia *Kullback-Leibler*
KLT	Transformada de Karhunen-Loève (*Karhunen-Loève Transform*)
KL-TNORM	*Kullback Leibler T-NORM*
K-NN	*K-Nearest-Neighbours*
LDA	Análisis Discriminante Lineal (*Linear Discriminant Analysis*)
LDC	*Linguistic Data Consortium*
LLR	Ratio de Verosimilitud *(Log-Likelihood Ratio)*
MAP	*Maximum A Posteriori*
MAVHOME	*Managing an Adaptive Versatile HOME*
MDE	Extracción de Metadatos (*Meta-Data Extraction*)
MFCC	Coeficientes Cepstrales en la escala de Mel (*Mel Frequency Cepstral Coefficients*)
MIT	*Massachusetts Institute of Technology*
ML	Máxima Verosimilitud (*Maximum-Likelihood*)
MLLR Kernel	Kernel de Transformadas de Regresión Lineal de Máxima Verosimilitud (*Maximum-Likelihood Linear Regression Transform Kernel*)
NAP	*Nuissance Attribute Projection*
NIST	Instituto Nacional de Estándares y Tecnología (*National Institute of Standards and Technology*)
NIST RT	Evaluación de Sistemas de Transcripción Enriquecida del Habla del NIST (NIST *Rich Transcription Evaluation*)
NIST SRE	Evaluación de Reconocimiento de Locutores del NIST **(**NIST *Speaker Recognition Evaluation*)
PCA	Análisis de Componentes Principales (*Principal Component Analysis*)
PLP	Percepción Lineal Perceptual (*Perceptual Linear Prediction*)
PRC	Precisión de un sistema de seguimiento/diarización
RCL	Cobertura (*recall*) de un sistema de seguimiento/diarización
RFID	*Radio Frequency Identification*
ROC	*Receiver Operating Characteristic*
SD	Algoritmo de selección de parámetros acústicos mediante GA
SCARS	*Speech Control and Audio Response System*
SCV	Vector de Caracterización del Locutor (*Speaker Characterization Vector*)
SDS	Servicio de detección de locutores (*Speaker Detection Service*)
STS	Servicio de seguimiento de locutores (*Speaker Tracking Service*)

SGA	Algoritmo Genético Simple (*Simple Genetic Algorithm*)
SOA	Arquitectura Orientada a Servicios (*Service Oriented Architecture*)
SOC	*Service Oriented Computing*
SPHERE	*SPeech HEader REsources*
SPO	Algoritmo de selección de parámetros mediante un raking de pesos óptimos estimados con GA
SSM	Modelo Superficial de Fuente (*Shallow Source Model*)
SSM	*Sample Speaker Model*
STT	Transcripción de voz a texto (*Speech-to-Text Transcription*)
SVM	Máquina de Vectores Soporte (*Support Vector Machine*)
SVM-GLDS	SVM con un GLDS kernel
TDOA	*Time Delay of Arrivals*
T-NORM	Normalización de test (*Test-NORMalization*)
UBM	Modelo Universal (*Universal Background Model*)
UPC	Universidad Politécnica de Catalunya
UPnP	*Universal Plug and Play*
UGM	*Universal Gender Model*
XML	*eXtensible Markup Language*
Z-NORM	Normalización cero (*Zero-NORMalization*)
ZSTS	Servicio de seguimiento de locutores de una zona *(Zone Speaker Tracking Service)*

Índice de Figuras

Figura 1.1. Clasificación de tareas relacionadas con el reconocimiento de locutores..... 4
Figura 2.1. Fase de entrenamiento de un sistema de reconocimiento de locutores..... 25
Figura 2.2. Fase de evaluación de un sistema de identificación de locutores..... 25
Figura 2.3. Fase de evaluación de un sistema de verificación de locutores..... 25
Figura 2.4. Esquema conceptual de una máquina de vectores soporte (SVM)..... 40
Figura 2.5. Esquema conceptual de una SVM con variables de holgura..... 42
Figura 3.1. Funcionamiento de un Algoritmo Genético Simple (SGA)..... 88
Figura 3.2. Representación esquemática de las particiones creadas sobre Albayzín y Dihana para la experimentación con transformaciones lineales..... 91
Figura 3.3. Error promedio de los conjuntos obtenidos con PCA en Albayzín y Dihana..... 93
Figura 3.4. Error promedio de los conjuntos obtenidos con LDA en Albayzín y Dihana..... 95
Figura 3.5. Error promedio de los conjuntos obtenidos con HLDA en Albayzín y Dihana..... 97
Figura 3.6. Error promedio de los conjuntos obtenidos con PCA, LDA_MAP, HLDA_EM y HLDA_MAP en Albayzín..... 98
Figura 3.7. Error promedio de los conjuntos obtenidos con PCA, LDA_MAP, HLDA_EM y HLDA_MAP en Dihana..... 99
Figura 3.8. Representación esquemática de las particiones creadas sobre Albayzín y Dihana para evaluar el rendimiento de las aproximaciones con GA..... 100
Figura 3.9. Error promedio de los subconjuntos óptimos obtenidos mediante SD en Albayzín..... 107
Figura 3.10. Error promedio de los subconjuntos óptimos obtenidos mediante SD en Dihana..... 108
Figura 3.11. Error promedio de los subconjuntos obtenidos con SD, SPO, SD_W (SD con pesos óptimos) y SPO_W (SPO con pesos óptimos) en Albayzín..... 113
Figura 3.12. Error Promedio de los subconjuntos obtenidos con SD, SPO, SD_W (SD con pesos óptimos) y SPO_W (SPO con pesos óptimos) en Dihana..... 114
Figura 3.13. Resumen de los experimentos de reducción de dimensionalidad en Albayzín..... 117
Figura 3.14. Resumen de los experimentos de reducción de dimensionalidad en Dihana..... 118
Figura 4.1 (a) Representación esquemática de la partición "AlbID 34/102/68" de Albayzín, definida para tareas de identificación de locutores en conjuntos abiertos. **(a+b)** Representación esquemática de "AlbID 34/102/68+Mismatched" que incluye un conjunto de test con señales no modeladas.en entrenamiento..... 123
Figura 4.2. Curvas DET de los sistemas de identificación GMM(EM) y GMM(MAP) en Albayzín.124
Figura 4.3. Curvas DET de los sistemas de identificación GMM(EM) y GMM(MAP) sobre Albayzín y un conjunto de test que contiene señales con fuentes no modeladas..... 124
Figura 4.4 (a) Representación esquemática de la partición "AlbVER 34/102/68" de Albayzín, definida para tareas de verificación de locutores. **(a+b)** Representación esquemática de la partición "AlbVER

34/102/68+Mismatched", que incluye un conjunto de test con señales no modeladas en entrenamiento 126
Figura 4.5. Curvas DET de los sistemas de verificación GMM(EM) y GMM(MAP) en Albayzín. . 126
Figura 4.6. Curvas DET de los sistemas de verificación GMM(EM) y GMM(MAP) sobre Albayzín y un conjunto de test que contiene señales con fuentes no modeladas..... 127
Figura 4.7. Curvas DET de los sistemas de identificación que emplean SSM como factor de normalización, evaluados sobre Albayzín..... 133
Figura 4.8. Curvas DET de los sistemas de verificación que emplean SSM como factor de normalización, evaluados sobre Albayzín y un conjunto de test con fuentes no modeladas. 134
Figura 4.9. Curvas DET de los sistemas de identificación GMM(EM), GMM(EM)/SSM, GMM(EM)/UBM+SSM y GMM(EM)/UBM+Uniform sobre Albayzín..... 136
Figura 4.10. Curvas DET de los sistemas de identificación GMM(EM), GMM(EM)/SSM, GMM(EM)/UBM+SSM y GMM(EM)/UBM+Uniform sobre Albayzín y un conjunto de test de señales con fuentes no modeladas..... 137
Figura 4.11. Curvas DET de los sistemas de identificación GMM(MAP), GMM(MAP)/UBM+SSM y GMM(MAP)/UBM+Uniform sobre Albayzín..... 139
Figura 4.12. Curvas DET de los sistemas de identificación GMM(MAP), GMM(MAP)/UBM+SSM y GMM(MAP)/UBM+Uniform sobre Albayzín y un conjunto de test de señales con fuentes no modeladas..... 140
Figura 4.13. Curvas DET de los sistemas de verificación GMM(EM), GMM(EM)/SSM, GMM(EM)/UBM+SSM y GMM(EM)/UBM+Uniform sobre Albayzín..... 141
Figura 4.14. Curvas DET de los sistemas de verificación GMM(EM), GMM(EM)/SSM, GMM(EM)/UBM+SSM y GMM(EM)/UBM+Uniform sobre Albayzín y un conjunto de test de señales con fuentes no modeladas..... 142
Figura 4.15. Curvas DET de los sistemas de verificación GMM(MAP), GMM(MAP)/UBM+SSM y GMM(MAP)/UBM+Uniform sobre Albayzín..... 144
Figura 4.16. Curvas DET de los sistemas de verificación GMM(MAP), GMM(MAP)/UBM+SSM y GMM(MAP)/UBM+Uniform sobre Albayzín y un conjunto de test de señales con fuentes no modeladas..... 145
Figura 4.17. Curvas DET de los sistemas GMM(MAP) y GMM-SVM sobre los conjuntos experimentales de la NIST SRE06..... 152
Figura 4.18. Curvas DET del sistema GMM(MAP) con Z-norm y T-norm sobre la tarea principal de la NIST SRE06..... 154
Figura 4.19. Curvas DET del sistema GMM(MAP) con Z-norm y T-norm sobre el subconjunto de experimentos en ingles de la tarea principal de la NIST SRE06..... 154
Figura 4.20. Curvas DET del sistema GMM-SVM con Z-norm y T-norm sobre la tarea principal de la NIST SRE06..... 155

Figura 4.21. Curvas DET del sistema GMM-SVM Z-norm y T-norm sobre el subconjunto de experimentos en ingles de la tarea principal de la NIST SRE06 155
Figura 4.22. Curvas DET del sistema GMM(MAP) con ZT-norm y TZ-norm sobre la tarea principal de la NIST SRE de 2006 157
Figura 4.23. Curvas DET del sistema GMM(MAP) con ZT-norm y TZ-norm sobre el subconjunto de experimentos en ingles de la tarea principal de la NIST SRE06 157
Figura 4.24. Curvas DET del sistema GMM-SVM con ZT-norm y TZ-norm sobre la tarea principal de la NIST SRE06 158
Figura 4.25. Curvas DET del sistema GMM-SVM con ZT-norm y TZ-norm sobre el subconjunto de experimentos en inglés de la tarea principal de la NIST SRE06 158
Figura 4.26. Curvas DET de los sistemas GMM(MAP), GMM-SVM y el sistema que resulta de la fusión – optimizada en cerrado – de ambos sobre la tarea principal de la NIST SRE05 163
Figura 4.27. Curvas DET de los sistemas GMM(MAP), GMM-SVM y el sistema que resulta de la fusión de ambos sistemas sobre la tarea principal de la NIST SRE06 164
Figura 4.28. Curvas DET de los sistemas GMM(MAP), GMM-SVM y el sistema que resulta de la fusión de ambos sobre el subconjunto de experimentos de la tarea principal de la NIST SRE06 164
Figura 4.29. Curvas DET de la fusión de: (1) GMM(MAP) y GMM-SVM; y (2) GMM(MAP) con Z-norm y GMM-SVM con T-norm, sobre la tarea principal de la NIST SRE06 167
Figura 4.30. Curvas DET de la fusión de: (1) GMM(MAP) y GMM-SVM; y (2) GMM(MAP) con Z-norm y GMM-SVM con T-norm, sobre el subconjunto de experimentos en inglés de la tarea principal de la NIST SRE06 167
Figura 5.1 Representación esquemática de las particiones creadas en la base de datos AMI para la evaluación del sistema de seguimiento continuo 176
Figura 5.2. Curvas DET del sistema de seguimiento de locutores continuo (con UBM-t y UBM-g) y el sistema de referencia sobre el conjunto AMI. 178
Figura 5.3. Curvas DET del sistema de seguimiento continuo (con UBM-g y UBM-t) y el sistema de referencia sobre el conjunto de evaluación de AMI. 178
Figura 5.4. Curvas DET del sistema de seguimiento continuo (con UBM-t) sobre el conjunto de desarrollo de AMI al aplicar dos tipos de suavizado 180
Figura 5.5. Curvas DET del sistema de seguimiento continuo (con UBM-t) sobre el conjunto de evaluación de AMI al aplicar dos tipos de suavizado 180
Figura 5.6. Curvas DET del sistema de seguimiento continuo (con UBM-g) sobre el conjunto de desarrollo de AMI, al aplicar dos tipos de suavizado 182
Figura 5.7. Curvas DET del sistema de seguimiento continuo (con UBM-g) sobre el conjunto de evaluación de AMI, al aplicar dos tipos de suavizado 182
Figura 5.8. Curvas DET del sistema de seguimiento continuo sobre el conjunto de desarrollo de AMI, incluyendo la evaluación de segmentos en los que intervienen más de un locutor 184

Figura 5.9. Curvas DET del sistema de seguimiento continuo sobre el conjunto de evaluación de AMI, incluyendo la evaluación de segmentos en lso que intervienen más de un locutor 184
Figura 5.10. Representación esquemática de un servicio de seguimiento de locutores integrable en las aplicaciones que siguen una arquitectura orientada a servicios (SOA). 188

Índice de Tablas

Tabla 3.1. Error promedio e intervalo de confianza del 95% obtenidos con los parámetros originales en Albayzín y Dihana. 92
Tabla 3.2. Error promedio e intervalo de confianza del 95% de los conjuntos obtenidos con PCA sobre Albayzín y Dihana 93
Tabla 3.3. Error promedio e intervalo de confianza del 95% de los conjuntos obtenidos con LDA sobre Albayzín y Dihana. 94
Tabla 3.4. Error promedio e intervalo de confianza del 95% de los conjuntos obtenidos con HLDA sobre Albayzín y Dihana 96
Tabla 3.5. Subconjuntos óptimos de parámetros acústicos determinados por SD sobre Albayzín y Dihana al emplear modelos GMM(EM) 104
Tabla 3.6. Subconjuntos óptimos de parámetros acústicos determinados por SD sobre Albayzín al emplear modelos e-VQ modelos GMM(EM)..... 105
Tabla 3.7. Subconjuntos óptimos de parámetros acústicos determinados por SD sobre Dihana al emplear modelos e-VQ y modelos GMM(EM)..... 106
Tabla 3.8. Error promedio e intervalo de confianza del 95% de los subconjuntos obtenidos con SD sobre Albayzín y Dihana 107
Tabla 3.9. Subconjuntos óptimos de parámetros determinados por SPO sobre Albayzín y Dihana .. 109
Tabla 3.10. Error promedio e intervalo de confianza del 95% de los subconjuntos obtenidos con SPO y SD en Albayzín 110
Tabla 3.11. Error promedio e intervalo de confianza del 95% de los subconjuntos obtenidos con SPO y SD en Dihana 110
Tabla 3.12. Pesos óptimos asignados por el GA a los subconjuntos óptimos SD, SPO, a los conjuntos de referencia MFCC, MFCC+E y el vector original d-dimensional de Albayzín..... 112
Tabla 3.13. Error promedio e intervalo de confianza del 95% de los subconjuntos obtenidos con SD, SPO, SD_W (SD con pesos óptimos) y SPO_W (SPO con pesos óptimos) en Albayzín..... 113
Tabla 3.14. Pesos óptimos asignados por el GA a los subconjuntos óptimos SD y SPO, a los conjuntos de referencia MFCC, MFCC+E y el vector original d-dimensional de Dihana 114
Tabla 3.15. Error promedio e intervalo de confianza del 95% de los subconjuntos obtenidos con SD, SPO, SD_W (SD con pesos óptimos) y SPO_W (SPO con pesos óptimos) en Dihana..... 115
Tabla 3.16. Error promedio e intervalo de confianza del 95% de los experimentos de reducción de dimensionalidad con SD, PCA, LDA y HLDA sobre Albayzín 117
Tabla 3.17. Error promedio e intervalo de confianza del 95% de los experimentos de reducción de dimensionalidad con SD, PCA, LDA y HLDA sobre Dihana..... 118
Tabla 4.1. Tasa EER y coste C_{DET}^{min} de los sistemas de identificación GMM(EM), GMM(EM)/SSM, GMM(EM)/UBM+SSM y GMM(EM)/UBM+Uniform sobre Albayzín. 136

Tabla 4.2. Tasa EER y coste C_{DET}^{min} de los sistemas de identificación GMM(EM), GMM(EM)/SSM, GMM(EM)/UBM+SSM y GMM(EM)/UBM+Uniform sobre Albayzín y un conjunto de test que contiene señales con fuentes no modeladas..... 138

Tabla 4.3. Tasa EER y coste C_{DET}^{min} de los sistemas de identificación GMM(MAP), GMM(MAP)/UBM+SSM y GMM(MAP)/UBM+Uniform sobre Albayzín..... 139

Tabla 4.4. Tasa EER y coste C_{DET}^{min} de los sistemas de identificación GMM(MAP), GMM(MAP)/UBM+SSM y GMM(MAP)/UBM+Uniform sobre Albayzín y un conjunto de test que contiene señales con fuentes no modeladas..... 140

Tabla 4.5. Tasa EER y coste C_{DET}^{min} de los sistemas de verificación GMM(EM), GMM(EM)/SSM, GMM(EM)/UBM+SSM y GMM(EM)/UBM+Uniform sobre Albayzín..... 142

Tabla 4.6. Tasa EER y coste C_{DET}^{min} de los sistemas de verificación GMM(EM), GMM(EM)/SSM, GMM(EM)/UBM+SSM y GMM(EM)/UBM+Uniform sobre Albayzín y un conjunto de test que contiene señales con fuentes no modeladas..... 143

Tabla 4.7. Tasa EER y coste C_{DET}^{min} de los sistemas de verificación GMM(MAP), GMM(MAP)/UBM+SSM y GMM(MAP)/UBM+Uniform sobre Albayzín..... 144

Tabla 4.8. Tasa EER y coste C_{DET}^{min} de los sistemas de verificación GMM(MAP), GMM(MAP)/UBM+SSM y GMM(MAP)/UBM+Uniform sobre Albayzín y un conjunto de test que contiene señales con fuentes no modeladas..... 145

Tabla 4.9. Tasa EER de los sistemas de identificación GMM(EM), GMM(EM)/UBM+SSM, GMM(MAP) y GMM(MAP)/UBM+SSM y GMM(MAP) sobre Albayzín 147

Tabla 4.10. Tasa EER de los sistemas de verificación GMM(EM), GMM(EM)/UBM+SSM, GMM(MAP) y GMM(MAP)/UBM+SSM sobre Albayzín..... 148

Tabla 4.11. Tasa EER y costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de los sistemas GMM(MAP) y GMM-SVM sobre los conjuntos experimentales de la NIST SRE06 152

Tabla 4.12. Tasa EER, y costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} del sistema GMM(MAP), sin aplicar ninguna normalización de probabilidades, y con Z-norm, T-norm, ZT-norm y TZ-norm, sobre los conjuntos experimentales de la NIST SRE06..... 160

Tabla 4.13. Tasa EER, y costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} del sistema GMM-SVM, sin aplicar ninguna normalización de probabilidades, y con Z-norm, T-norm, ZT-norm y TZ-norm, sobre los conjuntos experimentales de la NIST SRE06..... 161

Tabla 4.14. Tasa EER y costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de los sistemas GMM(MAP), GMM-SVM y el sistema que resulta de la fusión de ambos – optimizada en cerrado –, sobre la tarea principal de la NIST SRE05..... 163

Tabla 4.15. Tasa EER y costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de los sistemas GMM(MAP), GMM-SVM y el sistema que resulta de la fusión de ambos sobre los conjuntos experimentales de la NIST SRE06 ... 165

Tabla 4.16. Tasa EER y costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de los sistemas GMM(MAP) y GMM-SVM, sin aplicar ninguna normalización de probabilidades, y con Z-norm y T-norm, sobre la tarea principal de la NIST SRE05. Se incluye el rendimiento obtenido al fusionar, en cerrado, los sistemas de mejor rendimiento: GMM(MAP)+Z-norm y GMM-SVM+Tnorm. 166

Tabla 4.17. Tasa EER y costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de los sistemas GMM(MAP)+Znorm, GMM-SVM+T-norm y la fusión de ambos sobre los conjuntos experimentales de la NIST SRE06. 166

Tabla 5.1. Precisión, cobertura y medida-F del sistema de seguimiento de locutores continuo (con UBM-t y UBM-g) y el sistema de referencia sobre el corpus AMI. 179

Tabla 5.2. Precisión, cobertura y medida-F del sistema de seguimiento continuo (con UBM-t) sobre el corpus AMI, sin suavizado y con suavizado triangular..... 181

Tabla 5.3. Precisión, cobertura y medida-F del sistema de seguimiento continuo (con UBM-g) sobre el corpus AMI, sin suavizado y con suavizado triangular..... 183

Tabla 5.4. Precisión, cobertura y medida-F del sistema de seguimiento continuo (con UBM-g) sobre el corpus AMI, incluyendo los segmentos en los que intervienen más de un locutor..... 185

Tabla 5.5. Precisión, cobertura y medida-F del sistema de seguimiento continuo (con UBM-t) sobre el corpus AMI, incluyendo los segmentos en los que intervienen más de un locutor..... 186

Listado de las herramientas de software

A continuación, se presenta el listado de las herramientas de software que se han empleado en este trabajo y una breve descripción con sus características principales.

ECJ	API escrito en Java para el desarrollo de aplicaciones de computación evolutiva y programación genética http://cs.gmu.edu/~eclab/projects/ecj/
LNKnet	Herramienta en C para la ejecución y el desarrollo de aplicaciones de clasificación de patrones mediante métodos estadísticos, redes neuronales y aprendizaje automático http://www.ll.mit.edu/mission/communications/ist/lnknet/index.html
FoCal	Módulo de software que proporciona varios módulos en Matlab para la fusión y calibración de sistemas de detección automática de locutores y del idioma http://niko.brummer.googlepages.com/jfocal
Sautrela	Entorno de desarrollo de software versátil para tecnologías del habla, escrito en Java http://gtts.ehu.es/TWiki/bin/view/Sautrela
DETware	Aplicación de software desarrollado en Matlab para la creación de curvas DET http://www.itl.nist.gov/iad/mig//tools/DETware_v2.1.targz.htm
SRE Tools	Módulo de software desarrollado en el entorno R para la estimación de las métricas de rendimiento de referencia de un sistema de detección de locutores http://sites.google.com/site/sretools/
libSVM	Módulo de software desarrollado en C++ y Java para la clasificación de patrones mediante máquinas de vectores soporte http://www.csie.ntu.edu.tw/~cjlin/libsvm/

** En cuanto a Sautrela, no sólo se ha utilizado sino que se ha contribuido a su desarrollo, particularmente en lo que se refiere a la parte de la parametrización de señales.

** Se ha escrito una versión de DETware para Scilab que se hará pública próximamente en la web de GTTS.

Capítulo 1

Introducción

En este capítulo se introducen aspectos relacionados con el contexto del trabajo realizado, las motivaciones y los objetivos de la tesis. También se incluye, de manera resumida, la estructura y el contenido de esta memoria.

1.1 Contexto y motivación

Sistemas biométricos: identificación de un individuo mediante el reconocimiento de patrones

La naturaleza del ser humano engloba un conjunto de características que posibilita la distinción de un individuo frente a otro, o de un grupo de individuos frente a otro. Algunas de estas características son fisiológicas – características intrínsecas, como por ejemplo las facciones de la cara, la forma corporal, el olor o la voz – y, otras en cambio, conductuales – relacionadas con habilidades adquiridas, como la forma de andar, la forma de escribir, el modo de interacción con los elementos del entorno, etc. De hecho, son éstas las características que nosotros, los seres humanos, analizamos (en muchas ocasiones de forma inconsciente) a la hora de identificar a una persona conocida.

La disciplina científica conocida como reconocimiento de patrones hace referencia al desarrollo de máquinas o sistemas que reconocen de forma automática una forma, una secuencia, una escena, etc. Ello requiere, en general, una fase de aprendizaje previa a la fase de reconocimiento. En la fase de aprendizaje o entrenamiento se extraen las características y se estima el modelo o patrón del objeto a reconocer. La fase de reconocimiento trata de clasificar un objeto desconocido, extrayendo sus características y asignándole la etiqueta o categoría del patrón (o modelo) más probable o parecido.

Se denomina sistema biométrico a aquél que reconoce patrones humanos. Para ello, extrae de forma automática características fisiológicas y/o conductuales de una persona con objeto de identificarla. En los últimos años, probablemente debido al potencial de aplicación que ofrecen los sistemas biométricos para la seguridad y la actividad forense, su rango de aplicación ha aumentado considerablemente. Entre las aplicaciones más comunes se encuentran: el control de acceso a edificios y lugares privados; el control de acceso a información privilegiada; la autenticación de usuarios en transacciones bancarias; la monitorización (detección y seguimiento) de usuarios en un espacio concreto; etc.

La voz: una característica biométrica de gran potencial

Las características biométricas se clasifican como fisiológicas o conductuales. La voz, curiosamente, es una característica fisiológica y conductual. Presenta características inherentes al locutor (la geometría del tracto vocal y la frecuencia fundamental, entre otras) y características adquiridas (los patrones de pronunciación, el vocabulario, el uso de la lengua, etc.) que se combinan y aparecen reflejadas en el habla. Una señal de voz transporta diversa información: el mensaje, el idioma, el género del locutor, su identidad, su estado emocional, su estado de salud, etc. Un sistema biométrico debe analizar toda esa información utilizando sólo aquellas características que le permitan identificar al locutor.

La voz resulta influenciada por fuentes de variabilidad voluntaria e involuntaria. Se habla de variabilidad voluntaria cuando el usuario tratar de modificar su forma de hablar. La variabilidad involuntaria proviene de la incapacidad de pronunciar dos frases de forma idéntica, del esfuerzo vocal que depende del ruido del entorno o de una infección respiratoria. La variabilidad externa generada por el ruido del entorno o por el canal de transmisión o grabación, es también involuntaria.

Reconocimiento de locutores

Un sistema de reconocimiento de locutores es un sistema que trata de determinar o verificar, por medio de una máquina, la identidad de un locutor. Dependiendo del tipo de decisión a tomar, se dividen en: sistemas de identificación, que determinan la identidad del locutor; y, sistemas de verificación, que determinan si el locutor es quien se presupone. Si el locutor que aparece en la señal de test (conocido como locutor objetivo o *target*) pertenece a un grupo predefinido de locutores, se dice que la identificación se realiza en un conjunto cerrado. Por el contrario, si el sistema asume que la señal puede pertenecer a un locutor desconocido, un "impostor" en la terminología al uso, se dice que la identificación se realiza en un conjunto abierto. La decisión se toma a partir de la comparación de la señal de test con los modelos de locutor estimados en la fase de entrenamiento. Cada modelo define las características del locutor al que representa. Dichas características deberían permanecer estables en las señales emitidas por un mismo locutor, y tomar valores muy diferentes en las señales emitidas por otros locutores. En cuanto a la aplicación de estos sistemas, además de los propios de la biometría (seguridad y actividades forenses), destacan la indexación por locutores de recursos multimedia para la

búsqueda de información y la adaptación al locutor de los modelos acústicos para mejorar el rendimiento de los sistemas de reconocimiento del habla.

En los últimos años, concretamente en las evaluaciones del NIST (actividad y organismo de los que hablaremos inmediatamente a continuación), se ha definido una tercera tarea denominada detección de locutores. Consiste en determinar si un locutor está presente en la señal que se analiza, pero el sistema debe contemplar la posibilidad de que en dicha señal puedan tener presencia varios locutores. La tarea de detección se puede definir, por tanto, como la verificación de un conjunto de locutores sobre una pronunciación en la que habla más de un locutor (en el siguiente apartado se da una visión más completa de todas las tareas establecidas en relación con el reconocimiento de locutores).

Cabe destacar que el elevado grado de madurez que presentan hoy en día las tecnologías de identificación y verificación de locutores, se debe, en gran medida, al avance tecnológico propiciado por las infraestructuras creadas a lo largo de los últimos 20 años (bases de datos, métricas, etc.). Entre ellas, merecen especial atención las campañas de evaluación organizadas periódicamente por el NIST (*National Institute of Standards and Technology*), una agencia federal del departamento de Comercio de Estados Unidos cuya misión se centra en promover la innovación tecnológica y científica de la industria de su país y del resto del mundo. El grupo dedicado a las tecnologías del habla, con objeto de evaluar el estado de arte de las principales tecnologías, organiza evaluaciones de sistemas de reconocimiento de locutores desde 1996. Son evaluaciones abiertas a todos los grupos de investigación y entidades interesadas en el reconocimiento de locutores independiente del texto. La última evaluación ha tenido lugar en Brno, Chequia, en junio de 2010.

Las evaluaciones celebradas en los últimos años se han centrado en analizar la influencia que ejerce la variabilidad debida al canal de grabación y al idioma en el rendimiento de los sistemas, empleando principalmente grabaciones de conversaciones telefónicas correspondientes a cientos de locutores. De los resultados obtenidos se concluye que este tipo de variabilidad perjudica de forma considerable el rendimiento de los sistemas (véase [Przybocki07]). En la evaluación celebrada en 2008 se incluyó un nuevo escenario donde el locutor objetivo es entrevistado cara a cara, y las grabaciones se realizan mediante diversos micrófonos ubicados en la misma sala donde se realizan las entrevistas. De los resultados obtenidos en dicha evaluación, se deduce que el rendimiento de los sistemas se deteriora al emplear un canal de grabación diferente en las pronunciaciones de entrenamiento y test. No obstante, en comparación con evaluaciones previas, se observa una mejora del rendimiento con respecto a la variabilidad de canal (véanse [Martin09] y los resultados oficiales en [NIST SRE]). Con el objetivo de analizar de forma más exhaustiva este comportamiento, el NIST organizó una evaluación posterior (denominada *follow-up evaluation*) abierta a todos los participantes en la evaluación del 2008. La tarea difería únicamente en el conjunto de test, que contenía una mayor cantidad de pronunciaciones grabadas a través de micrófonos de mesa y una mayor diversidad en cuanto al tipo de micrófonos empleados. Según [Martin09], los resultados obtenidos en esta última evaluación vuelven a poner de manifiesto la necesidad de compensar de manera más efectiva la variabilidad no debida al locutor.

Seguimiento y diarización de locutores

Además de las tareas de identificación, verificación o detección de locutores, se han definido también tareas de seguimiento y diarización de locutores, que responden ambas a la misma pregunta *¿Quién habla cuando?* ("*who spoke when*" en inglés), aunque con una diferencia importante:

- La tarea se considera de diarización de locutores (*speaker diarization*) si previamente no se dispone de datos de los locutores objetivo (ni siquiera se sabe cuántos son). Cada segmento se asocia con una clase definida en el mismo procedimiento de diarización.
- La tarea se considera de seguimiento de locutores (*speaker tracking*) si previamente se dispone de datos de los locutores objetivo. En este caso, para cada segmento detectado se debe proporcionar una identificación absoluta.

En la siguiente **Figura 1.1** se resumen los diferentes tipos de tareas que se han identificado en relación con el reconocimiento de locutores. La clasificación se realiza en función de: (1) la constitución del universo, es decir, de si el espacio de posibles locutores está formado únicamente por locutores conocidos, desconocidos o de ambos tipos; y (2) el objetivo de la tarea, que puede ser la detección de presencia de un locutor (tarea que puede partir de un locutor candidato o hipótetico) o la detección temporal de los locutores.

<table>
<tr><th>Universo</th><th colspan="2">Tarea</th><th>Denominación</th></tr>
<tr><td rowspan="2">Locutores conocidos</td><td>Detección de presencia de locutor</td><td>Sin candidato</td><td>Identificacion en cerrado</td></tr>
<tr><td colspan="2">Detección temporal de locutores</td><td>Tracking en cerrado</td></tr>
<tr><td rowspan="3">Locutores conocidos y desconocidos (impostores)</td><td rowspan="2">Detección de presencia de locutor conocido</td><td>Sin candidato</td><td>Identificacion en abierto</td></tr>
<tr><td>Con candidato</td><td>Verificación</td></tr>
<tr><td colspan="2">Detección temporal de locutores conocidos</td><td>Tracking en abierto</td></tr>
<tr><td>Locutores desconocidos</td><td colspan="2">Detección temporal de locutores</td><td>Diarización</td></tr>
</table>

Figura 1.1. *Clasificación de tareas relacionadas con el reconocimiento de locutores.*

El seguimiento y la diarización surgen, principalmente, de la necesidad de desarrollar herramientas de búsqueda y de transcripción automática para recursos digitales de audio y video. Estas herramientas tratan determinar qué señales o fragmentos de una señal cumplen un requisito concreto. Por ejemplo, que sean fragmentos correspondientes a una misma fuente acústica (voz/no-voz, música/habla, etc.) – tarea denominada segmentación acústica –, o en lo que a la diarización y seguimiento de locutores respecta, que sean fragmentos producidos por un único hablante – tarea denominada segmentación de locutores. En la bibliografía se distinguen tres campos de aplicación para el seguimiento y la diarización: (1) noticias o programas difundidos por radio y televisión; (2) grabaciones de reuniones; y

(3) conversaciones telefónicas. Hoy en día, la segmentación de reuniones en turnos de locutor constituye uno de los más importantes campos de aplicación, debido a que, sin olvidar el potencial de mercado que ofrece la transcripción automática y la posibilidad de realizar búsquedas, proporcionan numerosas posibilidades de investigación en el campo de las tecnologías del habla (habla espontánea, solapamientos, canales variables, etc.).

Inteligencia ambiental

En los últimos años ha recibido especial atención lo que se conoce como inteligencia ambiental (*Ambient Intelligence,* AmI) concepto impulsado por el grupo ISTAG (*Information Society Technologies Program Advisory Group*) de la Comisión Europea, que propone una nueva visión de la Sociedad de la Información orientada al usuario, con servicios más eficientes y una interacción más amigable. Surge al inicio del V Programa Marco en 1999 y se define en un informe publicado a comienzos de 2001, que establece un conjunto de escenarios y recomendaciones.

La visión AmI trata de abordar el desarrollo de entornos digitales que integren numerosos dispositivos electrónicos embebidos y transparentes para el usuario, pero sensibles a su presencia y adaptables a sus necesidades, costumbres o emociones. Es decir, un entorno AmI debe ser proactivo, adaptando su funcionalidad de acuerdo a la situación y las necesidades del usuario, y a veces, incluso anticipándose a sus requerimientos. El bienestar del ciudadano es su principal objetivo, por lo que busca tecnologías amigables, aceptadas socialmente, que posibiliten una relación más natural, transparente segura, sostenible y productiva del individuo y su entorno.

En la visión AmI convergen tres tecnologías clave: la computación ubicua, las comunicaciones ubicuas y las interfaces amigables e inteligentes. El avance tecnológico de los últimos años ha posibilitado el desarrollo de muchas de estas tecnologías, la mayoría no materializadas cuando se definió la visión. Por ejemplo, hoy en día se dispone ya, entre otros, de tecnologías de comunicación ubicuas y de sistemas electrónicos portables (conocidos como *wearable systems*), dispositivos electrónicos miniaturizados a dimensiones que permiten su integración prácticamente en cualquier objeto cotidiano, como ropa, muebles, etc. Por lo tanto, aunque todavía presenta ciertas deficiencias, según [Aarts09], no es la tecnología el principal inconveniente para hacer realidad la visión AmI. Además de las propias tecnologías que posibilitan su realización, el concepto AmI supone un comportamiento inteligente de las mismas. Es decir, requiere herramientas que doten de inteligencia al entorno, como por ejemplo las herramientas de interpretación y representación del contexto, herramientas que posibiliten una interacción natural y transparente, herramientas de computación afectiva o emocional, etc. Estas herramientas presentan un menor grado de madurez que las tecnologías básicas descritas más arriba, y por tanto constituyen un campo de investigación muy atractivo.

Reconocimiento y seguimiento de locutores en entornos de inteligencia ambiental

Las tecnologías del habla presentan propiedades muy adecuadas para cumplir parte del comportamiento inteligente que requiere un entorno AmI. Posibilitan el desarrollo de interfaces

amigables, que ofrecen una interacción natural con el usuario. Además, permiten conocer elementos que definen gran parte del contexto en una interacción oral: el contenido y el idioma del mensaje emitido, la identidad, la localización y el estado emocional del locutor, etc. Es más, las herramientas necesarias para su adquisición son muy poco intrusivas, ya que la voz se puede capturar mediante micrófonos y dispositivos de grabación que pueden colocarse en cualquier sitio. Esto hace interesante abordar la integración y aplicación de los sistemas basados en tecnologías del habla en los entornos AmI (tales como reconocedores de locutores y del idioma, transcriptores voz-texto, etc.). En concreto, el estudio de determinados aspectos que se relacionan con la integración de sistemas de reconocimiento de locutores, para detectar la presencia de ciertos usuarios, y adaptar o adecuar la funcionalidad de los servicios prestados es el principal objetivo de esta tesis. Los sistemas de identificación y verificación permiten realizar detecciones puntuales, mientras que los sistemas de seguimiento o diarización permiten una monitorización constante. Es evidente que el usuario de un entorno AmI tiene que hablar para que el reconocimiento se pueda llevar a cabo, de modo que, sin olvidar que en estos entornos el análisis de voz es un sistema más entre otros que se complementan en la tarea global, en ocasiones se deberán diseñar modos de interacción específicos que inciten a hablar a los usuarios.

Por otro lado, además de la instalación de la infraestructura necesaria (micrófonos embebidos y sistemas de procesamiento de las señales de voz adquiridas), la integración de reconocedores de locutor en un entorno AmI deberá cumplir una serie de requisitos:

- El coste computacional debe ser el mínimo posible y proporcionar un rendimiento aceptable. Se trata de entornos que requieren sistemas de latencia baja, que den sus respuestas prácticamente en tiempo real. La transparencia y proactividad requeridas en estas situaciones impiden el aplazamiento de la interacción con el usuario, sobre todo si el retraso es debido al coste computacional que conlleva el reconocimiento. Por otro lado, es previsible que en un entorno de este tipo, la adquisición y procesamiento de las señales de voz se lleve a cabo en no pocas ocasiones a través de un dispositivo portátil o embebido, con limitaciones de almacenamiento y de capacidad computacional. Por tanto, se deben aligerar estos costes, tratando de producir el menor impacto posible en su eficiencia.
- La disminución del coste computacional conlleva el uso de metodologías de modelado y detección eficientes y fácilmente adaptables. En este sentido, no parece razonable realizar largas y complejas sesiones de adaptación, haciéndole repetir al usuario un gran número de elocuciones para obtener modelos robustos.
- La representación digital de las señales acústicas debe ser lo más compacta y lo menos redundante posible (sin correlaciones). A partir de una señal muestreada se estima una secuencia de vectores de parámetros acústicos. De esta forma, se obtiene una nueva representación más compacta, menos redundante y más adecuada a los modelos. La reducción del conjunto de parámetros acústicos conlleva la disminución del coste computacional de la clasificación (en cuanto a

velocidad y memoria), y en función de los datos empleados para la estimación de los modelos, puede incluso aumentar la robustez del sistema.

- Una de las características de diseño más importantes de un sistema AmI es la continuidad (característica denominada comúnmente funcionamiento *online*). La ejecución del sistema debe ser de duración indefinida, de forma que una vez iniciado no se pare prácticamente nunca. Además, deben ser sistemas auto-suficientes, de administración-cero (auto-configurables), y de hecho, deben adaptar su comportamiento a las necesidades y situaciones que vayan surgiendo. En lo que al reconocimiento de locutores respecta, deben ser sistemas continuos y de latencia baja, detectando de forma automática la presencia de nuevos locutores objetivo y la ausencia prolongada de locutores previamente detectados.

De todos los entornos de aplicación de la visión AmI, para esta tesis, y donde ha sido preciso, se ha elegido el hogar inteligente, dado que se dispone de un entorno de este tipo en nuestros laboratorios donde se podrían llegar a probar los resultados de este trabajo. El desarrollo de hogares inteligentes ha suscitado un gran interés en el mundo de la investigación, con numerosos proyectos y laboratorios de experimentación dedicados a ello en los últimos años, entre los que se destacan, entre otros: el proyecto *AwareHome* [AwareHome] del Instituto Tecnológico de Georgia; el laboratorio *Gator Tech Smart House* [Helal05] desarrollado por la Universidad de Florida; el laboratorio *PlaceLab* [Intille06] definido por el consorcio *House_n* del instituto MIT (*Massachusetts Institute of Technology*); el laboratorio *HomeLab* [Ruyter05] desarrollado por Philips; y el proyecto *Amigo* [Amigo] llevado a cabo por varias empresas y entidades de investigación europeas con el objetivo de abordar la integración de la visión AmI en hogares conectados. De todas formas, el trabajo realizado trata de tener siempre en cuenta la posible migración a otros escenarios de aplicación que van más allá del hogar inteligente.

1.2 Objetivos

Esta tesis se ha desarrollado en el Departamento de Electricidad y Electrónica de la Facultad de Ciencia y Tecnología de la Universidad del País Vasco (UPV/EHU), con la colaboración con el Centro de Investigaciones Tecnológicas Ikerlan. Es una tesis llevada a cabo en un marco de trabajo colaborativo universidad-empresa, marco que pretende abarcar no sólo la investigación básica, sino también la transferencia de tecnología. De hecho, se pretende priorizar el estudio y desarrollo de tecnologías que tienen aplicabilidad en el mercado.

En este sentido, a partir de las condiciones y requisitos que se han descrito en el apartado anterior, y habiendo establecido como objetivo principal el estudio de aspectos orientados a la integración de los sistemas de reconocimiento de locutores en los entornos AmI, esta tesis se centra en las tres líneas de trabajo que se mencionan a continuación:

(1) El estudio de la posibilidad de reducir la dimensionalidad de la representación acústica de las señales, mediante la búsqueda del conjunto óptimo de parámetros acústicos para el

reconocimiento de locutores. La mayoría de los sistemas de procesamiento del habla emplean conjuntos de parámetros acústicos muy similares, sean reconocedores del habla, reconocedores del locutor o idioma, etc. Es deseable contar únicamente con parámetros que sean relevantes para la clasificación de locutores, e ignorar características redundantes que aporten información innecesaria. Por tanto, en vez de analizar diferentes métodos de extracción de parámetros para determinar el más adecuado, tarea para la cual se pueden encontrar varios estudios comparativos en la bibliografía, en este trabajo se plantea la selección de los parámetros acústicos más relevantes para la clasificación de locutores, partiendo de un conjunto de parámetros acústicos estándar. Se trata de obtener conjuntos reducidos de parámetros que permitan ahorrar esfuerzo computacional con una baja degradación o incluso una mejora del rendimiento del sistema. Para ello, el proceso de búsqueda debe establecer un compromiso entre el coste computacional y el rendimiento de clasificación del sistema, que dependen respectivamente de la dimensión del conjunto de parámetros y la capacidad de discriminación del sistema al emplear el conjunto reducido. En la medida de lo posible, la reducción de parámetros no debe perjudicar la capacidad de discriminación del sistema. La reducción de dimensionalidad permitirá emplear conjuntos reducidos de parámetros en situaciones que así lo requieran. Por otro lado, esta parte del trabajo requiere el análisis de diferentes metodologías de reducción de dimensionalidad, o también diferentes procedimientos de selección de un subconjunto óptimo de parámetros.

(2) En relación con las tareas de identificación y verificación de locutores, se llevarán a cabo el análisis, desarrollo y evaluación de las metodologías de modelado, que son de referencia, y han sido definidas en los últimos años. Este estudio estará orientado a comparar el rendimiento de dichas metodologías, así como a analizar su robustez ante situaciones imprevistas, como por ejemplo la presencia de una señal cuya fuente es desconocida para el sistema. Asimismo, se deberán preparar las infraestructuras necesarias para la evaluación de un sistema de reconocimiento de locutores, haciendo especial hincapié en bases de datos y métricas, obtenidas mediante la participación en las evaluaciones organizadas a nivel mundial. Hoy en día han adquirido especial importancia las evaluaciones organizadas por el NIST, y de hecho, la mayoría de los trabajos de investigación relacionados con la detección de locutores utilizan las bases de datos definidas para estas evaluaciones. Es más, las métricas empleadas para analizar el rendimiento de los sistemas, son también generalmente las aplicadas en estas evaluaciones. Por lo tanto, el análisis y desarrollo de metodologías tiene como objetivo prioritario la participación en estas evaluaciones, en concreto, en la NIST SRE de 2008, debido a que establecen un marco de evaluación global, extendido y aceptado a nivel mundial.

(3) Se encuentra también entre las líneas de trabajo establecidas, el desarrollo de un sistema de seguimiento de locutores continuo, de latencia baja, diseñado para su aplicación en un hogar inteligente (como escenario AmI). El seguimiento de usuarios es una tarea básica en un sistema AmI, ya que conociendo la localización e identidad de los usuarios el sistema puede adecuar la

funcionalidad de los servicios prestados. Por tanto, se trata de desarrollar un sistema que monitorice por voz a los habitantes y usuarios presentes en la casa, para de esta forma proporcionar la información necesaria a los servicios de adaptación (adecuación) al usuario. Puesto que en un escenario de este tipo, la cantidad de locutores que debe monitorizar el sistema no suele ser muy elevada, la aproximación ideada parte de la condición de que los usuarios a seguir son conocidos: habitantes o invitados asiduos. Esta es la principal razón por la que se ha planificado una aproximación de seguimiento, y no de diarización. Por otro lado, entre los objetivos principales que marca esta línea de trabajo destacan, por un lado, el estudio de los sistemas de seguimiento de locutores de la bibliografía así como las infraestructuras necesarias para su evaluación (de nuevo, principalmente, las bases de datos, métricas y evaluaciones estándares), y por otro lado, el diseño y la implementación de una aplicación de software que ejecute la aproximación definida. La aplicación desarrollada seguirá el paradigma SOC (*Service Oriented Computing*), que engloba SOA (*Service Oriented Architecture*), una estructura lógica distribuida e interoperable para el diseño de aplicaciones a partir de servicios como componentes de software. En consecuencia, la aplicación desarrollada consistirá en un servicio de seguimiento de locutores basado en el estándar SOA. Por definición, este servicio se podrá integrar en cualquier otro servicio o aplicación del entorno que siga las especificaciones SOA.

1.3 Estructura de la memoria

Además de este capítulo introductorio y del de conclusiones y trabajo futuro, esta memoria consta de 4 capítulos, todos ellos orientados a la integración de los sistemas de reconocimiento de locutores en entornos AmI, si bien cada capítulo aborda una línea o materia concreta.

El **Capítulo 2** presenta, desde una perspectiva general, el estado del arte en las áreas de investigación estudiadas durante el desarrollo de esta tesis. Es un estado del arte de los elementos de que consta un sistema de reconocimiento de locutores, las metodologías de referencia definidas en los últimos años y las infraestructuras necesarias para evaluar su rendimiento. El estado del arte específico de las diferentes líneas de trabajo seguidas en la tesis se incluye en los capítulos correspondientes. Por tanto, este capítulo comienza con la descripción de la visión AmI y su relación con las tecnologías del habla, detallando cómo han evolucionado tanto la visión, como las tecnologías que engloba, en el periodo transcurrido desde su formulación inicial. Asimismo, se resumen las principales características de los proyectos de investigación elaborados en relación con hogares inteligentes en los últimos años y las aproximaciones más destacadas para el seguimiento de usuarios. El capítulo continúa con las tareas y los procedimientos relacionados con un sistema de reconocimiento de locutores, tales como la identificación y verificación de locutores, la parametrización acústica, y las metodologías de modelado y la clasificación de patrones (o toma de decisiones). Se describen también, tanto las técnicas de normalización de probabilidades propuestas recientemente, que buscan minimizar la sensibilidad a la variabilidad acústica de las señales y a la variabilidad de distintos locutores, como los procedimientos

definidos en los últimos años para la fusión de los sistemas de reconocimiento de locutores. El capítulo sigue con la descripción de las metodologías de segmentación y clasificación de locutores que emplean los sistemas de seguimiento de referencia. La mayoría de estos sistemas operan de forma *offline*, sobre un recurso de audio previamente grabado. Debido a que uno de los objetivos de la tesis consiste en desarrollar un sistema continuo (*online*), se analizan en la bibliografía, las aproximaciones continuas para el seguimiento de locutores. El capítulo concluye con el estudio de las infraestructuras necesarias para la evaluación de los sistemas: las bases de datos, métricas de rendimiento y campañas de evaluación. Se trata de elementos que establecen un marco de desarrollo común a toda la comunidad científica, y por ello se incluyen en el estado de arte.

El **Capítulo 3** se centra en la reducción de dimensionalidad del espacio acústico para tareas relacionadas con el reconocimiento de locutores. En la primera parte del capítulo, se describen diferentes estrategias para ello, entre las cuales se destacan las transformaciones lineales y la aplicación de algoritmos genéticos (*Genetic Algorithm*, GA) para la búsqueda del subconjunto óptimo de parámetros. Las transformaciones lineales constituyen una metodología muy empleada para la reducción de dimensionalidad de toda clase de representaciones en espacios vectoriales, ya que combinan la selección y extracción de parámetros en una única fase. En este estudio, se aplican las clásicas transformaciones lineales PCA (*Principal Component Analysis*) y LDA (*Linear Discriminant Analysis*); y una aproximación más reciente denominada HLDA (*Heteroscedastic Linear Discriminant Analysis*), que es una generalización de LDA. La aplicación de GA para la búsqueda del subconjunto óptimo de parámetros es una aproximación propuesta en esta tesis, ya que permite realizar búsquedas sin establecer límites a toda la combinatoria posible. Los GA son un conjunto de técnicas de búsqueda heurística y aleatoria inspiradas en las estrategias de la evolución biológica. En este trabajo, se definen dos aproximaciones: la selección directa mediante GA [Zamalloa08b], y de forma alternativa, la selección de parámetros mediante un ranking de pesos [Zamalloa06a], estimados mediante un proceso de búsqueda con GA de los pesos óptimos correspondientes a un conjunto concreto de parámetros acústicos [Zamalloa06b]. Con el objetivo de comparar el rendimiento de los GA y las transformaciones lineales como técnicas de reducción de dimensionalidad de la representación acústica de las señales, se define un conjunto experimental sobre las bases de datos de habla castellana Albayzín y Dihana que contienen distintos tipos de habla y provienen de canales diferentes [Zamalloa08b] [Zamalloa08c]. Los resultados obtenidos se describen en la segunda parte del capítulo, que finaliza con las principales conclusiones obtenidas de la experimentación realizada.

En el **Capítulo 4** se describen los sistemas desarrollados para analizar el rendimiento de diferentes metodologías de modelado en relación con las tareas de identificación y verificación de locutores. La evaluación se ha llevado a cabo sobre la base de datos Albayzín (habla castellana) grabada en condiciones de laboratorio, y el corpus definido para la evaluación del NIST del 2006, un corpus constituido por conversaciones telefónicas y que presenta una gran diversidad en cuanto a canales de grabación, locutores e idiomas. El capítulo incluye, en primer lugar, el estudio realizado sobre

Albayzín, que compara el rendimiento de dos metodologías de modelado básicas para las tareas de identificación en conjuntos abiertos y verificación: GMM (*Gaussian Mixture Model*) estimados mediante el algoritmo EM (*Expectation-Maximization*) y GMM adaptados a partir de un modelo universal (*Universal Background Model*, UBM) por medio del algoritmo MAP (*Maximum A Posteriori*). De forma adicional, se analiza el comportamiento de estas metodologías cuando hay un desajuste acústico (*mismatch*) entre las señales de entrenamiento y test. Para ello, se añade al conjunto de test de Albayzín, un subconjunto de señales compuesto por señales de música, conversaciones telefónicas y audio descargado al azar de internet. Dentro de esta línea de investigación, que trata de evaluar la robustez de las metodologías en la detección de señales cuya fuente no se ha modelado en entrenamiento, se ha propuesto una nueva metodología de modelado que pretende dar cobertura a fuentes no modeladas [Zamalloa08e][Zamalloa08f]. El estado del arte, la descripción y evaluación de la nueva metodología, denominada modelo superficial de fuente (*Shallow Source Model*, SSM), se encuentra en esta primera parte del tercer capítulo. A continuación, en la segunda parte, se detalla la experimentación realizada sobre el corpus del NIST del 2006. Se trata de una tarea de detección o verificación, concretamente la tarea principal de las evaluaciones del NIST. Los sistemas desarrollados se basan en dos técnicas de modelado que han demostrado muy buen rendimiento en los últimos años. Por un lado, los modelos de locutor GMM adaptados a partir de un UBM, y por otro lado, modelos de locutor consistentes en máquinas de vectores soporte (*Support Vector Machines*, SVM) definidas sobre el espacio vectorial formado por los parámetros de un GMM adaptado. Asimismo, se aplican las técnicas de normalización de probabilidades Z-norm, T-norm, y ZT-norm para tratar de determinar y comparar su rendimiento. La segunda parte del capítulo concluye con la fusión mediante regresión logística de los sistemas desarrollados. Se aplica con objeto de mejorar la capacidad de discriminación del sistema y calibrar debidamente las probabilidades obtenidas. Cabe destacar que los sistemas desarrollados en esta parte del trabajo sirven de soporte al trabajo desarrollado posteriormente en seguimiento de locutores.

El **Capítulo 5** es el dedicado al sistema de seguimiento de locutores continuo y de baja latencia, propuesto en esta tesis, para su aplicación en hogares inteligentes. Comienza con la descripción del sistema de seguimiento propuesto, que integra una cierta cantidad de módulos que dan cobertura a una función concreta del sistema de seguimiento (la parametrización, la detección, la calibración, el seguimiento y el suavizado) [Zamalloa10a]. Entre ellos, destaca el modulo de suavizado, propuesto y desarrollado de forma íntegra en esta tesis [Zamalloa10b], que trata de mejorar la robustez del sistema frente a la variabilidad local de las probabilidades estimadas. Asimismo, dentro del módulo de seguimiento, se han propuesto dos metodologías para la toma de decisiones: una sigue la empleada en tareas de identificación (el segmento analizado se asocia directamente con el locutor de máxima probabilidad), y la otra, en cambio, la empleada en tareas de verificación (si la probabilidad del locutor, dado el segmento de entrada, es superior a un umbral predeterminado, el sistema determina que dicho locutor está presente en el segmento). A continuación, con objeto de evaluar el rendimiento

del sistema (con todas sus posibles variaciones), el capítulo sigue con la descripción de la fase experimental, llevada a cabo sobre la base de datos de reuniones AMI, un conjunto de datos multimodal con interacciones humanas en tiempo real, en el contexto de reuniones grabadas en un entorno inteligente. En primer lugar, se especifica la configuración experimental, donde se detallan las particiones definidas para el desarrollo y la evaluación, las metodologías empleadas para modelar los locutores (todas ellas basadas en GMM adaptados a partir de un UBM), y el sistema de referencia elaborado como sistema de contraste de la propuesta realizada. El sistema de referencia sigue una aproximación estándar y opera de forma *offline*. Después de la configuración experimental, se describen los resultados obtenidos, comparando el rendimiento del sistema propuesto (empleando tanto la aproximación de identificación como la de verificación) con el rendimiento del sistema de referencia. A continuación, se analiza el comportamiento del módulo de suavizado, y por último, se estudia el rendimiento del sistema con la aproximación de verificación para la detección de segmentos solapados en los que hablan dos o más locutores. El capítulo finaliza con el proceso de implantación del sistema, donde se describen la arquitectura y la fase de implementación de la aplicación desarrollada. Consta de un servicio de seguimiento de locutores basado en el estándar SOA que, a su vez, se compone de varios tipos de servicios.

Capítulo 2

Estado del arte

2.1 Inteligencia Ambiental

La inteligencia ambiental (AmI) es un término impulsado por el grupo ISTAG de la Comisión Europea. Propone una nueva visión de la *Sociedad de la Información* orientada al usuario, con servicios más eficientes y una interacción más amigable. Para ello, hace referencia a entornos sensibles a la presencia de usuarios y adaptables a sus necesidades [Cook09]. Su principal objetivo es mejorar la calidad de vida del ser humano en cuanto al cuidado, el bienestar, la educación y la creatividad. Ello requiere el desarrollo de productos y servicios socialmente aceptados y fáciles de utilizar [Aarts09].

La visión AmI generaliza el concepto planteado por Mark Weiser en 1991, denominado "Computación Ubicua" *(Ubiquitous Computing)*, que se fundamenta en la integración de los resultados tecnológicos obtenidos en varias áreas de la informática y la física, entre las que cabe destacar la comunicación y computación ubicua, las interfaces inteligentes y la domótica [Weiser91]. Weiser defiende que las tecnologías más potentes son aquellas que desaparecen, las que se entrelazan en el tejido de la vida cotidiana hasta que se hacen invisibles. De hecho, a partir de la miniaturización, plantea la integración de la tecnología computacional en objetos cotidianos. Define espacios rodeados de dispositivos omnipresentes dedicados a tareas cotidianas del hogar y el trabajo, de forma transparente para el usuario, es decir, de forma invisible y no-intrusiva. Con un enfoque ligeramente diferente, la computación ubicua es también llamada "pervasiva" (*Pervasive Computing*). Es un término que ha introducido la industria, debido a que la definición de Weiser tiene un enfoque académico e idealista. Básicamente, la computación pervasiva consiste en aplicar la computación ubicua en las tecnologías Web [Hansmann01]. Por otro lado, el concepto AmI y la computación ubicua, tal como especifica ISTAG, difieren en que AmI está más orientada a la usabilidad de la tecnología que a la propia tecnología [ISTAG03].

La visión AmI surge al inicio del V Programa Marco en 1999 [ISTAG99], y se define en un informe publicado a comienzos de 2001 que determina un conjunto de escenarios y recomendaciones [ISTAG01].Los escenarios describen la integración de AmI en situaciones de la vida cotidiana. A modo de ejemplo, a continuación se describe parte de un escenario donde una mujer llamada Carmen planifica su jornada:

> "Por las mañanas, después de levantarse, Carmen entra en la cocina AmI y planifica sus traslados que siempre realiza en transporte público. Querría salir hacia el trabajo en media hora; solicita a AmI, mediante comandos de voz, que inicie la búsqueda de un servicio de transporte compartido para el trayecto. AmI encuentra un vehículo que pasará en cuarenta minutos. El biosensor integrado en el vehículo detecta que el chofer no fuma (uno de los requisitos de Carmen) por lo que confirma la reserva. A partir de este momento, Carmen y el chofer permanecen en contacto para dar solución a cualquier imprevisto que surja (retrasos, accidentes, etc). Mientras desayuna, Carmen recuerda que esta noche tiene invitados. Se lo comunica a AmI y le menciona que querría cocinar una tarta para ellos. AmI muestra la lista de la compra y la receta en el monitor del e-frigorífico. En la receta ilumina los ingredientes que faltan (leche y huevos). Carmen completa la lista de la compra (incluye la leche y los huevos) y pide que sea depositada en el punto de distribución más cercano de la vecindad (Carmen podrá recoger la compra en cualquier momento). Todos los productos poseen una etiqueta identificativa inteligente, por lo que Carmen podrá seguir y chequear su compra desde cualquier dispositivo habilitado y desde cualquier lugar (en casa, en la oficina, etc). En todo momento, AmI le mantendrá informada sobre el estado y posibles alternativas de su compra, sugiriéndole cambios en base al stock y las ofertas detectadas... (continúa)"

AmI proviene de la convergencia de tres tecnologías clave:

(1) La **computación ubicua**: se requiere que el hardware no sea intrusivo y esté integrado en objetos cotidianos como los envases, la ropa, los materiales inteligentes o incluso las partículas decorativas como la pintura, etc..

(2) Las **comunicaciones ubicuas**: en un entorno AmI la comunicación debe ser ubicua, llevada a cabo mediante infraestructuras transparentes. Las redes serán dinámicas y distribuidas, formadas por un número indefinido de dispositivos. Algunas redes serán cableadas, y otras, en cambio, inalámbricas, algunas de ellas serán redes móviles, y muchas otras fijas. Es imprescindible que las redes se puedan configurar *ad-hoc*, adecuándose a una tarea específica, probablemente de corta duración. Esta complejidad pone de manifiesto la necesidad de soluciones "Plug and Play" inalámbricas.

(3) Las **interfaces amigables e inteligentes**: uno de los principales pilares de AmI, es el diseño de sistemas intuitivos que ofrezcan una interacción transparente y natural mediante interfaces amigables. Las interfaces han de ser multiusuario, multilingües y multimodales.

Los informes publicados por el ISTAG, además de proporcionar una descripción conceptual de la visión AmI, identifican las tecnologías necesarias para afrontar dicho reto [ISTAG03]:

- Materiales inteligentes
- Sistemas micro-electromecánicos
- Tecnologías sensoriales
- Sistemas embebidos
- Comunicaciones ubicuas
- Tecnologías de interacción
- Softwares adaptativos

Según [Aarts09], en los últimos diez años, estas tecnologías han evolucionado considerablemente, y aunque todavía presentan ciertas deficiencias, éstas no son el principal inconveniente para hacer realidad la visión AmI. Hoy en día, se dispone ya (obviamente, con diferentes grados de madurez tecnológica), de sistemas electrónicos portables (conocidos como *wearable systems*), comunicaciones ubicuas e interfaces de diálogo, entre otros. Asimismo, en lo que a sistemas embebidos respecta, se puede decir que el diseño y la producción de dispositivos electrónicos (con capacidad de procesamiento, comunicación y almacenamiento) han alcanzado prácticamente el nivel de miniaturización necesario para su integración en cualquier objeto cotidiano, como la ropa, los muebles, los coches o el hogar, lo que posibilita la creación de espacios inteligentes. No obstante, no toda la tecnología necesaria está desarrollada. Queda mucho por hacer en varios aspectos: la autonomía y necesidades energéticas de los dispositivos tecnológicos, el control distribuido de las redes inalámbricas, el desarrollo de sistemas o algoritmos de software inteligentes auto-suficientes de duración indefinida (*never-ending*), etc.

Pero el concepto AmI, además de las propias tecnologías que posibilitan su realización, engloba un comportamiento inteligente de las mismas, y de hecho, requiere la presencia de herramientas que doten al entorno de inteligencia. Entre estas herramientas, en los informes del ISTAG se listan las siguientes:

- Gestión y administración de los medios del entorno
- Interacciones naturales
- Inteligencia computacional
- Conocimiento del contexto
- Computación afectiva o emocional

En cuanto al potencial de aplicación, son herramientas que presentan un menor grado de madurez que las tecnologías descritas más arriba. Se trata de herramientas muy relacionadas con la inteligencia artificial (*Artificial Intelligence*, AI), lo que pone de manifiesto el vínculo estrecho entre AmI y AI. En

[Ramos08] se argumenta que el concepto AmI supone un nuevo reto para AI, y que dicho concepto guiará la evolución de varias metodologías dentro del campo. En el mismo trabajo se prevé que la aceptación social de AmI se hará realidad con una equilibrada combinación de tecnologías básicas y tecnologías AI. Se pone como ejemplo determinadas tareas planteadas en los entornos o escenarios AmI (como los escenarios definidos en [ISTAG01]) que requieren la participación de metodologías y técnicas del campo AI. Entre ellas destacan la interpretación del contexto, la interacción con las personas, los sistemas multi-agente, la representación del conocimiento, etc. (véase [Ramos08] para un listado de tareas más extenso).

En relación con la interacción y la interpretación del contexto, se debe tener en cuenta que un entorno AmI debe ser amigable y proactivo, adaptando su funcionalidad, de acuerdo a la situación del momento y las necesidades de los usuarios, a veces, incluso anticipándose a sus requerimientos. Para poder adaptarse, el sistema debe conocer el contexto. Por contexto entendemos cualquier información que caracterice la situación de una entidad. Por entidad entendemos una persona, un lugar, o un objeto considerados relevantes para la interacción entre el usuario y la aplicación (incluso el propio usuario y la propia aplicación) [Dey01]. Al mismo tiempo, es imprescindible que el sistema monitorice al usuario y determine las acciones a llevar a cabo, aplicando para ello estrategias de aprendizaje automático.

La voz es una de las principales herramientas para la comunicación humana, por lo que constituye un medio natural y poco intrusivo para la personalización y adaptación de un sistema. La voz permite identificar características que definen gran parte del contexto en una interacción oral: el contenido y el idioma del mensaje emitido, la identidad, la localización y el estado emocional del locutor, etc. Además, generalmente, el ser humano se comunica mediante el habla, por lo que cabe esperar que pretenda interaccionar de forma similar con un entorno AmI. De todas formas, cabe considerar otros modos de interacción como por ejemplo la vista o el tacto. Se pueden aplicar metodologías de visión por ordenador para reconocer gestos o expresiones del rostro facial, y así detectar las necesidades o el estado emocional del usuario. Asimismo, se pueden emplear interfaces hápticos (que analizan el sentido del tacto) para percibir las necesidades del usuario y guiar su interacción.

De la transparencia requerida en un entorno AmI, surge la necesidad de abordar la integración y aplicación de los sistemas que procesan el habla (reconocedores del locutor y del idioma, transcriptores voz-texto, etc). Esta tesis se centra en la integración de sistemas de reconocimiento de locutores en entornos AmI. La detección del locutor puede realizarse de forma continua (seguimiento/monitorización en tiempo real) o bien sólo en momentos puntuales (identificación o verificación). En cualquier caso, el coste computacional que conlleva la identificación no debe retrasar la interacción con el usuario, y tampoco parece razonable realizar largas y complejas sesiones de adaptación, haciéndole repetir al usuario un gran número de elocuciones para obtener un perfil robusto. De hecho, el usuario podría haber accedido al servicio prestado a través de un dispositivo

portátil o embebido, con limitaciones de almacenamiento y capacidad computacional, lo que plantea la necesidad de aligerar el coste computacional con el menor impacto posible en la eficiencia.

El concepto AmI abarca diversos entornos de aplicación; tanto los relacionados con la industria y el comercio, como los relacionados con la vida cotidiana, tales como el hogar, la calle, los servicios de transporte y servicios de ocio públicos. De estos entornos, el escenario objeto de esta tesis es el seguimiento de usuarios en hogares inteligentes. Cabe destacar que el hogar es un escenario con gran potencial de aplicación del concepto AmI, debido a que es el lugar principal de descanso y ocio.

A continuación, se describen algunos de los trabajos de investigación más destacados en relación con el desarrollo de aplicaciones para hogares inteligentes. Se añade también un pequeño estado del arte de las principales tecnologías que engloban las aplicaciones de seguimiento de usuarios.

2.1.1 Hogares inteligentes

En los últimos años, se han llevado a cabo numerosos proyectos de investigación que tienen como objetivo desarrollar las tecnologías necesarias para hacer realidad el concepto de hogar inteligente. Entre ellos, destacan los siguientes:

- El proyecto *Aware Home* [AwareHome], desarrollado por el instituto tecnológico de Georgia, trata de definir sistemas con conocimiento del contexto para hogares inteligentes. Para ello, han construido un hogar inteligente con dos espacios habitables. Cada espacio consta de dos dormitorios, dos baños, una oficina, una cocina, un comedor, una sala de estar, un lavadero, un área común de entretenimiento y una habitación de control para la centralización de todos los servicios informáticos instalados en el espacio. Entre las diversas tecnologías desarrolladas para el hogar inteligente destaca el seguimiento de usuarios mediante sensores de presión embebidos en el suelo [Orr00]. Esta tecnología trata de monitorizar la posición del usuario y en la medida de lo posible identificarlo.
- El entorno experimental *Gator Tech Smart House*, desarrollado por la Universidad de Florida [Helal05], está diseñado principalmente para el cuidado de personas discapacitadas y personas mayores. El principal objetivo del proyecto consiste en desarrollar entornos de asistencia que detecten su propia situación, así como la de sus usuarios, para proporcionar servicios de intervención y monitorización remota. El hogar se compone de varios tipos de sensores y actuadores, entre los que caben destacar: etiquetas RFID *(Radio Frequency Identification*) para el control de acceso al hogar o para controlar el lavado o estado de las ropas; sensores ultrasónicos instalados en la sala de estar para detectar los movimientos, la localización y la orientación de los usuarios presentes; sensores de presión ubicados en el suelo de la cocina y la sala de entretenimiento para realizar el seguimiento de usuarios, así como para detectar caídas; y un asistente cognitivo que pretende guiar las tareas a realizar mediante mensajes visuales y orales.
- El proyecto MAVHOME *(Managing an Adaptive Versatile HOME)*, desarrollado por la Universidad de Texas en Arlington, emplea un sistema de aprendizaje para automatizar la toma de

decisiones en un hogar inteligente [Youngblood07]. El entorno AmI se entiende como un agente inteligente que percibe la situación del entorno mediante sensores y actúa en él empleando controladores de línea eléctrica. Para la implementación y evaluación de las tecnologías desarrolladas se han elaborado dos espacios inteligentes: el primero, denominado *MavLab*, simula un entorno de oficinas con cabinas, un área de descanso, una sala de estar y una sala de conferencias; el segundo, denominado *MavPad*, es un apartamento en el que viven tres personas. Para percibir el contexto del entorno se emplean diferentes sensores: sensores de luz, temperatura, humedad, movimiento, apertura/cierre de ventanas y puertas, etc.

- El laboratorio IDORM (*Intelligent DORMitory*), desarrollado por la Universidad de Essex (Gran Bretaña) se creo con el objetivo de analizar la integración de agentes de forma embebida en espacios inteligentes [Hagras04]. El laboratorio está compuesto por diferentes áreas de descanso, trabajo y ocio, y equipado con diversos sensores (de temperatura, ocupación, temperatura, humedad, etc.), actuadores, procesadores y redes. Uno de los 3 agentes definidos en el sistema se encarga de controlar el entorno físico, analizando para ello la temperatura, la iluminación o la actividad que está desarrollando el usuario. La actividad desarrollada por el usuario se detecta a partir de sensores de presión instalados en la cama y la silla del escritorio. Asimismo, para analizar si el usuario está trabajando o viendo algún vídeo de entretenimiento en el ordenador de la habitación, se analiza el tráfico de la red.
- El laboratorio *PlaceLab* es un hogar inteligente con habitantes reales, definido y construido por el consorcio *House_n* del MIT a mediados de 2004 [Intille06]. El laboratorio se define con objeto de recolectar datos sensoriales sobre actividades domésticas. PlaceLab es un apartamento con una sala de estar, un comedor, una cocina, una pequeña oficina un dormitorio y dos baños. El apartamento está equipado con varios tipos de sensores y tecnologías que recogen todas las actividades en la casa: interruptores que detectan los eventos de apertura/cierre; sensores de humedad, luz, presión; sensores que miden el flujo de gas y el flujo de agua; acelerómetros inalámbricos portados por los habitantes que controlan el movimiento de las extremidades; receptores que perciben el movimiento de objetos, etc.; y diversas cámaras y micrófonos distribuidos a lo largo de todo el apartamento que, junto con el resto de tecnologías, pretenden realizar el seguimiento de los habitantes (véase [Intille06] para obtener el listado específico de las tecnologías aplicadas).
- El laboratorio *HomeLab*, desarrollado por Philips, es un hogar inteligente que consta de una sala de estar, dos dormitorios, un baño y un estudio [Ruyter05]. Está equipado con diversas cámaras y micrófonos ubicados en el techo para localizar e interactuar con los usuarios. Su labor se centra en la interacción entre el usuario y el entorno, con objeto de mejorar la calidad de vida con respecto a la comunicación entre los usuarios y el exterior, y el equilibrio y la organización de tareas cotidianas. Una de las principales tecnologías analizadas en HomeLab son las pantallas interactivas de gran tamaño que se proyectan en espejos, ventanas, paredes etc. de todo el espacio

inteligente. El espacio incluye dispositivos de control remotos y locales para monitorizar la iluminación de la casa. De la fusión de estas dos tecnologías ha surgido el concepto *Living Light* definido por Philips, que consiste en insertar efectos de luz artificial y ambiental en las emisiones de música y vídeo. Otras tecnologías desarrolladas son espejos interactivos con interfaces visuales y táctiles, juguetes interactivos y robots personales de asistencia al usuario [HomeLab]. Por último, para localizar a los usuarios se utilizan de técnicas de visión por ordenador.

- El proyecto europeo *Amigo*, llevado a cabo por varias empresas y entidades de investigación europeas, trata de integrar la visión AmI en hogares conectados [Amigo]. Una de las principales características del proyecto es la definición de un sistema de red para el hogar cuya arquitectura posibilita la integración y composición de todo tipo de dispositivos y servicios presentes en la casa. Cabe destacar que establece pocas restricciones tecnológicas para la interacción, ya que se emplean metodologías semánticas para la interoperabilidad, integradas en la propia arquitectura del sistema.
- El sistema SCARS (*Speech Control and Audio Response System*), desarrollado por el Departamento de Electrónica de la Universidad Tecnológica de Tampere (Finlandia) [Kaila09], se basa en el denominado *Smart Home*, similar a un apartamento común, diseñado para labores de investigación en hogares inteligentes. SCARS es un sistema de control basado en interfaces de voz. Puesto que conoce la localización del usuario, el sistema puede contextualizar la interacción mediante voz. El escenario planteado para la aplicación del sistema requiere que el usuario porte un dispositivo de mano que consta de una placa base, batería recargable, un módulo de transmisión de señales de radio FM, un botón (de pulsar), y un receptor de radiación infrarroja. Como tecnología de localización se emplea la radiación infrarroja, recibida por el receptor instalado en el dispositivo portado por cada usuario.

2.1.2 Seguimiento de usuarios

El seguimiento de usuarios es una tarea básica en un sistema AmI [Cook09], ya que conociendo la localización e identidad de los usuarios, el sistema puede adecuar la funcionalidad de los servicios prestados.

El seguimiento se puede realizar mediante tecnologías de identificación basadas en dispositivos muy pequeños, tales como etiquetas RFID o dispositivos *i-Button*, que se pueden embeber prácticamente en cualquier objeto cotidiano (como ropa, joyas, etc.):

- RFID es una tecnología inalámbrica de identificación y captura de datos que emplea señales de radio [Finkenzeller03][Glover06]. Generalmente, un sistema RFID consta de tres componentes principales: etiquetas RFID, lectores RFID y un servidor *back-end*. El usuario u objeto a identificar debe portar la etiqueta RFID, que responde a las ondas de radio-frecuencia emitidas por el lector. El lector, además de detectar la presencia de etiquetas RFID, lee la información proporcionada por los mismos. El lector re-direcciona la información de identificación

proporcionada por la etiqueta al servidor *back-end*, el cual contiene una base de datos con toda la información pertinente a las etiquetas.

- Un i-Button es un circuito integrado digital, encapsulado en una pequeña carcasa de acero inoxidable de 16mm de diámetro [iButtons]. Gracias a su tamaño y robustez, el dispositivo se puede embeber prácticamente en cualquier objeto cotidiano, tales como llaveros, anillos, relojes, etc. El i-Button emplea su propia carcasa de acero inoxidable como interfaz de comunicación electrónica. Cada i-Button contiene una única dirección digital grabada en su chip, por lo que se puede emplear como herramienta de identificación. La transferencia de información de un i-Button a un dispositivo computacional (como un PC, una PDA, un micro-controlador, etc.) se lleva a cabo mediante el contacto físico del i-Button y un receptor conectado al dispositivo. Obviamente, estas metodologías requieren que el usuario porte la etiqueta o el chip de identificación. Desde un punto de vista práctico, los i-Button presentan el inconveniente de que para su lectura deben acercarse al receptor.

Estas tecnologías de identificación presentan buen rendimiento pero requieren que el usuario lleve consigo la tecnología de identificación y pase por un punto suficientemente cercano al dispositivo lector [Cook09].

En la bibliografía, para la localización y seguimiento de los usuarios se ha propuesto también el uso de sensores que capturan movimiento, como por ejemplo sensores ultrasónicos [Helal05], sensores y receptores de radiación infrarroja [Youngblood07] [Intille06] [Kaila09], sensores de presión embebidos en el suelo o en los objetos del entorno [Helal05] [Hagras04], etc. Estas tecnologías no proporcionan ninguna información sobre la identidad de la fuente que ha generado el movimiento. De hecho, para que el seguimiento se pueda realizar de forma personalizada, la mayoría de los sistemas AmI definidos en la bibliografía combinan tecnologías de localización con tecnologías de identificación:

- En el laboratorio *Smart House* [Kaila09], por ejemplo, el usuario a seguir debe portar un dispositivo de mano con un código único de identificación, el cual contiene también un receptor de señales infrarrojas emitidas por el entorno para llevar a cabo la localización.
- El sistema definido en MAVHOME emplea también radiación infrarroja para localizar a los usuarios [Youngblood07]. La posición espacial se estima a partir de sensores de radiación infrarroja acoplados a los terminales embebidos en el entorno y a los dispositivos portados por el usuario. Cuando el sistema quiere localizar a un usuario concreto, emite una señal a todas las posibles zonas donde se pueda encontrar. Todos los terminales presentes en las zonas de emisión perciben la señal, pero sólo el terminal correspondiente emitirá una señal en respuesta. Según los autores, es una tecnología eficiente de localización (con una tasa de detección del 95%), pero depende de la cantidad de zonas y habitantes a seguir, ya que su rendimiento decrece cuando aumenta el número de zonas o habitantes.

- El sistema de localización (únicamente de posición y orientación espacial) del *Gator Tech House* emplea transceptores de ultrasonidos colocados en el techo que emiten frecuencias moduladas. Los usuarios deben portar también chalecos con transceptores, los cuales escuchan las frecuencias moduladas y responden de acuerdo a la señal recibida [Helal05]. Según los autores, es una tecnología de gran precisión, pero muy intrusiva ya que el usuario debe portar un chaleco. Como alternativa, los autores proponen el empleo de sensores de presión colocados en el suelo (a costa de perder información relativa a la orientación del usuario).
- En [Ince07] se describe un sistema de asistencia médica en el hogar para personas con discapacidades cognitivas y lesiones cerebrales. El sistema debe detectar la tarea o actividad desarrollada por el paciente en todo instante. Para ello, detectan y monitorizan la localización de los habitantes a partir de acelerómetros inalámbricos embebidos en el entorno y en una muñequera portada por cada paciente. Los datos extraídos por ambas tecnologías sensoriales se parametrizan en tiempo y en frecuencia, y se modelan mediante mezclas de gaussianas.

Otra tecnología aplicable para el seguimiento de usuarios en un entorno AmI es el reconocimiento de locutores [Zamalloa10a] [Zamalloa10b], campo que engloba un conjunto de técnicas y metodologías para determinar la identidad de hablante en base a las propiedades acústicas o lingüísticas observadas en una señal de voz. Contando con uno o varios locutores objetivo y una señal de entrada, se pueden aplicar técnicas del campo de la verificación o identificación de locutores, para determinar en qué momento habla un locutor hipotético (tarea de verificación) o alguno de los locutores objetivo (tarea de identificación). Si la señal a analizar se adquiere mediante un micrófono embebido en el entorno AmI, el seguimiento de locutores se restringe al entorno abarcado por el micrófono. Asimismo, si el entorno AmI se delimita por zonas, con uno o varios micrófonos por cada zona, el mismo procedimiento de seguimiento de locutores se puede aplicar para determinar la posición o localización relativa de los locutores. Otro posible escenario de aplicación del reconocimiento de locutores es el seguimiento de un usuario ya localizado mediante cualquier otra tecnología de localización no intrusiva. En [Potamitis04], por ejemplo, se emplea un array de micrófonos y la metodología TDOA (*Time Delay of Arrivals*) para estimar la localización espacial de una fuente acústica (véase también el sistema multimodal [Busso05] que se describe más abajo).

El reconocimiento de locutores es una metodología muy poco intrusiva y completamente transparente, ya que como tecnología sensorial emplea únicamente micrófonos embebidos en el entorno. Es una tecnología apropiada para entornos privados, no muy ruidosos, en los que de antemano se conocen las características de los usuarios a reconocer. A diferencia de las tecnologías sensoriales de movimiento, o las tecnologías de identificación RFID e i-Button, libera al usuario de tener que llevar algún dispositivo de captación o identificación, debido a que el seguimiento se realiza en base a la propia voz de los usuarios, una característica intrínseca. Obviamente, presenta el inconveniente de que el usuario tiene que hablar para que el seguimiento se pueda llevar a cabo.

No obstante, se ha de tener en cuenta que el habla es el medio de comunicación más natural del ser humano, por lo que permite una interacción no intrusiva entre personas y máquinas, siempre que las máquinas del entorno contengan módulos de reconocimiento y síntesis del habla. Tal como se argumenta en [Abowd05], tan importante como el avance de las tecnologías, es que su aplicación sea sencilla y natural (es decir, lo más semejante posible a las interacciones humanas). En esta línea, una de las características más novedosas del sistema SCARS citado anteriormente (véase [Kaila09]), es que la interacción entre el usuario (localizado mediante sensores infrarrojos) y el sistema se realiza por voz. Para activar la interacción por voz, el usuario debe pulsar el botón de un dispositivo de mano, funcionalidad algo intrusiva y que debería ser reemplazada con alguna tecnología de monitorización continua, como por ejemplo, el seguimiento de locutores continuo propuesto en este trabajo [Zamalloa10], o simplemente un detector de voz sobre las señales adquiridas mediante los micrófonos embebidos en el entorno.

Por último, cabe destacar que existen aproximaciones de seguimiento alternativas, que, al igual que el reconocimiento de locutores, resultan muy cómodas para el usuario. Entre estas aproximaciones destacan el seguimiento de usuarios mediante sensores de presión embebidos en el suelo [Orr00] [Helal05] [Valtonen09] y el seguimiento de usuarios a través de cámaras embebidas en el entorno [Krum00][PlaceLab][HomeLab], en cuyas imágenes se trata de reconocer rostros, patrones de movimiento, gestos o insignias identificativas portadas por los usuarios. Dentro de esta última línea de investigación, y posiblemente influenciados por el interés que ha suscitado la transcripción de reuniones como escenario de aplicación en el campo de las tecnologías del habla, se han desarrollado sistemas multimodales de video y voz para identificar o realizar el seguimiento de los participantes de una reunión [Luque07] [Salah08] [Busso05] [Gatica-Perez07][Bernardin07]. Estos sistemas se evalúan mediante transcripciones detalladas de grabaciones realizadas en salas de reuniones equipadas con cámaras de video y micrófonos:

- En [Luque07] los autores definen un sistema multimodal para determinar la identidad de los participantes en grabaciones realizadas en la sala de reuniones inteligente de la UPC (Universidad Politécnica de Catalunya) que consta de seis cámaras colocadas en puntos estratégicos de la sala, un array de micrófonos colocado cerca de la pared principal de la sala, tres clusters de micrófonos de 4 canales en forma de T colocados en las otras tres paredes de la sala y 8 micrófonos de mesa. El sistema multimodal consiste en la fusión de dos sistemas: un sistema de reconocimiento de locutores que opera sobre los segmentos de audio; y un sistema de reconocimiento de rostros que opera sobre los segmentos de vídeo. En los experimentos realizados por los autores, el sistema de reconocimiento de locutores presenta mejor rendimiento que el sistema de reconocimiento de rostros.
- En [Salah08] se define un sistema multimodal de identificación y localización de los participantes en una reunión. Al igual que en el trabajo anterior, las grabaciones se realizan en la sala de reuniones inteligente de la UPC. Los módulos combinados son los siguientes: (1) un módulo de

reconocimiento de rostros; (2) un módulo de seguimiento de usuarios que opera sobre las señales de vídeo; (3) un módulo de localización espacial del usuario en base a las fuentes acústicas detectadas en la sala (en caso de que haya múltiples fuentes, el sistema detecta la fuente dominante); y (4) un módulo de identificación de locutores. Este último módulo realiza la identificación sobre la suma de las señales acústicas adquiridas por los diferentes micrófonos de la sala, e incluye sistemas de detección de voz/no-voz y de segmentación acústica. En los experimentos realizados por los autores, el módulo de reconocimiento de rostros presenta mejor rendimiento que el basado en los segmentos de voz.

- En [Busso05] se presenta un sistema multimodal que trata de determinar la localización espacial y la identidad de los participantes en una reunión grabada en una sala inteligente equipada con un array de 16 micrófonos omnidireccionales, cuatro cámaras de vídeo 3D colocadas en las esquinas de la sala y una cámara omnidireccional de ángulo circular (de 360º) colocada en el centro de la mesa de reuniones. El sistema realiza la fusión en tiempo real de varios sistemas de localización e identificación: un sistema de localización acústica que trata de determinar la ubicación de la fuente dominante a partir de las señales grabadas por los micrófonos del array; un sistema de identificación de locutores que, a partir de la señal creada con la suma de las señales adquiridas por los diferentes micrófonos, emite una probabilidad para cada participante en la reunión; un sistema de localización que, a partir de las señales de vídeo (sincronizadas) adquiridas por las diversas cámaras, trata de determinar la localización espacial de cada participante; y un sistema de reconocimiento de rostros que, en base a las imágenes adquiridas por la cámara de vídeo omnidireccional, trata de determinar la identidad del hablante.
- En [Gatica-Perez07] se propone una aproximación probabilística para realizar simultáneamente la localización y la detección de actividad de voz de cada uno de los participantes en una reunión grabada en una sala equipada con diversas cámaras de vídeo y un array de micrófonos. A partir de las señales adquiridas por el array de micrófonos, el sistema determina la localización espacial 2D del participante que está hablando. A continuación, partiendo de la información de localización, el sistema aplica un módulo de reconocimiento de rostros que analiza, entre otras cosas, las manchas de piel y la forma de la cabeza.
- En [Bernardin07] se propone un sistema multimodal de identificación. El sistema consta de varios módulos, de los cuales el principal es un módulo de seguimiento que opera sobre las imágenes captadas por las 4 cámaras fijas colocadas en las esquinas de la sala y una cámara circular ubicada en el centro. De forma paralela, se realiza la localización e identificación del locutor mediante grupos de micrófonos en forma de T instalados en las paredes de la sala. La localización del locutor se realiza mediante todos los micrófonos, empleando la metodología TDOA. La identificación de locutores, que incluye los módulos de detección de voz y segmentación, se aplica únicamente sobre los segmentos de duración superior a un segundo. Partiendo de la información

obtenida de los sistemas de seguimiento, localización e identificación, aplica un sistema de reconocimiento de rostros que trata de localizar a los participantes activos en cada instante.

2.2 Identificación y Verificación de locutores

En campo del procesamiento del habla, la tarea denominada reconocimiento de locutores, es la tarea consistente determinar o verificar, por medio de una máquina, la identidad del locutor a partir de una señal de voz. En este sentido, un sistema de reconocimiento de locutores es aquél que determina o verifica la identidad del hablante. Son sistemas catalogados como biométricos ya que abordan la detección de características fisiológicas y conductuales que permiten identificar a las personas.

En la bibliografía pueden encontrarse numerosos trabajos que revisan las técnicas y resultados más importantes obtenidos en los últimos 15 años en el área del reconocimiento del locutor [Rosenberg08] [Faundez-Zanuy05] [JPCampbell97]. Para identificar locutores es preciso obtener un conocimiento previo sobre los mismos, que se obtiene a través del proceso denominado entrenamiento y consiste en la recopilación de sus señales de voz. El desarrollo de un sistema de reconocimiento se completa con la evaluación, el proceso de comparar las decisiones tomadas por el sistema con las decisiones correctas y establecer una medida del rendimiento.

Los sistemas de reconocimiento de locutor, dependiendo del tipo de decisión a tomar, se dividen en: (1) *sistemas de identificación*, que determinan la identidad del locutor; y (2) *sistemas de verificación*, que determinan si el locutor es quien dice ser. Si el locutor que aparece en la señal de evaluación (conocido como locutor objetivo o *target*) pertenece a un grupo predefinido de locutores, se dice que la identificación se realiza en un conjunto cerrado y recibe el nombre de *closed-set speaker identification* [Atal76] [Doddington85]. Es decir, el sistema debe identificar a qué locutor de un conjunto de N locutores conocidos pertenece la observación de entrada. Por el contrario, si el sistema asume que la señal de entrada puede pertenecer a un locutor desconocido, un impostor, se dice que la identificación se realiza en un conjunto abierto y recibe el nombre de *open-set speaker identification*.

En las Figuras 2.1 y 2.2 se muestran las fases de entrenamiento y evaluación de un sistema de identificación. La **Figura 2.1** resume la fase de entrenamiento que parte de la señal muestreada de un determinado locutor. En primer lugar, se extrae un conjunto de vectores de parámetros acústicos que generalmente consiste en realizar el análisis espectral de pequeños tramos de señal consecutivos. A partir de esta representación, se obtiene un modelo estadístico que define un patrón específico del locutor. Este proceso se repite para todos los locutores que se pretende identificar o verificar. Por tanto, suponiendo que se desea construir un sistema de reconocimiento para N locutores, la fase de entrenamiento consiste en estimar N modelos de locutor.

Figura 2.1. *Fase de entrenamiento de un sistema de reconocimiento de locutores*

El proceso de evaluación (**Figura 2.2**) parte de una señal de voz muestreada, de la que de nuevo, se extrae una secuencia de vectores de parámetros. Los parámetros se emparejan con cada modelo de locutor i ($i \in \{1, ..., N\}$), de donde se obtienen N puntuaciones, y habitualmente se escoge el modelo del locutor m que proporciona la máxima puntuación s_m. Si la identificación se realiza en un conjunto cerrado se determina que el locutor objetivo es el locutor m. Si se realiza en un conjunto abierto y s_m supera un umbral predeterminado se decide que m es el locutor objetivo; en caso contrario, se asume que la señal proviene de un locutor desconocido.

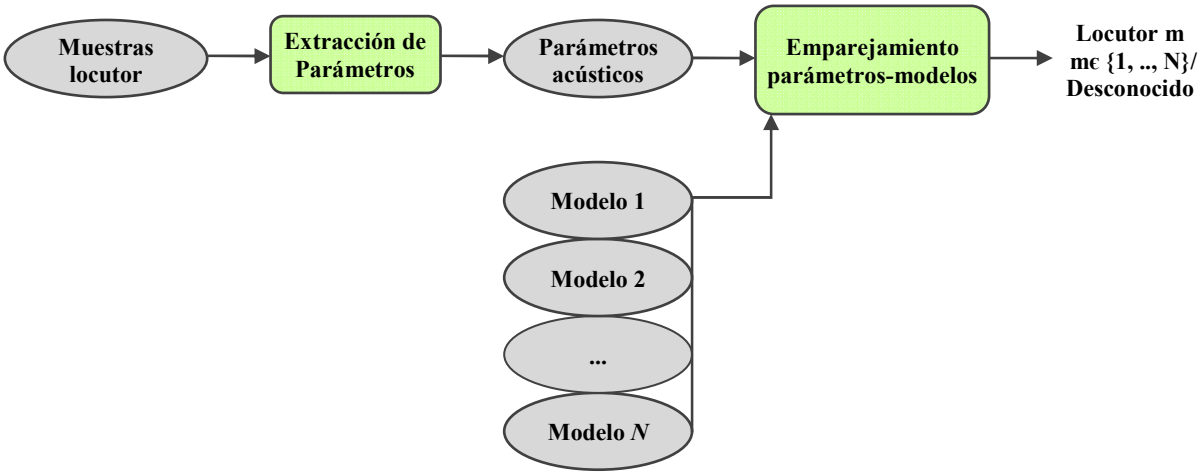

Figura 2.2. *Fase de evaluación de un sistema de identificación de locutores*

En un sistema de verificación, en cambio, la evaluación se realiza sobre un par formado por la observación de entrada y una identidad [Rosenberg76][Bimbot04]. En este caso, la señal de entrada sólo se evalúa para la identidad suministrada. Si la verificación es satisfactoria, se acepta la identidad; en caso contrario, se rechaza. La verificación es un caso especial de la identificación en un conjunto abierto para N=1. En la **Figura 2.3** se muestra la fase de evaluación de un sistema de verificación.

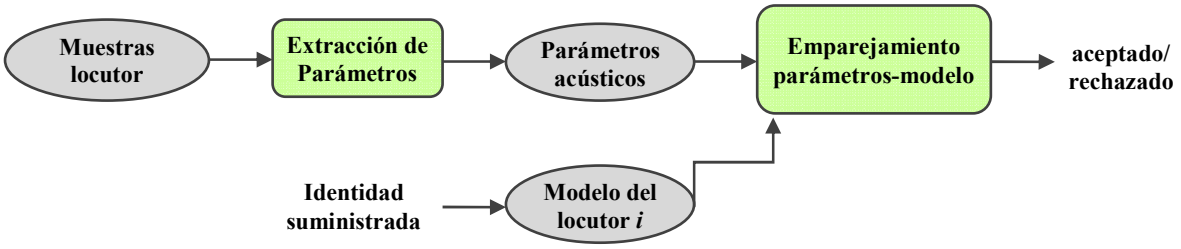

Figura 2.3. *Fase de evaluación de un sistema de verificación de locutores*

Los sistemas de identificación y verificación se dividen en otras dos categorías [Doddington85]: (1) sistemas independientes del texto – *text-independent* –; y (2) sistemas que dependen del texto de la observación– *text-dependent*. Los sistemas independientes del texto constituyen el caso más general [Reynolds08] [Gish94]; no pretenden que el locutor pronuncie un texto determinado, ni al entrenarlo, ni para reconocerlo. Esta modalidad se aplica en casos en los que el locutor no participa voluntariamente en una operación de identificación (por ejemplo, en aplicaciones forenses), o simplemente cuando el reconocimiento se emplea como servicio adicional (por ejemplo, para la adaptación y personalización de servicios). La independencia del texto añade flexibilidad pero también dificultad a la tarea de reconocimiento. La modalidad dependiente del texto, en cambio, requiere que el locutor pronuncie un texto explícito que bien conoce de antemano o le ha indicado el sistema en ese instante [Hébert08] [Higgings91]. Se dispone de modelos acústicos de cada locutor, por lo que la señal de entrada se empareja con un modelo específico para ese locutor. Por ello, el rendimiento de clasificación es generalmente superior al de los sistemas independientes del texto. Los sistemas dependientes del texto se emplean en aplicaciones de identificación muy exigentes (no sólo por la mejora de su fiabilidad, sino por su robustez ante intentos de fraude mediante grabaciones) o en sistemas de diálogo.

Los sistemas de reconocimiento de locutores se emplean principalmente en aplicaciones de seguridad, monitorización y forenses [Rosenberg08]. En los últimos años, debido a la expansión de Internet, se está manejando un gran volumen de recursos multimedia, por lo que surge la necesidad de desarrollar herramientas de búsqueda eficientes y versátiles [Tranter06] [Bordel09]. Estos recursos contienen habla de muchísimos locutores, que podría ser interesante indexar, donde surge otra aplicación interesante de los sistemas de identificación de locutores [Wilcox94] [Siegler97]. Por último, los sistemas de reconocimiento de locutores se emplean también para mejorar el rendimiento de los reconocedores del habla, mediante la adaptación al locutor de los modelos acústicos [Solomonoff98] [Rodriguez-Fuentes04] [Liu05].

2.3 Parametrización Acústica

La parametrización acústica, también denominada extracción de parámetros, consiste en calcular una secuencia de vectores de parámetros a partir de una señal de voz muestreada. El objetivo principal de este procedimiento es la obtención de una nueva representación más compacta, menos redundante y más adecuada a los modelos estadísticos (sin correlaciones). Generalmente, el habla se muestrea a 8kHz o 16kHz, empleando 8 o 16 bits por muestra, lo que produce un conjunto de datos de gran volumen. No obstante, las características esenciales del habla varían lentamente en el tiempo, por lo que su representación digital se puede reducir de forma significativa. La extracción de parámetros trata de producir un conjunto de vectores que disminuya la dimensión del conjunto de datos de entrada y, al mismo tiempo, mantenga toda la información relevante. Por otro lado, en general, se supone que los vectores de un locutor ocupan una región determinada en el espacio acústico. Aunque estas regiones se

solapan, se pueden distinguir y, de hecho, clasificar. Con el objetivo de simplificar la clasificación, habitualmente se asume que los vectores son independientes, aunque se sabe que, en realidad, hay correlación entre vectores consecutivos.

La información relativa al locutor se puede extraer a muchos niveles, pero se distinguen dos grandes tipos de características o parámetros [JPCampbell03]:

(1) Parámetros de bajo nivel, que hacen referencia fundamentalmente a la acústica (los sonidos y sus secuencias)

(2) Parámetros de alto nivel, que hacen referencia a la morfología, la sintaxis y la pragmática, es decir, el uso que el locutor hace de la lengua: su forma de construir palabras y frases.

En los dos siguientes apartados se describen ambos tipos de características con algo más de detalle.

2.3.1 Parámetros de bajo nivel

Los parámetros de bajo nivel están relacionados con la zona periférica de percepción del habla en el cerebro, representan características fisiológicas del locutor (la forma del tracto vocal, la mandíbula, etc), es decir, la geometría del sistema de producción de habla humano. Se reflejan en el espectro de la señal, en la ubicación y ancho de banda de las formantes o la frecuencia fundamental (*pitch*). La extracción de estos parámetros es relativamente sencilla y robusta. Su principal inconveniente es que pueden ser fácilmente corrompidos por el ruido ambiental y otras fuentes de distorsión.

En la práctica, en vez de extraer explícitamente los formantes de la señal, se miden las amplitudes espectrales, que también capturan implícitamente las características fisiológicas del locutor, y además presentan la ventaja de que su estimación es muy sencilla [Rabiner93]. Por tanto, son los parámetros espectrales los más empleados. Aunque en la bibliografía se pueden encontrar varias formas de representar parámetros espectrales [Reynolds94], los coeficientes cepstrales en la escala de Mel (*Mel Frecuency Cepstral Coefficients*, MFCC), que en un principio se desarrollaron para el reconocimiento del habla independiente del locutor [Davis80], constituyen la representación más usada. En los parámetros MFCC, la resolución espectral disminuye drásticamente para frecuencias superiores a 2KHz, lo que parece perjudicial para el reconocimiento de locutores. Sin embargo, en la práctica, ha resultado ser una de las mejores representaciones. Los parámetros espectrales capturan las diferencias entre sonidos, pero también las diferencias entre patrones de pronunciación, por lo que se emplean tanto en reconocimiento del locutor como en reconocimiento del habla, sobrepasando a veces el rendimiento de la percepción auditiva humana. Además de los MFCC, la representación basada en coeficientes cepstrales estimados mediante técnicas de predicción lineal [Fant70] y la percepción lineal perceptual (*Perceptual Linear Prediction*, PLP) [Hermansky90] son también parametrizaciones muy comunes.

La extracción de los parámetros MFCC comienza con la aplicación de una ventana de análisis, a partir de la cual se estima un vector espectral. La ventana se especifica mediante dos longitudes; el tamaño de la ventana, y su desplazamiento en el tiempo. Generalmente, se utilizan ventanas de 20 o 30

milisegundos y el desplazamiento se fija a unos 10 milisegundos, de forma que dos ventanas consecutivas se solapen entre 10 y 20 milisegundos. Una vez fijados estos dos valores, hay que elegir el tipo de ventana que atenue los valores de la señal en los extremos para aliviar el efecto negativo que tiene el uso de una señal finita recortada por un escalón en el procesamiento posterior: la ventana Hamming es la más empleada en reconocimiento de locutores.

A continuación, se calcula la Transformada Rápida de Fourier (*Fast Fourier Transform*, FFT) de la señal ventaneada. Para ello, hay que determinar la cantidad de puntos N de la FFT, una potencia de 2 mayor que la cantidad de muestras de la ventana. Por tanto, N depende de la frecuencia de muestreo f (generalmente, $N = 256$ con $f = 8$ kHz; y $N = 512$ con $f = 16$ kHz). Al espectro de la señal se le aplica un banco de filtros para obtener la envolvente del espectro, lo cual reduce la dimensionalidad de los vectores espectrales y elimina ciertas fluctuaciones. Un banco de filtros consiste en una serie de filtros pasa-banda solapados que abarcan todo el rango de frecuencias. Los filtros pueden ser triangulares o de cualquier otro tipo. En la mayoría de los trabajos de la bibliografía, se emplea la escala de Mel o Bark para definir la distribución frecuencial de los filtros. Se trata de una escala que simula la percepción auditiva humana con mayor resolución en frecuencias bajas que en las altas. Por último, se calcula el logaritmo de los coeficientes del banco de filtros y se realiza una Transformada Discreta de Cosenos (*Discrete Cosine Transform*, DCT), que produce los MFCC.

Los coeficientes cepstrales son muy sensibles a las variaciones del canal y del transductor empleado en la grabación. Una de las técnicas más empleadas para compensar las distorsiones del canal es la normalización de la media cepstral (*Cepstral Mean Subtraction*, CMS) [Rosenberg94], un método simple y efectivo. Consiste en transformar los coeficientes cepstrales para que su media en el tiempo sea cero. Asume que la media de una señal grabada en condiciones de laboratorio es cero, y debido a que la media cepstral de una señal suficientemente larga se aproxima a la media de un canal invariante en el tiempo, al restar la media cepstral se supone que se compensa la distorsión del canal. En la bibliografía se pueden encontrar otras técnicas de normalización, entre las que cabe destacar RASTA [Hermansky94], y de forma especial, *Feature Warping* (FW) [Pelecanos01] que mapea la distribución observada a una distribución normal mediante una ventana deslizante, y ha mostrado un buen rendimiento en las últimas competiciones del NIST.

Asimismo, diversos trabajos han demostrado que la información dinámica mejora de forma significativa la eficiencia de los sistemas de reconocimiento del habla, por lo que además de los MFCC, generalmente se utilizan también aproximaciones de sus primeras y segundas derivadas [Furui81]. Así, se añade información sobre la evolución en el tiempo de los MFCC. La energía, calculada como el coeficiente cero de los cepstrales y establecida como la media del logaritmo de la energía de la señal, también suele incluirse en el vector de parámetros, junto con su primera y segunda derivada.

2.3.2 Parámetros de alto nivel

La mayoría de los sistemas de reconocimiento de locutor, y de forma especial los primeros sistemas desarrollados, emplean parámetros espectrales. Hasta el momento, estos parámetros han presentado el mejor rendimiento. Sin embargo, la señal de voz transporta otra información que permite distinguir a un locutor de otro, como la prosodia, los patrones de pronunciación, el uso de la lengua, etc. [Reynolds03] [Mary06] [Ferrer07].

La prosodia representa el acento, el tono y la entonación de una expresión oral [Sonmez98]. Para su extracción se analiza la variación de la frecuencia fundamental, la energía y la duración de fonemas y silencios. Para el reconocimiento de patrones de pronunciación, en cambio, se analizan secuencias de fonemas que supuestamente reflejan la pronunciación y elección de las palabras más frecuentes mediante bigramas y trigramas [Boakye04]. Generalmente, cada locutor pronuncia cada fonema de una manera particular. Esta variabilidad posibilita la comparación de frecuencias y grupos de fonemas. Así, se capturan las características idiolectales del locutor que, entre otras, pueden reflejar sus rasgos culturales y geográficos. Esto requiere entrenar y aplicar modelos acústicos de fonemas [Gauvain95] para obtener la decodificación acústico-fonética de la señal de entrada. Por último, el uso de la lengua es también un rasgo característico del locutor [Doddington01]. En este caso, las secuencias se forman con palabras obtenidas mediante un reconocedor [Newman96]. A través de la probabilidad de secuencias de N palabras consecutivas se representa el uso de la lengua, es decir, las expresiones más utilizadas por el locutor.

Los parámetros de alto nivel son robustos frente al ruido pero requieren una gran cantidad de datos para estimar los modelos acústicos y de lenguaje. Su extracción es también más compleja, dado que la mayoría de parámetros requieren el empleo de un reconocedor del habla para obtener la secuencia de palabras o fonemas. Además, son muy sensibles a los errores del reconocedor y los modelos tienden a caracterizar el contexto de la conversación junto con los rasgos del locutor. Esto supone una complejidad excesiva para los escenarios AmI que se plantean en esta tesis. Por lo tanto, se utilizarán parámetros acústicos de bajo nivel, ya que su extracción es muy sencilla, no requieren un reconocedor y basta con una relativamente pequeña cantidad de datos para estimar buenos modelos.

2.4 Reconocimiento de locutores

El reconocimiento de patrones, campo en el que se ubica el reconocimiento de locutores, consiste en la estimación de los modelos que definen los patrones a reconocer (en este caso los locutores), y la clasificación de las observaciones mediante la aplicación de estos modelos [Duda00].

La estimación de modelos y la clasificación se realiza sobre las señales parametrizadas. Se emplean modelos probabilísticos, que representan una distribución probabilística de las observaciones y permiten calcular la verosimilitud, o probabilidad condicional, de una observación dado el modelo. En el campo del reconocimiento de locutores, los primeros sistemas desarrollados empleaban modelos

deterministas basadas en plantillas (*Template Matching*) [Higgins93] [Soong85]. Los modelos deterministas asumen que la observación es una réplica del modelo, y tratan de minimizar la distancia entre los vectores de la observación y los vectores del modelo. Aunque en ciertas circunstancias todavía se emplean modelos deterministas, los modelos probabilísticos presentan un mejor rendimiento. Los modelos deterministas son muy sensibles a las variaciones del canal y al ruido del entorno.

2.4.1 La clasificación de locutores

El punto central del reconocimiento de locutores es la clasificación de una señal en función de sus características [Duda00]. Al emplear modelos de locutor probabilísticos, el espacio acústico de un locutor se describe mediante una distribución de probabilidad, y las decisiones se determinan a partir de probabilidades o funciones de verosimilitud. Supóngase que $p(\mathbf{x}|l)$ es la función de densidad de probabilidad condicional de la observación **x** (un vector acústico con la representación espectral correspondiente a un tramo de señal) dado el locutor *l,* el cual pertenece a un conjunto de *L* locutores conocidos. La densidad de probabilidad conjunta del locutor *l* y la observación de entrada **x**, es la siguiente:

$$p(l,\mathbf{x}) = p(l|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|l)P(l) \tag{2.1}$$

donde, siguiendo la notación estándar, *p* es la función de densidad de probabilidad y *P* la función de probabilidad. A partir de la fórmula anterior obtenemos la expresión comúnmente conocida como fórmula de Bayes, que determina la probabilidad a posteriori de que, dada la observación **x**, el locutor sea *l*:

$$p(l|\mathbf{x}) = \frac{p(\mathbf{x}|l)P(l)}{p(\mathbf{x})} \tag{2.2}$$

donde *P(l)* es la probabilidad a priori de que el locutor sea *l* y *p(***x***)* la probabilidad de la observación **x**. Si el conjunto de locutores es cerrado, *p(***x***)* es la suma ponderada por locutores de las densidades de probabilidad de todos los locutores:

$$p(\mathbf{x}) = \sum_{l=1}^{L} p(\mathbf{x}|l)P(l) \tag{2.3}$$

Además, si, como en la mayoría de casos, todos los locutores son equi-probables ($P(l) = 1/L$), el locutor más probable como emisor de la observación **x** a posteriori ($\hat{l}$) es el locutor de máxima probabilidad condicional. Si suponemos una secuencia de observaciones independientes $X=\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, $\hat{l}$, calculado en espacio logarítmico, es:

$$\hat{l}(X) = \arg\max_{l=1,..,L}\{\ln p(X|l)\} = \arg\max_{l=1,..,L}\left\{\ln\left(\sqrt[N]{\prod_{n=1}^{N} p(\mathbf{x}_n|l)}\right)\right\} =$$

$$= \arg\max_{l=1,..,L}\left\{\frac{1}{N}\sum_{n=1}^{N}\ln p(\mathbf{x}_n|l)\right\} \tag{2.4}$$

Pero, si el conjunto de locutores del sistema de identificación es abierto, para tomar una decisión, $\hat{l}(X)$ se compara con un umbral predeterminado θ_I. En consecuencia, la probabilidad de la observación $p(\mathbf{x})$ (ecuación 2.3) puede desempeñar un papel importante como factor de normalización: $p(\mathbf{x})$ hace que la tasa obtenida sea una verdadera probabilidad (suponiendo que el espacio de posibles locutores es el del conjunto de entrenamiento), y además, permite comparar la verosimilitud de cada locutor con la del conjunto de locutores. En resumen, dada una observación *X*, un sistema de identificación sobre un conjunto de locutores abierto, toma su decisión de la siguiente manera:

$$\text{decisión} = \begin{cases} \hat{l} & \ln p(X|\hat{l}) - \ln \mathrm{p}(X) \geq \theta_\mathrm{I} \\ \text{impostor} & \text{en caso contrario} \end{cases} \tag{2.5}$$

donde la log-probabilidad de la secuencia de observaciones es la siguiente ($p(\mathbf{x}_n)$ se estima mediante la ecuación 2.3):

$$\ln p(X) = \frac{1}{N}\sum_{n=1}^{N} \ln p(\mathbf{x}_n) \tag{2.6}$$

Los sistemas de verificación, en cambio, parten una observación *X* y un locutor hipotético *l*. Es una clasificación binaria; el sistema debe determinar si es o no *l* quien ha pronunciado *X*. Generalmente, se asume que *X* contiene el habla de un único locutor. La tarea de verificación es un contraste (test) de hipótesis [Fukunaga90]:

(1) H_0: *X* es del locutor *l*

(2) H_1: *X* no es del locutor *l*

La hipótesis H_0 se representa mediante la probabilidad a posteriori $p(H_0|X) = p(l|X)$. La hipótesis alternativa H_1 se representa mediante $p(H_1|X) = p(\bar{l}|X)$. Para tomar la decisión, se estima el ratio de probabilidades a posteriori $\Lambda(X, l)$ de estas dos hipótesis, por conveniencia, en el espacio logarítmico:

$$\Lambda(X, l) = \ln \frac{\mathrm{p}(l|X)}{\mathrm{p}(\bar{l}|X)} = \ln \frac{\mathrm{p}(X|l)\mathrm{P}(l)}{\mathrm{p}(X|\bar{l})\mathrm{P}(\bar{l})} =$$

$$= \left(\frac{1}{N}\sum_{n=1}^{N} \ln \mathrm{p}(\mathbf{x}_n|l) + \ln \mathrm{P}(l)\right) - \left(\frac{1}{N}\sum_{n=1}^{N} \ln \mathrm{p}(\mathbf{x}_n|\bar{l}) + \ln \mathrm{P}(\bar{l})\right) \begin{cases} \geq \theta_V, & se\ acepta\ H_0 \\ < \theta_V, & se\ acepta\ H_1 \end{cases} \tag{2.7}$$

donde θ_V es el umbral de aceptación del sistema de verificación. El ratio calculado en el espacio logarítmico se denomina comúnmente *log-likelihood ratio* (LLR).

La hipótesis H_0 está bien definida y se modela con los datos de entrenamiento del locutor *l*. No obstante, la hipótesis alternativa debe representar el espacio de posibles alternativas al locutor *l*, por lo que su definición no es tan obvia. La bibliografía cuenta con dos aproximaciones principales: (1) utilizar una cohorte de locutores; y (2) utilizar un modelo universal.

La primera aproximación consiste en emplear un conjunto de locutores externos para dar cobertura al espacio de la hipótesis alternativa [Higgins91] [Rosenberg92]. Dado un conjunto de *G* locutores $B=\{b_1, b_2, \ldots, b_G\}$, que generalmente no incluye el locutor objetivo *l*, la hipótesis alternativa se representa como:

$$p(X|\bar{l}) = f(p(X|b_1), \ldots, p(X|b_G)) \tag{2.8}$$

donde f es una función (por ejemplo, de maximización o promedio) de las probabilidades condicionales de X dado B. La mayoría de los trabajos consultados en la bibliografía señalan que se ha obtenido un rendimiento mejor al emplear un conjunto de locutores específico B_l para cada locutor l con locutores muy similares a l.

La segunda aproximación, en cambio, consiste en estimar un único modelo, denominado modelo general [Carey91] o modelo universal (UBM) [Reynolds97], a partir de las muestras de un amplio conjunto de locutores que representan el espacio global de locutores. Se asume que el UBM da cobertura a todo el espacio acústico, por lo que se debe estimar a partir de un conjunto de locutores variado y equilibrado en cuanto a fonética, canal de grabación y transmisión, características de locutores (genero, edad, etc), etc. Aunque se puede definir un UBM para cada locutor, o tipo de locutor, el empleo de un único modelo es la aproximación dominante.

2.4.2 Modelos de locutor

El reconocimiento de patrones trata de definir modelos (patrones) a partir de las muestras de entrenamiento. Cada modelo describe, de forma matemática, las características de una clase en particular. Habitualmente, los modelos se estiman mediante un algoritmo que minimiza el coste (el error) de clasificación de las muestras de entrenamiento [Duda00]. Esta es la aproximación más empleada, y también la que se ha seguido en los sistemas desarrollados en este trabajo. Se denomina aprendizaje supervisado debido a que el sistema cuenta con un conjunto de muestras de entrenamiento y conoce la categoría de cada muestra. Por el contrario, si el sistema no conoce las categorías de las muestras y agrupa "a ciegas" las observaciones en un conjunto indefinido de clases, se dice que el aprendizaje es no-supervisado. En este último caso, el proceso de aprendizaje trata de definir el mejor algoritmo de agrupación (o *clustering*) mediante el cual se definen las categorías.

En función de la estrategia de modelado, se distinguen [Fukunaga90]:

- Modelos no-paramétricos frente a modelos parametricos
- Modelos generativos frente a modelos discriminativos

Los modelos no-paramétricos no realizan, prácticamente, ninguna suposición sobre la estructura de la clase y presentan un buen rendimiento cuando los modelos se estiman con muchos datos que se ajustan a los datos de evaluación. Los modelos paramétricos, en cambio, parten de una estructura predefinida y caracterizada mediante un conjunto de parámetros. Esta aproximación presenta ciertas ventajas: se pueden estimar modelos más robustos con menos datos, realizar inferencias estadísticas y modelar los cambios de datos mediante cambios en los parámetros. De todas formas, puede resultar perjudicial que la estructura sea demasiado restrictiva. Todas las técnicas analizadas en esta tesis son paramétricas, con un grado considerable de flexibilidad.

Los modelos discriminativos, como por ejemplo las máquinas de vectores soporte (SVM), tratan de minimizar el número de errores de clasificación de las muestras de entrenamiento. Determinan los límites entre clases (es decir, las fronteras) e ignoran las pequeñas fluctuaciones. Para ello, el

entrenamiento se realiza con muestras positivas (correspondientes a la clase objetivo) y muestras negativas (que representan todo el espacio alternativo). Estos clasificadores modelan la distribución a posteriori, es decir tratan de estimar directamente la probabilidad a posteriori $p(l|\mathbf{x})$. Por otro lado, los modelos generativos, como por ejemplo las mezclas de Gaussianas (GMM), estiman la densidad de probabilidad de cada clase de forma independiente: el modelo se estima únicamente con muestras positivas. Estos modelos tratan de capturar todas las fluctuaciones y variaciones de los datos. Modelan la probabilidad condicional de **x** dada la clase l, $p(\mathbf{x}|l)$, y los parámetros del modelo se estiman por máxima verosimilitud para las muestras de entrenamiento de la clase. Una vez estimado el modelo, para tomar una decisión, se aplica la regla de Bayes y se selecciona el modelo de máxima probabilidad a posteriori.

La elección de la metodología de modelado, así como el tipo de modelo, dependen de la parametrización de las muestras y las limitaciones de la aplicación. En aplicaciones de identificación de locutores dependiente del texto, es preferible que los modelos, típicamente modelos ocultos de Markov (*Hidden Markov Models*, HMM) representen los fonemas, palabras o frases a detectar para de esta forma capturar las características temporales. Por el contrario, en aplicaciones de identificación de locutores independiente del texto, la observación de test puede contener fonemas o palabras que no han aparecido en entrenamiento, por lo que en estos casos resulta más efectivo que los modelos representen el espacio acústico del locutor. Debido a que las aplicaciones de un entorno AmI son independientes del texto, en este trabajo se ha seguido esta última aproximación: los modelos de locutor representan el espacio acústico de cada locutor, y por tanto, se estiman a partir de parámetros acústicos de bajo nivel.

A continuación, se describen las metodologías de modelado utilizadas en este trabajo, que incluyen algunas de las metodologías de referencia de los últimos años (y que son también más eficientes):

(1) Distribuciones de etiquetas de cuantificación vectorial
(2) GMM estimados por máxima verosimilitud mediante el algoritmo EM
(3) GMM estimados por adaptación bayesiana del UBM
(4) SVM con un kernel secuencial lineal discriminante generalizado
(5) SVM con un kernel secuencial de vectores soporte GMM

Cabe también destacar la metodología de modelado denominada VQ-UBM propuesta recientemente [Hautamäki08], basada en la cuantificación vectorial, y muy similar a los modelos GMM estimados por adaptación bayesiana del UBM. En [Kinnunen09] se presenta una comparativa del rendimiento de esta metodología, con respecto a los modelos GMM adaptados por adaptación bayesiana del UBM y los modelos SVM con un kernel lineal discriminante generalizado.

2.4.2.1 Distribuciones de etiquetas de cuantificación vectorial (e-VQ)

La cuantificación vectorial (*Vector Quantization*, VQ) es una aproximación estadística que trata de representar los datos de entrada mediante un conjunto de vectores de referencia [Linde80]. La cuantificación se define como una función q, que asigna un vector de referencia $\mathbf{c}=q(\mathbf{x}) =(c_1, .., c_D)$,

denominado *codeword* o centroide, a cada vector del conjunto original $\mathbf{x}=(x_1, ..., x_D)$. El conjunto de centroides es finito $C=\{\mathbf{c}_i;\ i=1, ..., N_c\}$ y se denomina *codebook*. La cuantificación define una partición $S=\{S_i;\ i=1,..., N_c\}$ del conjunto de entrada, donde $S_i=\{\mathbf{x}:\ q(\mathbf{x})=\mathbf{c}_i\}$. Para ello, trata de encontrar el mapeo óptimo, $\hat{q}$, que minimiza una función objetivo (generalmente, la distorsión media) que representa el error de cuantificación. Para determinar la distorsión se pueden emplear diferentes métricas, entre las cuales predomina la distancia euclídea.

Típicamente, los modelos de locutor basados en cuantificación vectorial (que en este trabajo se denominan simplemente *modelos VQ*), representan el espacio acústico de cada locutor mediante un codebook [Soong85]. Es decir, cuantifican los vectores parametrizados del conjunto de entrenamiento para generar L codebooks $CL=\{C_l;\ l=1,\ ..,\ L\}$, correspondientes a los L locutores del sistema (obviamente, el codebook de cada locutor se estima únicamente con sus vectores de entrenamiento). El codebook se estima de forma que la distancia entre sus centroides y los vectores de entrenamiento sea la mínima posible. Para clasificar una observación de test $Y=\{\mathbf{y}_0, ..., \mathbf{y}_T\}$, se suman los errores de cuantificación (la distancia al centroide más cercano) de cada vector con respecto a cada codebook C_l de CL. Dada la observación Y, la distorsión media $D^{(l)}$ para el codebook C_l se estima de la siguiente manera:

$$D^{(l)} = \frac{1}{T}\sum_{t=1}^{T} \min_{1\le n\le N_c} d(\mathbf{y}_t, \mathbf{c}_n^{(l)}) \tag{2.9}$$

De ahí, se obtienen L puntuaciones (distorsiones), una para cada locutor, eligiendo a aquel locutor que proporcione una distorsión mínima:

$$\hat{l} = \operatorname{argmin}_{1\le l\le L} D^{(l)} \tag{2.10}$$

En los sistemas de verificación, la distorsión sólo se calcula con respecto al codebook del locutor objetivo, y para tomar una decisión este valor se compara con un umbral predefinido.

En este trabajo, los modelos VQ aplicados siguen una aproximación distinta a la metodología estándar. Nuestros modelos son distribuciones de etiquetas de cuantificación vectorial (a partir de ahora, se denominarán *modelos e-VQ*). Es una metodología sencilla que ha mostrado muy buen rendimiento en adaptación al locutor mediante agrupamiento de locutores en reconocimiento del habla [Rodriguez04]. Consiste en estimar un único codebook C_g con todo el conjunto de entrenamiento, codebook que comparten todos los locutores. El codebook se estima mediante el algoritmo e-LBG [Patane01] y minimiza la distorsión media (distancia euclídea) en la cuantificación de los vectores de entrenamiento. Una vez estimado el codebook, cada vector de entrenamiento se reemplaza por el índice del centroide más cercano. Este proceso se repite con las observaciones de test, donde a cada vector parametrizado se le vuelve a asignar el índice del centroide más cercano de C_g, obteniendo una secuencia equivalente de etiquetas acústicas. Los modelos de locutor son distribuciones de etiquetas: como parámetros específicos de cada locutor se almacenan las frecuencias de las etiquetas observadas en las muestras de entrenamiento de cada locutor.

Supóngase que $X^{(l)}$ es el conjunto de entrenamiento correspondiente al locutor l, $c^{(l)}$ el número de etiquetas que aparece en $X^{(l)}$ y $c(k,l)$ el número de veces que aparece la etiqueta k en $X^{(l)}$. La probabilidad de la etiqueta k dado el locutor l, se puede calcular empíricamente de la siguiente manera:

$$P(k|l) = \frac{c(k,l)}{c(l)} \tag{2.11}$$

Finalmente, suponiendo independencia estadística entre etiquetas sucesivas, la probabilidad de la secuencia de etiquetas $Y=\{Y_1, ..., Y_T\}$ correspondiente a la observación de test $Y=\{\mathbf{y}_0, ..., \mathbf{y}_T\}$, dado el locutor l, puede calcularse como sigue:

$$P(Y|l) = \prod_{t=1}^{T} P(Y_t|l) \tag{2.12}$$

Aunque los modelos VQ y e-VQ son eficientes y todavía se emplean en determinadas circunstancias, existe evidencia empírica de que para el reconocimiento de locutores su rendimiento es inferior al de las técnicas basadas en GMM o SVM.

2.4.2.2 Mezclas de gaussianas estimadas por el algoritmo EM (GMM(EM))

Un GMM es una combinación lineal de M funciones de densidad de probabilidad (M componentes gaussianas, cada componente ponderada por un coeficiente W_m), que representa la densidad de probabilidad de la clase a modelar (en este caso, un locutor l):

$$p(\mathbf{x}|\lambda_l) = \sum_{m=1}^{M} W_m \frac{1}{(2\pi)^{d/2}|\Sigma_m|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_m)^T {\Sigma_m}^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)\right) \tag{2.13}$$

donde $p(\mathbf{x}|\lambda_l)$ es la probabilidad condicional de la observación $\mathbf{x}$ (d-dimensional), dado el modelo λ_l del locutor l. Cada componente gaussiana se parametriza mediante el vector de medias $\boldsymbol{\mu}_m$, de dimensión $d \times 1$, y la matriz de covarianza Σ_m, de dimensión $d \times d$. Los pesos W_m toman valores positivos y deben sumar 1: $\sum_{m=1}^{M} W_m = 1$. En la bibliografía se pueden encontrar diversas técnicas para estimar los parámetros W_m, $\boldsymbol{\mu}_m$ y Σ_m de un GMM, entre los que la más empleada es la estimación por máxima verosimilitud (*Maximum Likelihood*, ML) que trata de estimar los parámetros que maximizan la probabilidad de los datos de entrenamiento, dado el modelo. Dada una secuencia de entrenamiento $X=\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, compuesta por N vectores, la probabilidad condicional de X dado el modelo λ_l es:

$$p(X|\lambda_l) = \prod_{n=1}^{N} p(\mathbf{x}_n|\lambda_l) \tag{2.14}$$

Como esta función es una expresión no-lineal de los parámetros del modelo y no se puede realizar una maximización directa, generalmente se aplica el algoritmo esperanza-maximización (EM) [Dempster77] para estimar los parámetros del modelo. El algoritmo EM busca estimaciones de máxima verosimilitud de los parámetros en modelos probabilísticos que dependen de variables no observables. Alterna el cálculo de la esperanza matemática (paso E) de la probabilidad de la estimación en curso y el reajuste de parámetros que maximizan (paso M) el cálculo de la esperanza. Básicamente, el algoritmo EM parte de un modelo inicial, $\lambda^{(0)}$, y estima iterativamente nuevos modelos, $\lambda^{(k)}$, que cumplen la siguiente condición: $p(X|\lambda^{(k)}) \geq p(X|\lambda^{(k-1)})$. El modelo nuevo $\lambda^{(k)}$ se

convierte en el modelo inicial de la siguiente iteración, y el proceso se repite hasta que se cumpla algún criterio de convergencia. Hay dos factores críticos en la estimación de un GMM por EM: (1) la determinación del orden M del modelo; y (2) la inicialización de los parámetros W, $\boldsymbol{\mu}$ y Σ.

Los trabajos [Reynolds95a] [Reynolds95b] detallan la aplicación de GMM estimados por EM (para los que utilizaremos el acrónimo *GMM(EM)* en este trabajo) en tareas de identificación y verificación de locutores. Generalmente, se emplean matrices de covarianza diagonales porque simplifican la parametrización de los modelos, sin apenas afectar el rendimiento. Aunque los parámetros no son estadísticamente independientes, no es necesario que las matrices de covarianza sean no diagonales. La combinación lineal de la mezcla modela las correlaciones de los vectores parametrizados. Además, existe evidencia empírica de que se puede igualar el rendimiento de un GMM con matrices de covarianza no-diagonales, mediante otro GMM con matrices de covarianza diagonales y un número mayor de componentes. Como criterio de convergencia del algoritmo EM se emplea un número máximo de iteraciones (normalmente, el algoritmo converge en 10-15 iteraciones).

Una aproximación muy empleada, sobre todo en tareas de verificación, es el empleo de dos GMM, en vez de uno [Reynolds95b]. El primer GMM modela el espacio acústico del locutor objetivo y el segundo GMM, es un modelo universal de background (UBM), independiente del locutor, que trata de modelar las características comunes a todos los locutores. Por tanto, con objeto de verificar positivamente a un locutor objetivo, es deseable que los vectores $\mathbf{x}_n$ correspondientes al locutor obtengan una puntuación alta, y los vectores $\mathbf{x}_n$ que son comunes, obtengan una puntuación baja. Para ello, se estima el ratio entre las probabilidades condicionales obtenidas con el modelo λ_l del locutor l, y el modelo universal λ_{UBM} (véase la estimación del ratio en la ecuación 2.7). Supóngase un tramo de señal que obtiene una verosimilitud alta, s_l, para λ_l. Si el tramo es común en muchos locutores, λ_{UBM} proporcionara también una verosimilitud s_{UBM} alta. En consecuencia, al calcular el ratio $s=s_l/s_{\mathrm{UBM}}$, resultará $s\sim1$. Por el contrario, si el tramo corresponde al locutor objetivo, será $s_l \gg s_{\mathrm{UBM}}$, y en consecuencia $s \gg 1$. Esta metodología de modelado ha proporcionado buen rendimiento cuando las señales que se modelan y clasifican provienen de la misma fuente acústica, es decir, cuando los modelos dan cobertura a los tramos a reconocer. Obviamente, al aplicar el ratio se obtienen mejores resultados que con un único modelo, porque las puntuaciones están normalizadas.

El número de componentes de un modelo de locutor de tipo GMM(EM) depende de la cantidad de datos de entrenamiento, pero generalmente se definen mezclas de entre 64 y 256 componentes gaussianas. Son modelos de muy fácil implementación y computacionalmente eficientes para el reconocimiento de locutores independiente del texto. Su principal inconveniente es que no capturan las características temporales del habla. No obstante, este es un aspecto que aporta mejoras muy limitadas cuando el contenido lingüístico de las observaciones a evaluar no coincide con el contenido lingüístico de las observaciones de entrenamiento.

2.4.2.3 Mezclas de gaussianas estimadas por adaptación bayesiana del UBM (GMM(MAP))

El GMM de un locutor se puede estimar mediante adaptación bayesiana de los parámetros de las componentes gaussianas de un GMM de referencia (generalmente un UBM). La adaptación se realiza a partir de los datos de entrenamiento del locutor y se conoce como Aprendizaje Bayesiano o estimación *Máximum A Posteriori* (MAP) [Gauvain94]. En este trabajo se denominarán *modelos GMM(MAP)*. El término de adaptación bayesiana se utiliza por su similitud con la adaptación al locutor en reconocimiento automático del habla. Los modelos GMM(MAP) ajustan sus parámetros a partir de los parámetros del UBM, que razonablemente se habrán estimado con muchos datos y de forma robusta. En consecuencia, el modelo de locutor puede entenderse como el desplazamiento frente al UBM, y el LLR puede interpretarse como la medida de ese desplazamiento de la observación hacia el locutor objetivo.

La adaptación de parámetros sigue las ecuaciones de estimación MAP de un GMM que se describen en el trabajo de Gauvain y Lee [Gauvain94]. En el artículo de Reynolds [Reynolds00] pueden encontrarse más detalles sobre la aplicación de estas ecuaciones y los modelos de locutor GMM(MAP) en la verificación de locutores. Al igual que el algoritmo EM, la adaptación MAP consta de dos pasos:

(1) El primer paso es idéntico al paso de E del algoritmo EM: calcula la esperanza matemática de la probabilidad de cada componente del UBM para los datos de entrenamiento del locutor (la probabilidad, y los primeros y segundos momentos necesarios para estimar el peso, la media y la varianza de las componentes adaptadas).

(2) El segundo paso es diferente al paso M del algoritmo EM: los parámetros estimados en el paso anterior se combinan con los parámetros del UBM mediante un coeficiente de adaptación. El coeficiente de adaptación se define de forma que orienta las componentes de mayor probabilidad hacia los parámetros re-estimados, y por el contrario, las componentes de menor probabilidad hacia los parámetros del UBM. El coeficiente de adaptación se controla mediante un parámetro r denominado factor de relevancia y la probabilidad re-estimada de cada componente; cuando mayor es r, mayor debe ser la probabilidad del componente para que la adaptación tenga lugar. Es decir, determina el grado de adaptación de las componentes; o dicho de otro modo, la proporción de datos que deben estar próximos a una componente para que los parámetros re-estimados reemplacen a los parámetros del UBM.

Aunque mediante la adaptación MAP se pueden re-estimar todos los parámetros del modelo (la media, el peso y la matriz de covarianza, en la práctica se obtiene mejor rendimiento (tanto en cuanto al reconocimiento como al cómputo) cuando únicamente se adaptan las medias [Reynolds00].

Por otro lado, los modelos adaptados permiten la aplicación de una metodología muy eficiente para estimar el LLR de un modelo de locutor. En vez de evaluar la verosimilitud de los dos modelos (el modelo adaptado y el UBM), que puede resultar costoso si las mezclas contienen un número elevado de componentes, se emplea una técnica de evaluación rápida, basada en las dos apreciaciones

siguientes: (1) cuando se evalúa la verosimilitud de un vector para una mezcla que consta de un número elevado de componentes la verosimilitud es muy próxima a la estimada con las *C* componentes de mayor probabilidad; y, (2) las componentes del modelo de locutor están muy relacionadas con las del UBM, por lo que es razonable asumir que las componentes que contribuyen en el modelo de locutor serán las mimas que las contribuyen en el UBM. Partiendo de estas dos apreciaciones, el LLR de un modelo adaptado se calcula como sigue: para cada vector de la observación, se determinan las *C* componentes de mayor probabilidad del UBM y se estima su verosimilitud empleando únicamente estas *C* componentes. La verosimilitud del locutor se estima, de nuevo, con las mismas *C* componentes pero correspondientes al modelo de locutor. Con esta aproximación, para un modelo UBM de *M* componentes el número de cálculos se reduce de 2*M* a *M*+*C*.

Los resultados encontrados en la literatura indican que los modelos GMM(MAP) presentan mejor rendimiento que los modelos GMM(EM). Se cree que podría ser debido a que los modelos adaptados son más robustos frente a las fuentes no-modeladas, es decir frente a las clases acústicas que no han aparecido en el entrenamiento de un locutor. En términos generales, considerando que el UBM da cobertura a todo el espacio de clases acústicas de forma independiente al locutor, la adaptación es un "ajuste" al locutor. Los parámetros de las componentes que no se han observado en entrenamiento se copian directamente del UBM, y las que se han observado, se ajustan a las características del locutor. En consecuencia, los vectores correspondientes a las clases que no se han observado en entrenamiento producen un LLR prácticamente nulo en reconocimiento, con un factor de normalización que no va en contra, ni a favor, del locutor hipotético. El modelo de locutor producirá una probabilidad baja para las clases acústicas que no han aparecido en entrenamiento, y por tanto, se obtendrán LLR inferiores. Aunque este factor es favorable para el habla que no proviene del locutor, es perjudicial para los casos en los que la observación a evaluar proviene del locutor y contiene alguna característica no presente en entrenamiento.

2.4.2.4 Máquinas de vectores de soporte (SVM)

Las SVM son sistemas de clasificación de gran eficiencia que han adquirido mucha popularidad en los últimos años. Las SVM modelan los límites entre la clase objetivo (el locutor) y la población global (el conjunto de impostores), a diferencia de las metodologías anteriormente citadas (GMM(EM), GMM(MAP)), que modelan de forma separada la distribución de probabilidad del locutor objetivo y la de la población global.

Las SVM aplican un algoritmo de clasificación binaria [Cristianini00] [Burges98], que realiza previamente el mapeo $\mathbf{x}: \rightarrow f(\mathbf{x}, \mathbf{w})$, donde f es una función de aproximación y $\mathbf{w}$ un vector de parámetros estimado con los datos de entrenamiento con objeto de minimizar el número y magnitud de los errores clasificación, $e_i - f(\mathbf{x}_n, \mathbf{w})\ \forall n$, para la categoría e_n del vector de entrada $\mathbf{x}_n$. Dado un conjunto de entrenamiento compuesto por *N* observaciones, el riesgo empírico estima la esperanza de error del clasificador (es decir, de la función objetivo *f*):

$$R_{emp}(w) = \frac{1}{N}\sum_{n=1}^{N} C(e_n - f(\mathbf{x}_n, \mathbf{w})) \quad (2.15)$$

donde C es una función que establece la magnitud de error (el coste de error) de clasificación. Una de las principales características de las SVM respecto a los métodos tradicionales de aprendizaje, es que no se basan en la minimización del riesgo empírico, sino en el principio de minimización del riesgo estructural [Vapnik95], que trata de establecer el mejor compromiso entre el riesgo empírico y la capacidad de generalización del modelo (que está relacionada con la complejidad de la función de decisión). Se busca un modelo que estructuralmente tenga poco riesgo de cometer errores con los datos de test, a costa de que se puedan cometer más errores con los datos de entrenamiento. Consiste en minimizar el riesgo empírico junto con una constante que mide la complejidad del clasificador, T_w, que aumenta cuando decrece el riesgo empírico. El riesgo estrucutral R_s es por tanto:

$$R_s(\mathbf{w}) = T_w + c\, R_{emp}(\mathbf{w}) \quad (2.16)$$

donde c es una constante. R_S trata de establecer el mejor compromiso entre el rendimiento de clasificación de las muestras de entrenamiento y la complejidad del clasificador.

Dadas dos clases categorizadas mediante las etiquetas +1 y -1, las muestras de entrenamiento se pueden representar de la siguiente manera: $\{\mathbf{x}_n, e_n\}$, $n = 1, \ldots, N$, $e_i \in \{-1, +1\}$. Supóngase que en el espacio de las muestras existe un hiperplano multidimensional (hiperplano de separación) que separa las muestras por clase. Los puntos **x** que se sitúan en el hiperplano cumplen la siguiente condición:

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{w} \cdot \mathbf{x} + b = 0 \quad (2.17)$$

donde **w** es el vector normal al hiperplano y $\frac{|b|}{\|\mathbf{w}\|}$ la distancia perpendicular del hiperplano al origen ($\|w\|$ es la norma euclidea de $\boldsymbol{w}$). Sean d_+ y d_- las distancias más cortas del hiperplano de separación a las muestras de la clase positiva y negativa, respectivamente. El "*margen*" del hiperplano es la suma $d_+ + d_-$. Cuando los datos se pueden separar por un hiperplano (linealmente), el algoritmo SVM busca simplemente aquél que determina el margen máximo. Estos conceptos, que se muestran esquemáticamente en la **Figura 2.4**, cumplen la siguiente formulación para todas las muestras del conjunto de entrenamiento:

$$f(\mathbf{x}, \mathbf{w}) = \begin{cases} \geq +1 & e_n = +1 \\ \leq -1 & e_n = -1 \end{cases} \quad (2.18)$$

que se puede representar en una única ecuación de la siguiente manera:

$$e_n.(\mathbf{x}_n \cdot \mathbf{w} + b) - 1 \geq 0 \quad \forall n \quad (2.19)$$

Las muestras que cumplen la igualdad en la ecuación 2.19 (aquellas que se han introducido en un rectángulo en la Figura 2.4), y cuya eliminación cambiaría la resolución del SVM, se denominan *vectores de soporte*. Son los vectores que delimitan el margen. Por tanto, en la fase de entrenamiento, el algoritmo SVM trata de determinar los vectores que definen el margen óptimo. El límite de decisión es aquel que divide por la mitad el margen: el hiperplano de separación $f(\mathbf{x}, \mathbf{w}) = 0$. En consecuencia, las muestras de evaluación **y** que cumplan la condición $f(\mathbf{y}, \mathbf{w}) > 0$ se clasificarán como positivas, y viceversa, las muestras que cumplan la condición $f(\mathbf{y}, \mathbf{w}) < 0$, como negativas.

Para los puntos ubicados en el hiperplano $H_1: f(\mathbf{x}, \mathbf{w}) = -1$ (los vectores soporte de la clase -1) con vector normal **w**, la distancia perpendicular al origen es $\frac{|b-1|}{\|\mathbf{w}\|}$. De forma análoga, la distancia al origen de los puntos ubicados en el hiperplano $H_2: f(\mathbf{x}, \mathbf{w}) = +1$ (los vectores soporte de la clase +1) con vector normal **w**, es $\frac{|b+1|}{\|\mathbf{w}\|}$. Por tanto, $d_+ = d_- = \frac{1}{\|\mathbf{w}\|}$, y el margen es $\frac{2}{\|\mathbf{w}\|}$.

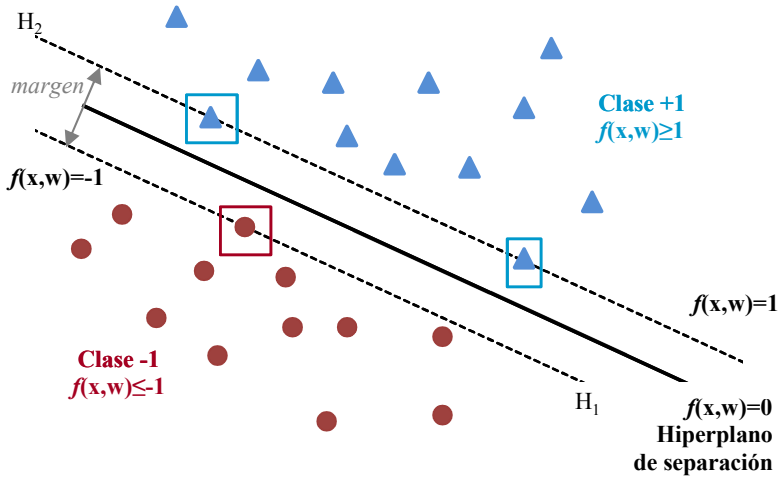

Figura 2.4. *Esquema conceptual de una máquina de vectores soporte (SVM)*

Véase como H_1 y H_2 son paralelos (tienen el mismo vector normal), por lo que el SVM debe encontrar los hiperplanos H_1 y H_2 que maximizan el margen: el $\|\mathbf{w}\|^2$ mínimo bajo las restricciones de la ecuación 2.19. La optimización se puede resolver introduciendo un multiplicador de Lagrange, $\alpha_i \geq 0$, por cada restricción de la ecuación 2.19, que conduce a la siguiente ecuación de Lagrange:

$$L_P = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} \alpha_n e_n (\mathbf{x}_n \cdot \mathbf{w} + b) \tag{2.20}$$

L_P se debe minimizar con respecto a **w** y b. De su derivación, se obtienen las dos siguientes condiciones:

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n e_n \mathbf{x}_n \tag{2.21}$$

$$\sum_{n=1}^{N} \alpha_n e_n = 0 \tag{2.22}$$

Al sustituir las ecuaciones 2.21 y 2.22 en 2.20, se eliminan **w** y b, y se llega al siguiente problema dual de optimización:

$$L_D = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j e_i e_j \, \mathbf{x}_i \cdot \mathbf{x}_j \tag{2.23}$$

bajo las condiciones $\alpha_n \geq 0$ para todo n y $\sum_{n=1}^{N} \alpha_n e_n = 0$. La maximización de L_D es un problema de programación cuadrática, que consiste en ajustar los multiplicadores de Lagrange, α, y produce la siguiente función objetivo mediante técnicas de resolución estándares (para más detalles, véase [Burges98]).

$$f(\mathbf{x}) = \sum_n \alpha_n e_n (\mathbf{x}_n \cdot \mathbf{x}) + b \tag{2.24}$$

En la ecuación 2.24 se observa que el límite de decisión es una combinación lineal de los vectores de entrada y sus etiquetas (para los vectores con $\alpha_i > 0$). De las condiciones de *Karush-Kuhn-Tucker* (KKT) [Fletcher87] se obtiene:

$$\alpha_n [e_n (\mathbf{x}_n \cdot \mathbf{w} + b) - 1] = 0 \ \ \forall n \tag{2.25}$$

Cualquier solución de un problema de optimización con restricciones cumple las condiciones KKT; la intersección del conjunto de direcciones válidas y el conjunto de direcciones descendentes coincide con la intersección del conjunto de direcciones válidas para las restricciones lineales y el conjunto de direcciones descendentes. Esta condición se cumple en todos los algoritmos SVM dado que las restricciones son lineales. Además, hay que tener en cuenta que la función objetivo de un SVM es convexa, y en este caso las condiciones KKT son suficientes y necesarias para que **w**, b y α formen una solución. Por tanto, el hecho de resolver un problema SVM es equivalente a encontrar una solución para las condiciones KKT. Para las muestras del corpus de entrenamiento que no están en el margen (para las que la expresión entre corchetes no es 0) α_i debe ser necesariamente cero. Asimismo, α_i debe ser no nulo para los vectores soporte; es decir, para los vectores que el clasificador debe memorizar porque son los que definen el límite de decisión. Aunque en este momento no quede del todo claro por qué se deben memorizar los vectores soporte y no el vector normal **w** que se puede calcular en entrenamiento, es un concepto que se aclarará al tratar de solucionar los problemas que no admiten un clasificador lineal. Por otro lado, hay que destacar que el umbral b se puede resolver mediante la condición complementaria KKT (de la ecuación 2.25), por medio de una muestra x_n para la que $\alpha_n \neq 0$.

Por otro lado, supóngase que los datos de entrenamiento no se pueden separar linealmente por lo que siempre resultan mal clasificadas algunas muestras de entrenamiento al definir el hiperplano de separación. En estas situaciones, el algoritmo debe asumir la posibilidad de cometer errores en la clasificación y penalizarlos. Para ello, se introducen variables de holgura (*slack variables*), ξ_n, en la función objetivo. Estas variables permiten que algunas muestras resulten mal clasificadas o ubicadas dentro del margen:

$$f(\mathbf{x}, \mathbf{w}) = \begin{cases} \geq +1 - \xi_n & e_n = +1 \\ \leq -1 + \xi_n & e_n = -1 \end{cases} \tag{2.26}$$

para $\xi_n \geq 0 \ \forall n$. De nuevo, estas condiciones se puede representar mediante la siguiente ecuación:

$$e_n . (\mathbf{x}_n \cdot \mathbf{w} + b) \geq 1 - \xi_n \ \ \forall n \tag{2.27}$$

En la **Figura 2.5** se muestra el funcionamiento de un SVM con variables de holgura. Los puntos mal clasificados se indican mediante un recuadro, y se incluye una barra que reflejaría la distancia al hiperplano correspondiente a su categoría. Esta distancia es generalmente el valor ξ_n de la variable de holgura de la muestra (valor que normalmente se pondera con un coste de error). Sólo los vectores que están dentro del margen o resultan mal clasificados tendrán variables de holgura no cero. La optimización del margen se realiza teniendo en cuenta la penalización de estos errores. Ahora, además

de los puntos que se sitúan en el hiperplano de separación de cada clase, las muestras mal clasificadas y las que están dentro del margen son también vectores de soporte. Recuérdese que se denominan vectores de soporte aquellos puntos cuya eliminación cambia la resolución del SVM. En este caso, son aquellos puntos que maximizan el margen y minimizan la distancia entre los hiperplanos y las muestras mal clasificadas (de ahí, la ecuación 2.25 pasa a ser $\alpha_n[e_n(\mathbf{x}_n \cdot \mathbf{w} + b) - 1 + \xi_n] = 0 \ \forall n$ donde $\xi_n \geq 0$).

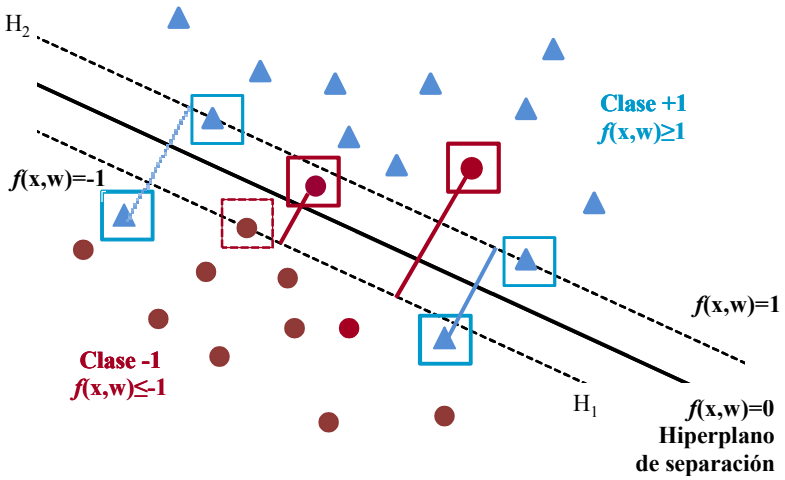

Figura 2.5. *Esquema conceptual de una máquina de vectores soporte (SVM) con variables de holgura*

La descripción proporcionada hasta ahora se ha centrado en la resolución de problemas que se pueden resolver mediante un clasificador lineal. No obstante, en la mayoría de aplicaciones reales, el espacio de muestras no es linealmente separable. En estos casos, se emplea una estrategia muy sencilla que consiste en aplicar una transformación no-lineal sobre el espacio de entrada, para obtener un espacio de mayor dimensión donde las clases, presumiblemente, sí son linealmente separables. El algoritmo SVM sigue buscando la maximización del margen, pero en estos problemas, la maximización se lleva a cabo sobre el espacio transformado, denominado espacio de características (*feature space*). El hiperplano óptimo se calcula sobre el espacio de características y, de hecho, es donde se realiza la clasificación lineal. Como se puede observar en la ecuación 2.24, la función objetivo de un problema lineal es una suma de productos escalares ponderados, entre la muestra (vector) de entrada y los vectores de entrenamiento. En problemas no lineales, en vez de realizar el producto escalar explícito con los vectores expandidos, se emplea una función denominada *kernel* [Scholkopf02]. El kernel representa la combinación de expansión y producto realizados sobre los vectores de entrada, directamente.

Por ejemplo, dados los vectores de entrada bi-dimensionales $\mathbf{x}_1=(x_{11}, x_{12})$ y $\mathbf{x}_2=(x_{21},x_{22})$, sus vectores de características son $\bar{\mathbf{x}}_1=(x_{11}, x_{12},x_{11}x_{12})$ y $\bar{\mathbf{x}}_2=(x_{21}, x_{22}, x_{21}x_{22})$, que incluyen el producto

de sus coeficientes, el producto escalar de los vectores expandidos se puede resolver de la siguiente manera:

$$\bar{\mathbf{x}}_1 \cdot \bar{\mathbf{x}}_2 = \begin{pmatrix} x_{11} \\ x_{12} \\ x_{11}x_{12} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} x_{21} \\ x_{22} \\ x_{21}x_{22} \end{pmatrix} = \mathbf{x}_1 \cdot \mathbf{x}_2 + x_{11}x_{12}x_{21}x_{22} \tag{2.28}$$

En la ecuación 2.28 se observa cómo el producto escalar $\bar{\mathbf{x}}_1 \cdot \bar{\mathbf{x}}_2$ se puede definir sobre los vectores de entrada $\mathbf{x}_1$ y $\mathbf{x}_2$. El kernel de este ejemplo sería:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2 + x_{11}x_{12}x_{21}x_{22} \tag{2.29}$$

De ahí, que en problemas no-lineales, el límite de decisión (ecuación 2.24) se calcule mediante un kernel:

$$f(\mathbf{y}) = \sum_{n=1}^{N} \alpha_i e_i K(\mathbf{y}, \mathbf{x}_n) + b \tag{2.30}$$

Esta ecuación justifica la decisión de tomar los vectores soporte como solución del problema (y no el vector normal **w)**. En general, el vector normal del hiperplano de separación del espacio de características no se puede representar mediante una transformación no-lineal de un único vector en el espacio de entrada. Sí, en cambio, mediante el producto escalar de dos vectores del espacio de características, tal como se ha visto al aplicar el kernel.

Hasta ahora se ha analizado cómo se puede realizar la clasificación definiendo una superficie de separación plana. Para ello, se ha aplicado un kernel $K(\mathbf{x}, \mathbf{y})$ lineal, que es el tipo de kernel más básico (y también el más aplicado en las aproximaciones de referencia en el campo del reconocimiento de locutores). En los espacios de características donde la superficie de separación no se puede definir mediante un hiperplano, existe la posibilidad de emplear un kernel de otro tipo. Entre los más conocidos y empleados destacan el kernel polinómico ($(a\mathbf{x} \cdot \mathbf{y} + b)^n$) y el radial ($\exp\left(-\frac{1}{2\sigma}(\mathbf{x} - \mathbf{y})^2\right)$).

Formalmente, el kernel lineal se define de la siguiente manera:

$$K(\mathbf{x}, \mathbf{y}) = b(\mathbf{x})^{\mathrm{T}} b(\mathbf{y}) \tag{2.31}$$

donde $b(\mathbf{x})$ es la transformación al espacio de características. Para que una función kernel sea válida debe cumplir la condición de Mercer [Vapnik95][Courant53],que exige que, dada cualquier función $g(\mathbf{x})$ para la cual la siguiente integral sea finita:

$$\int g(\mathbf{x})^2 d\mathbf{x} \tag{2.32}$$

el kernel debe cumplir la siguiente condición:

$$\int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0 \tag{2.33}$$

En general no es fácil comprobar si se cumple o no la condición de Mercer, ya que la ecuación 2.33 se debe cumplir para todas las funciones $g(\mathbf{x})$ de cuadrado integrable. No obstante, puede comprobarse que cualquier kernel que se pueda expresar como $K(\mathbf{x}, \mathbf{y}) = \sum_{p=0}^{\infty} c_p (\mathbf{x} \cdot \mathbf{y})$, donde c_p es un coeficiente

real positivo y la serie es convergente uniforme, cumple la condición de Mercer. Con todo ello podemos decir que el kernel es el componente crítico de un SVM: debe definir una métrica adecuada para la clasificación en el espacio de características y debe ser computacionalmente eficiente. Aún y todo, el comportamiento del método es complejo y se conocen circunstancias como, por ejemplo, que si el conjunto de entrenamiento produce una matriz Hessiana $K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)\ \forall i,j$ semi-definida positiva, el algoritmo SVM converge incluso para los kernels que no cumplen la condición de Mercer.

Como clasificadores binarios, los SVM resultan muy apropiados para la verificación de locutores: se entrena un modelo con las señales de entrenamiento del locutor objetivo como muestras positivas, y un conjunto de impostores, denominado conjunto de locutores externos (o de *background*), como muestras negativas. El conjunto de background debe representar el espacio acústico de todos los posibles impostores, donde se deben incluir, de forma especialmente importante, locutores de gran semejanza al locutor del modelo.

Debido a que todo sistema de clasificación se puede resolver con un clasificador binario mediante la estrategia de "divide y vencerás", los SVM son aplicables también en problemas de clasificación multi-clase, como la identificación de locutores. Generalmente, se entrenan tantos clasificadores como clases para la identificación. Todas las muestras de entrenamiento que no corresponden al locutor se emplean como muestras negativas, y obviamente, como muestras positivas las muestras de entrenamiento del locutor. De esta forma, se estima un modelo SVM para cada locutor.

En los sistemas de reconocimiento de locutores que emplean SVM, el kernel se aplica sobre toda la secuencia de observaciones, lo que se denomina kernel de secuencia (*sequence kernel*) [WMCampbell02]. En vez de modelar los vectores que representan los tramos individuales de una señal, este método modela toda la secuencia de vectores, es decir toda la pronunciación. Básicamente, trata de comparar dos pronunciaciones, *X* e *Y*, utilizando el kernel. Si el kernel es lineal se realiza de la siguiente manera:

$$K(X,Y) = b(X)^{\mathrm{T}} b(Y) \tag{2.34}$$

Existen diversas variaciones, entre las que destacan el kernel de secuencia de discriminación lineal generalizada (*Generalized Lineal Discriminant Sequence Kernel*, GLDS kernel) [WMCampbell02], el kernel de n-gramas [WMCampbell04], el kernel de transformadas de regresión lineal de máxima verosimilitud (*Maximum-Likelihood Linear Regression transform kernels*, MLLR kernel) [Stolcke05] y el kernel de vectores de soporte GMM [WMCampbell06a]. Los MLLR kernel y kernel de n-gramas se aplican sobre parámetros de alto nivel que no se han empleado en este trabajo. Los sistemas SVM analizados e implementados en este trabajo, cuya breve descripción se ofrece a continuación, son los basados en el GLDS kernel y el kernel de vectores de soporte GMM.

SVM con un GLDS kernel (SVM-GLDS)

En los clasificadores SVM con un GLDS kernel (denominados SVM-GLDS en este trabajo) [WMCampbell02], la función objetivo es una función polinómica discriminante de grado *n*

(generalmente *n*=2 o *n*=3). Suponiendo un vector bi-dimensional $\mathbf{x} = (x_1\ x_2)$ y una expansión polinómica de grado 2, el vector de características sería:

$$b(\mathbf{x}) = (1, x_1, x_2, {x_1}^2, x_1x_2, {x_2}^2) \tag{2.35}$$

Supónganse dos secuencias de observaciones $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ e $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_T\}$. El polinomio de discriminación se entrena minimizando el error cuadrático medio entre los vectores de características, que da como resultado el siguiente kernel:

$$k(X, Y) = b(X)^T R^{-1} b(Y) \tag{2.36}$$

donde $b(\cdot)$ es la media de la expansión polinómica de todas las observaciones de la secuencia:

$$b(X) = \frac{1}{N}\sum_{n=1}^{N} b(\mathbf{x}_n) \tag{2.37}$$

y la matriz *R*, es una matriz de correlación de un conjunto de datos de referencia. Se calcula sobre un conjunto global expandido (conjunto de *background*) $Z = \{\mathbf{z}_{1,\dots,}\mathbf{z}_K\}$, y generalmente sólo se extraen los elementos diagonales:

$$R = \begin{bmatrix} b(\mathbf{z}_1)^{\mathrm{T}} \\ b(\mathbf{z}_2)^{\mathrm{T}} \\ \vdots \\ b(\mathbf{z}_K)^T \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} b(\mathbf{z}_1)^{\mathrm{T}} \\ b(\mathbf{z}_2)^{\mathrm{T}} \\ \vdots \\ b(\mathbf{z}_K)^T \end{bmatrix} \tag{2.38}$$

Para más detalles sobre el origen de estas ecuaciones véase [WMCampbell02]. En [WMCampbell06b] se describe la aplicación de sistemas SVM-GLDS en tareas de reconocimiento de locutores y verificación de la lengua.

SVM con un kernel secuencial de vectores soporte GMM (GMM-SVM)

Los sistemas de clasificación SVM con un kernel secuencial de vectores soporte GMM (denominados GMM-SVM en este trabajo) aplican un sistema de clasificación SVM sobre el espacio vectorial definido por parámetros GMM adaptados a partir de un GMM base [WMCampbell06a]. En definitiva, son modelos discriminativos (SVM) definidos sobre modelos generativos (GMM) [Jaakkola98].

Dada una pronunciación X_a, se estima un GMM que adapta las medias de un GMM(MAP) (el GMM base). Dado un UBM de *M* componentes, los vectores de soporte GMM se forman apilando los parámetros adaptados. Habitualmente, sólo se utilizan las medias adaptadas: $\overline{\boldsymbol{\mu}} = (\boldsymbol{\mu}_1^a, \boldsymbol{\mu}_2^a, \dots, \boldsymbol{\mu}_M^a)$. Los parámetros adaptados se comparan mediante una aproximación de la divergencia Kullback-Leibler, que trata de calcular la distancia entre dos modelos y da como resultado el siguiente kernel lineal:

$$K(X_a, X_b) = \sum_{m=1}^{M} \left(\sqrt{W_m}\Sigma_m^{-\left(\frac{1}{2}\right)}\boldsymbol{\mu}_m^a\right)^{\mathrm{T}} \left(\sqrt{W_m}\Sigma_m^{-\left(\frac{1}{2}\right)}\boldsymbol{\mu}_m^b\right) \tag{2.39}$$

donde W_m y Σ_m son el peso y la matriz de covarianza, respectivamente, de la componente *m* del UBM. La descripción de esta metodología, así como del origen del kernel, se pueden encontrar en [WMCampbell06a]. Hay que destacar que los modelos GMM-SVM, como técnica de modelado, constituyen una de las metodologías más novedosas y más eficientes dentro del área del

reconocimiento de locutores, ya que combinan de forma inteligente las características de los modelos generativos y los modelos discriminativos.

2.4.3 Normalización de probabilidades

Todo sistema de verificación debe tomar la decisión definitiva de aceptar o rechazar la hipótesis propuesta, que en este caso es la identidad del locutor objetivo. Para ello, la probabilidad a posteriori del locutor objetivo se compara con un umbral, generalmente fijado de antemano. Si la probabilidad es superior al umbral, se decide que el locutor objetivo se corresponde con la señal de test; en caso contrario, se decide que se trata de un impostor. La determinación del umbral es una tarea difícil y problemática, debido a la variabilidad que presentan las probabilidades en diferentes experimentos. Esta variabilidad proviene de diversas fuentes:

- La variabilidad acústica de las señales, en cuanto al contenido fonético, duración, canal de grabación, ruido ambiental, etc.
- La disparidad (*mismatch*) acústica entre los datos de entrenamiento y los datos de test, debida a dos factores principales: (1) variabilidad intra-locutor (debida al estado emocional, estado de salud, etc.); y (2) la variabilidad debida al canal de transmisión, al contenido de las señales o al entorno acústico.
- La variabilidad inter-locutor, variabilidad relacionada con las características acústicas de la voz de diferentes locutores. Esta variabilidad, necesaria para distinguir unos locutores de otros, requiere especial atención en sistemas que pretenden establecer un umbral de decisión único para todo el sistema, es decir, un umbral independiente del locutor.

Al emplear el término "normalización de probabilidades", se hace referencia al conjunto de técnicas que buscan minimizar la sensibilidad a la variabilidad acústica de las señales y a la variabilidad intra-locutor, para de esta forma, optimizar la toma de decisiones. Se aplican sobre las probabilidades estimadas con modelos probabilísticos, tales como las GMM o SVM, entre otros. En definitiva, son metodologías que tratan de resolver los errores ocasionados por la variabilidad estadística de las probabilidades de clasificación, mapeando las probabilidades a una distribución global donde se pueda establecer un umbral de clasificación independiente del locutor y de la señal de evaluación.

La mayoría de las técnicas de normalización de probabilidades propuestas en la bibliografía se basan en el trabajo de Li y Porter [Li88]. En este trabajo, debido a que se observa una gran variabilidad en las distribuciones de probabilidades de los locutores objetivo e impostores, se propone la normalización de la distribución de impostores para, de esta forma, reducir la varianza de la distribución global (distribución de locutores e impostores). La experimentación se aplica sobre 26 locutores objetivo, donde para cada locutor, los 25 restantes actúan como impostores. La normalización pretende centrar la distribución de impostores y se aplica sobre las probabilidades estimadas mediante los modelos de locutor. Dada la probabilidad a posteriori $p(\lambda_l|X)$, su valor normalizado $\tilde{p}(\lambda_l|X)$ se calcula de la siguiente manera:

$$\tilde{p}(\lambda_l|X) = \frac{p(\lambda_l|X) - \mu_{\lambda_l}}{\sigma_{\lambda_l}}, \tag{2.40}$$

donde μ_{λ_l} y σ_{λ_l} son, respectivamente, la media y la desviación estándar de las probabilidades estimadas con la distribución de impostores del locutor *l* para un conjunto de pronunciaciones. Esta metodología de normalización se basa en que, para los segmentos analizados, se observa una gran correlación entre las probabilidades obtenidas con el modelo de un locutor objetivo y las probabilidades obtenidas con el correspondiente modelo de impostores.

A partir del trabajo de Li et. al. se han propuesto varias técnicas similares de normalización de probabilidades. Entre ellas destacan las denominadas Z-norm, T-norm y ZT-norm:

- Z-NORM: La normalización cero (*Zero-NORMalization*, Z-norm) consiste en evaluar el modelo de locutor para un conjunto de señales de impostores. Las probabilidades obtenidas para este conjunto se modelan mediante una distribución gaussiana unimodal. Los parámetros de normalización, dependientes del locutor, son la media y varianza de esta distribución. Estos parámetros, siguiendo la ecuación 4.8, se aplican sobre la probabilidad del locutor objetivo dada la pronunciación de entrada. Una de las principales ventajas de la Z-norm es que los parámetros de normalización se pueden estimar offline en la fase de entrenamiento.
- T-NORM: La normalización de test (*Test-NORMalization*, T-norm), propuesta en [Auckenthaler00], al igual que la Z-norm, estima la media y la varianza de una distribución gaussiana unimodal de probabilidades de impostores, pero difiere en que, en vez de un conjunto de señales de test de impostores, se emplea un conjunto de modelos de impostores. En este caso, la distribución de impostores se estima a partir de las probabilidades de los modelos de impostores dado el segmento a evaluar. Los parámetros de normalización son la media y la varianza de esta distribución de probabilidades, que se aplican de la misma forma que la Z-norm, siguiendo la ecuación 4.8. La Z-norm se puede considerar como una normalización del locutor (o más concretamente, del modelo de locutor). La T-norm, en cambio, es una técnica dependiente de la señal de test, es decir, del propio conjunto de evaluación, y pretende normalizar la variabilidad de la propia señal a evaluar. La T-norm estima los parámetros de normalización sobre la misma señal con la que se obtiene la probabilidad del locutor objetivo, por lo que resuelve los problemas ocasionados por una posible disparidad acústica entre la señal a evaluar y las señales empleadas para la normalización (como puede suceder en la Z-norm). Sin embargo, los parámetros de la T-norm no se pueden estimar offline, sino que han de calcularse en la misma fase de evaluación, por lo que su coste computacional es superior.
- ZT-NORM: Esta aproximación consiste en aplicar secuencialmente las dos aproximaciones descritas más arriba: Z-norm y T-norm. Concretamente, se aplica una T-norm sobre probabilidades que primero se han normalizado mediante una Z-norm (o viceversa, en algunos trabajos se aplican en orden inverso). De esta forma, se explotan las ventajas de ambas aproximaciones: se normaliza la variabilidad inter-locutor (mediante la Z-norm) y la variabilidad de la pronunciación a evaluar (mediante la T-norm). Esta aproximación ha adquirido mucha

popularidad en los últimos años, sobre todo en sistemas basados en GMM. Se ha empleado en las últimas NIST SRE, como, por ejemplo, en los sistemas desarrollados por, entre otros el MIT e IBM para la NIST SRE06 [WMCampbell07]; y Loquendo y la Universidad Politécnica de Torino[Dalmasso09], el SRI [Kajarekar09] y la Universidad de Ciencia y Tecnología de China [Guo09] para la NIST SRE08. Por otro lado, MacLaren et. al. [MacLaren09] han presentado recientemente un trabajo en el que se trata de determinar conjuntos de impostores refinados para las normalizaciones T-norm, Z-norm y ZT-norm. El conjunto refinado se forma a partir de un conjunto extenso de impostores, en función de la frecuencia con que cada señal del conjunto de impostores es seleccionada como vector de soporte de los modelos GMM-SVM entrenados para clasificar un conjunto de locutores objetivo de desarrollo. En los resultados obtenidos sobre las bases de datos de las NIST SRE de 2006 y 2008 empleando modelos GMM(MAP), el rendimiento de los sistemas mejora al emplear conjuntos refinados con Z-norm, T-norm y ZT-norm, siendo esta última la que proporciona un mejor rendimiento.

Existen otras aproximaciones, similares a las anteriores, pero con características específicas interesantes. Entre ellas, cabe destacar las siguientes:

- La H-norm (*Handset-NORMalization*), propuesta en [Reynolds98], consiste en aplicar una Z-norm de forma separada para cada tipo de terminal de teléfono empleado en el conjunto de evaluación. Se basa en que, al evaluar señales telefónicas, los modelos presentan una gran dependencia del tipo de terminal empleado. En los experimentos del artículo se distingue entre micrófonos de carbono y micrófonos de condensador electret. Por tanto, se estiman parámetros de normalización específicos utilizando señales grabadas con cada tipo de micrófono. Se definen dos conjuntos de parámetros de normalización (media y varianza) para cada locutor, uno para señales que emplean micrófonos de carbono, y otro para los que emplean micrófonos de condensador electret. En evaluación, primero se ha de detectar el tipo de micrófono utilizado en el segmento a evaluar. La normalización se realiza con los parámetros que corresponden a ese tipo de micrófono, de acuerdo, como siempre, a la ecuación 4.8.
- La D-norm (*Distance-NORMalization*), propuesta en [Ben02], trata de eludir la necesidad de definir un conjunto adicional de locutores para la estimación de los parámetros de normalización. Para ello, se generan datos de locutores e impostores mediante el método Monte-Carlo, empleando para ello los modelos del sistema. A partir de estos datos, se calcula la distancia simétrica KL entre el modelo del locutor objetivo y los modelos de referencia. El ratio entre estos modelos se normaliza frente a esta distancia. La principal ventaja de esta aproximación es que no requiere ningún conjunto adicional de locutores para calcular el parámetro de normalización. En los experimentos realizados por los autores, su rendimiento es similar al de la Z-norm.
- La AT-norm (*Adaptive T-norm*), propuesta en [Sturim05], se basa en la T-norm descrita más arriba. Consiste en emplear una cohorte de impostores dependiente del locutor objetivo para estimar los parámetros de la T-norm. La cohorte se define con los modelos más parecidos al

modelo del locutor objetivo. Para ello, se define un conjunto formado por diversas señales de impostores, y un conjunto de modelos T-norm que incluye todo tipo de canales de grabación, duración de las señales, cantidad de sesiones, etc. Dados estos modelos, se estiman las probabilidades de las señales del conjunto de impostores, y con estas probabilidades se define un conjunto de vectores *I,* cuya dimensión coincide con la cantidad de señales del conjunto de impostores, y su volumen con la cantidad de modelos T-norm. Paralelamente se construye otro vector **v** con las probabilidades de las señales del conjunto de impostores para el modelo del locutor objetivo. En el trabajo mencionado, como cohorte se seleccionan los k=55 vectores del conjunto de vectores *I* más parecidos a **v**, siguiendo el conocido algoritmo K-NN (*K Nearest Neighbours*) con la distancia euclidea.

- Recientemente, se ha definido una nueva aproximación denominada KL-Tnorm (*Kullback Leibler T-norm*), que, al igual que la AT-norm, define una cohorte de impostores dependiente del locutor para estimar una T-norm [Ramos-Castro07]. Se emplea una aproximación de la distancia KL para calcular la distancia entre el modelo de locutor y cada modelo del conjunto de modelos T-norm. La selección de modelos para la cohorte se realiza, de nuevo, por medio del algoritmo K-NN, que permite determinar los k modelos de impostor más próximos al modelo del locutor objetivo.

En algunos trabajos, el uso de un UBM o de una cohorte de locutores se clasifican dentro de este tipo de técnicas de normalización de probabilidades. Sin embargo, aunque efectivamente buscan la normalización del modelo de locutor y persiguen objetivos parecidos a las metodologías descritas más arriba, en la formulación seguida en este trabajo, están dirigidas más bien a modelar la hipótesis alternativa en un contraste de hipótesis (ecuación 2.7), y no se consideran como metodologías de normalización de probabilidades.

2.4.4 Fusión de Sistemas

Una práctica común en tareas de clasificación de patrones es la combinación de información proveniente de diversas fuentes de evidencia. En el campo del reconocimiento de locutores, esto se conoce como "fusión de sistemas" [Ramachandran02] [JPCampbell03] [Brummer07]. Consiste en combinar las probabilidades estimadas o las decisiones tomadas por diferentes sistemas de clasificación de locutores que tratan de dar solución a una misma tarea de clasificación. Generalmente, los sistemas considerados aplican diferentes metodologías de extracción de parámetros, diferentes metodologías de modelado o distintos conjuntos de entrenamiento [Ramachandran02], lo que a fin de cuentas se reduce a definir más de un modelo para cada locutor. No obstante, el rendimiento del sistema global puede también mejorar al fusionar sistemas que emplean diferentes metodologías de compensación de canales o normalización de probabilidades, o sistemas que se han desarrollado mediante diferentes herramientas [Brummer07].

Obviamente, no se puede garantizar que mediante la fusión se vaya a mejorar el rendimiento de los sistemas combinados. Llevado al caso más extremo, supóngase que todos los subsistemas dan la

misma salida para las muestras a clasificar. En esta situación, sea cual sea el procedimiento de fusión empleado, resulta imposible mejorar el rendimiento. La fusión debe compensar el comportamiento de los sistemas combinados, de forma que un sistema pueda corregir los errores cometidos por otro, y viceversa. Por lo tanto, es importante (y deseable) que los sistemas a fusionar no estén correlacionados (véase [Ferrer08a] para una justificación más extensa), debido a que el rendimiento de la fusión depende de la capacidad de los sistemas de complementarse-corregirse unos a otros.

La fusión, en su formulación más simple, es una combinación lineal de las probabilidades estimadas por cada sistema. Supóngase que X es el segmento de voz a evaluar, procesado de forma paralela por N sistemas, y $s_i(X,l)$ la puntuación que recibe la señal para el locutor objetivo l con el sistema i, tal que las puntuaciones están relacionadas positivamente con la probabilidad a posteriori del locutor objetivo, es decir, cuanto mayor es la puntuación, más probable es el locutor objetivo. Bajo estas condiciones, la fusión de sistemas mediante combinación lineal se realiza de la siguiente manera:

$$s_f(X,l) = f(X,l,\mathbf{w}) = w_0 + \sum_{i=1}^{N} \mathbf{w}_i s_i(X,l) \tag{2.41}$$

donde $s_f(X,l)$ es la puntuación global (del sistema fusionado), y $\mathbf{w}$ es un vector de pesos con valores reales. El peso asignado a cada subsistema determina su contribución en la clasificación global. Estos pesos se optimizan sobre un conjunto de desarrollo. Entre las metodologías de optimización propuestas en la bibliografía para la búsqueda de los pesos óptimos en sistemas de reconocimiento de locutor, destacan las siguientes:

- Las redes neuronales [JPCampbell03].
- Las máquinas de vectores soporte [Ferrer06].
- La regresión logística lineal [Brummer07].

De ellas, la más utilizada es la regresión logística, probablemente debido a que Niko Brummer (el autor que propuso la aplicación de regresión logística en la búsqueda de los pesos óptimos) ha desarrollado un software en Matlab que incluye todas las herramientas necesarias para la fusión de sistemas mediante regresión logística. El paquete de software se denomina FoCal (*Fusion and Calibration Toolkit*), y se ofrece bajo una licencia libre [FoCal]. Todos los experimentos de fusión de sistemas llevados a cabo en este trabajo tesis se han realizado mediante FoCal.

El proceso de optimización del vector de pesos mediante regresión logística tiene dos objetivos principales [Brummer07]: (1) mejorar la capacidad de discriminación del sistema, es decir, que la curva DET del sistema fusionado sea mejor que la curva de cualquier subsistema (las curvas y métricas se describen en el apartado 2.6.3.2); y (2) calibrar el sistema debidamente, de forma que las puntuaciones estimadas mediante la fusión se puedan interpretar como probabilidades a posteriori de los locutores y sus rasgos de variación sean coherentes. Al calibrar debidamente el sistema, se puede determinar un umbral teórico para un amplio rango de puntos de operación, donde cada punto corresponde a una aplicación concreta con costes y probabilidades a priori específicos [Brummer06] (véase 2.6.3.2, sección "una métrica independiente de la aplicación"). Por último, en [Brummer07] se defiende que la fusión basada en regresión logística ha mostrado muy buen rendimiento en los mejores

sistemas desarrollados para la NIST SRE de 2006, los cuales presentaban una diferencia muy pequeña entre los costes C_{DET} y C_{DET}^{min}, lo cual indica que estaban debidamente calibrados.

De forma ideal, los parámetros de la fusión se deberían estimar de forma separada para cada locutor. En la práctica no es posible, dado que generalmente se dispone de pocos datos de entrenamiento para los locutores objetivo, y es mejor emplearlos en la estimación de modelos que para optimizar la fusión. En consecuencia, los parámetros de fusión se estiman de forma simultánea para todos los locutores objetivo, empleando un conjunto de desarrollo compuesto por una cantidad elevada de experimentos objetivo y experimentos impostor. Este conjunto de desarrollo debe ser diverso (en cuanto a la fuente acústica de las señales, el género de los locutores, la duración de los segmentos de entrenamiento y test, etc.), de forma que dé cobertura a experimentos reales del conjunto de evaluación.

Con objeto de mejorar el rendimiento de las metodologías clásicas de fusión, algunos autores han propuesto integrar información auxiliar (*side information*) referente a los segmentos a clasificar [Solewicz07][Ferrer08b]. Una aproximación similar fue propuesta en [Garcia-Romero04], mediante la incorporación de métricas de calidad (*Quality Measures*) en la fase de entrenamiento y evaluación de un sistema de reconocimiento de locutores, así como en la fusión de sistemas. Las métricas de calidad son etiquetas informativas que determinan la calidad de una señal con respecto a cualquier factor que influya en el rendimiento del sistema (relación señal/ruido, etc.). En [Solewicz07] se propone la incorporación de ciertos atributos relacionados con la fuente acústica de los segmentos en el proceso de optimización de la fusión. A partir de dichos atributos, se definen conjuntos de desarrollo diferentes para estimar los parámetros de la fusión. Estos conjuntos se optimizan de forma separada, obteniendo parámetros de fusión específicos para cada uno de ellos. Prácticamente en la misma época, Ferrer et.al. [Ferrer08b] proponen una aproximación muy similar, donde de nuevo, en contraste con las aproximaciones más clásicas, la fusión no se aplica de forma uniforme a partir de todo el conjunto de desarrollo, sino que se adapta a ciertas características de la señal de test (como, por ejemplo, si el locutor es nativo o no). La aproximación propuesta en este último trabajo se integra en la fusión mediante regresión lineal, motivo que ha impulsado su empleo en muchos de los sistemas desarrollados para la última NIST SRE de 2008. De hecho, Niko Brummer ha incluido esta aproximación en el software FoCal, definiendo una función para la fusión con información auxiliar adicional (véase [FoCal]).

2.5 Seguimiento de Locutores

En los últimos años, debido a la expansión de la banda ancha y la difusión de contenidos digitales, en particular, recursos de audio y video, ha surgido la necesidad de desarrollar herramientas de búsqueda basadas en el contenido de los recursos. Son recursos que contienen audio de diversas fuentes acústicas: música, anuncios comerciales, etc. (sobre todo en las grabaciones que provienen de televisión o radio). Las herramientas de búsqueda deben identificar qué señales o fragmentos de

señales encajan con un determinado requerimiento. Cuando las grabaciones, además de voz, contienen ruido o música, es necesario pre-procesar el audio, para seleccionar únicamente los fragmentos de voz. Una vez se dispone de fragmentos continuos de habla, éstos suelen dividirse en fragmentos homogéneos internamente. El proceso de división, que habitualmente consiste en identificar los turnos de locutor, es decir fragmentos producidos por un único hablante, se conoce como segmentación de locutores [Dunn00]. Está muy relacionada con la segmentación acústica o detección de cambios acústicos que trata de determinar fragmentos continuos que provienen de una misma fuente acústica (voz/no voz; música/habla; etc.).

En el caso de la segmentación de locutores, una vez determinadas las fronteras, es necesario clasificar los segmentos resultantes, de acuerdo a la identidad del locutor. Se distinguen dos tipos de tareas: diarización y seguimiento de locutores. Ambas tareas responden a la pregunta "*¿Quién habla cuándo?"* ("*who spoke when?"*), pero parten de precondiciones distintas:

- Si previamente no se dispone de datos de los locutores objetivo (ni siquiera se sabe cuántos son), la tarea se denomina diarización de locutores (*speaker diarization*) [Reynolds05]. En este caso, se aplica un algoritmo de clasificación no supervisada, en el que cada segmento se asocia con una clase definida en el mismo procedimiento de diarización. En [Kotti08] [Tranter06] [Anguera06] se presentan las principales aproximaciones de segmentación y agrupamiento aplicadas a la diarización de locutores.
- Si previamente se dispone de datos de los locutores objetivo, la tarea se conoce como seguimiento de locutores (*speaker tracking*) [Martin01]. El sistema debe proporcionar una identificación absoluta para los segmentos detectados. Para ello, previamente se estiman modelos a partir de los datos de los locutores objetivo y se determina si cada segmento detectado corresponde a alguno de ellos o proviene de un locutor desconocido. El seguimiento de locutores se definió por primera vez en la evaluación del NIST de 1999 (1999 NIST *Speaker Recognition Evaluation,* NIST SRE) [Przybocki99]. A partir de 2002, el NIST sustituyó las tareas de seguimiento por las de diarización, debido en gran parte al interés suscitado por el indexado no supervisado de reuniones.

En la bibliografía se distinguen tres campos de aplicación principales para el seguimiento y la diarización de locutores [Reynolds06]:

(1) Noticias difundidas por radio y televisión

(2) Grabaciones de reuniones

(3) Conversaciones telefónicas

Se trata de escenarios en los que, antes de comenzar el procesamiento, el sistema cuenta con todo el recurso de audio a procesar. En consecuencia, la mayoría de las aproximaciones propuestas en la bibliografía son iterativas o multi-fase, es decir, los datos de entrada se procesan de principio a fin de forma cíclica. Este hecho supone que son aproximaciones que no se pueden aplicar en escenarios que requieran una resolución continua (*online*) y en tiempo real, como, por ejemplo, en aplicaciones de monitorización y adaptación continua al locutor (en entornos AmI), o en sistemas de dialogo.

Suponiendo que el usuario hablará durante un tiempo suficiente, es posible emplear sistemas de seguimiento o diarización para ello, reto que plantea la necesidad de abordar la adecuación de dichas aproximaciones a los requerimientos que establecen estas aplicaciones (baja latencia, procesamiento continuo de la señal de entrada).

2.5.1 Metodologías para la segmentación y la clasificación de locutores

Las aproximaciones clásicas de seguimiento y diarización se realizan en dos fases, generalmente desacopladas:

(1) Segmentación acústica, que produce una secuencia de segmentos.
(2) Clasificación de segmentos, procedimiento que consiste en asignarle una etiqueta de locutor a cada segmento detectado en la fase anterior.

Generalmente, como pre-proceso, se incluye también un módulo que detecta tramos de voz y no-voz, de los cuales sólo se conservan los tramos de voz [Tranter06] [Meignier06] [Barras06] [Moraru05] [Istrate05].

En tareas de diarización, la clasificación de segmentos se lleva a cabo mediante un algoritmo de agrupamiento no supervisado, el cual determina qué segmentos son hipotéticamente del mismo locutor [Tranter06] [Meignier06] [Barras06]. En tareas de seguimiento, en cambio, se estiman modelos probabilísticos a partir de los datos de los locutores objetivo, y a cada segmento detectado se le asigna el locutor más probable [Moraru05] [Bonastre00]. En [Collet06] se propone un sistema de seguimiento que, con objeto de aumentar el rendimiento de la detección de locutores, incluye un algoritmo no supervisado de agrupamiento. En [Istrate05] se realiza la comparación de dos sistemas de seguimiento: (1) un sistema que consiste en un módulo de segmentación seguido de otro de detección de locutores; y (2) un sistema que, además de los dos módulos mencionados anteriormente, incluye un modulo adicional de agrupamiento. Los resultados obtenidos en [Istrate05] indican que si una fase (sea de segmentación, agrupamiento o detección) parte de información errónea, su comportamiento es impredecible, y en consecuencia, generalmente perjudica el rendimiento global del sistema de seguimiento. En [Moraru05] se llega a la misma conclusión.

2.5.1.1 Segmentación acústica

Desde una perspectiva global, las metodologías para la segmentación acústica se pueden dividir en dos categorías principales [Kotti08][Anguera06]:

(1) Metodologías basadas en métricas estadísticas
(2) Metodologías basadas en modelos probabilísticos

Las metodologías que comparan los fragmentos de señ3al empleando métricas estadísticas son probablemente las más aplicadas hasta la fecha. Calculan una distancia espectral entre dos segmentos continuos (generalmente solapados) para decidir si provienen o no de la misma fuente. Para determinar si hay un punto de cambio, se estima el máximo/mínimo local de la función de distancia en

el fragmento considerado, valor que se compara con un umbral predeterminado. Entre las aproximaciones a la segmentación acústica mediante métricas estadísticas destacan las siguientes:

- La aproximación denominada *Bayesian Information Criterion* (BIC) es probablemente la aproximación dominante [Chen98]. BIC es un criterio de selección de modelos que, de entre todos los posibles candidatos, trata de identificar el mejor modelo paramétrico para describir los datos de entrada. Obviamente, cuanto mayor es el número de parámetros del modelo, mayor es la probabilidad de los datos con los que se ha estimado el modelo. Pero una cantidad excesiva de parámetros lo puede dar lugar a lo que se conoce como "sobre entrenamiento". Por ello, BIC penaliza la complejidad del modelo, específicamente con respecto a la cantidad de parámetros. En la segmentación, BIC se aplica como contraste de hipótesis: se trata de determinar si los segmentos a comparar se modelan mejor mediante dos distribuciones gaussianas distintas o mediante una única distribución común (que corresponde a la unión de los segmentos).
- Otra aproximación, también muy empleada, es la denominada *Generalized Likelihood Criterion* (GLR) [Delacourt00]. Es un contraste de hipótesis donde: (1) la hipótesis nula, H_0, considera que los dos segmentos a analizar provienen de la misma fuente; y (2) la hipótesis alternativa, H_1, en cambio, considera que cada segmento proviene de una fuente distinta. La distancia entre ambos se representa mediante el LLR de las dos hipótesis, donde, al igual que en la aproximación BIC, los modelos se estiman a partir de los propios datos de los segmentos analizados. En [Delacourt00] la aproximación GLR se emplea como primera fase de la segmentación. El procedimiento finaliza con una segunda fase, basada en la aproximación BIC, que trata de refinar los resultados obtenidos en la primera. Como alternativa a GLR, se analiza el empleo de la distancia *Kullback-Leibler* (KL) en esta segmentación de dos fases. En los experimentos realizados por los autores se observa que el rendimiento de la aproximación GLR es superior al de la aproximación KL.
- En [Ajmera04] los autores proponen una aproximación basada en BIC pero sin penalización, ya que la cantidad de parámetros de los modelos de las dos hipótesis (H_0: no hay frontera, H_1: hay frontera) es, por definición, la misma. Así, sus probabilidades se pueden comparar directamente, y no es necesario estimar un factor de penalización específico para los datos de entrada (factor que penaliza la complejidad de los modelos). La hipótesis H_0 se modela mediante una GMM de 2 componentes. Los dos modelos que forman la hipótesis H_1, consisten en una única gaussiana. Los autores defienden que al estimar modelos directamente comparables, se obtiene un umbral de decisión natural en torno al cero, robusto, el cual no depende de la fuente acústica de los datos. Esta aproximación es equivalente al método GLR con la restricción de igualar la complejidad de los modelos.
- Recientemente, se ha propuesto una aproximación probabilística para la segmentación que también se basa en estimar sendos GMM (específicamente GMM(MAP)) para modelar los segmentos [Malegaonkar07]. El método se basa en comprobar si la probabilidad *a posteriori* del modelo generado con uno de los segmentos, dado el otro segmento, y viceversa, superan un

umbral predeterminado. Esta aproximación es muy similar a la conocida como *Cross Likelihood Ratio* (que se describe en el siguiente apartado), muy empleada para el agrupamiento de locutores.

El segundo grupo de métodos de segmentación acústica, en vez de comparar fragmentos de señal aplicando métricas estadísticas, estima modelos generativos de las clases acústicas (de forma *offline*), a partir de un conjunto de entrenamiento cerrado. Estos modelos representan el canal, el género, la fuente acústica (música/habla/silencio), o sus combinaciones, y se emplean para clasificar los fragmentos continuos, asociando el tramo analizado con la clase que obtiene la máxima probabilidad [Gauvain98] [Kemp00]. Los cambios en el modelo de máxima probabilidad determinan las fronteras. A continuación se describen brevemente algunos trabajos basados en estas ideas:

- Wu et. al. [Wu03a] proponen una aproximación para la detección de cambios acústicos en tiempo real. Estiman un UBM offline que clasifica como habla/habla-dudosa/no-habla los vectores de los dos segmentos consecutivos a analizar. En base a esta información, definen una función que calcula la distancia KL entre los dos segmentos. Para decidir si en el fragmento analizado hay una frontera, la distancia máxima local se compara con un umbral (que se establece de forma dinámica). En un trabajo posterior [Wu03b], los mismos autores añaden una fase de refinamiento: una vez determinada una frontera acústica entre los segmentos *i* e *i*+1 (mediante el algoritmo descrito más arriba), se estima un modelo de locutor ISA (*Incremental Speaker Adaptation*) con los datos del segmento *i* (en caso de que entre los segmentos *i-1* e *i* no haya habido cambio, se actualiza el modelo ISA del segmento *i*-1). Si la probabilidad del modelo ISA dado el segmento i+1 supera el umbral dinámico, se confirma la frontera. En caso contrario, se descarta.
- Kwon y Narayanan presentan una aproximación que también emplea modelos GMM estimados offline [Kwon05]. En una primera fase, se aplica la aproximación GLR para la segmentación. A continuación, en una segunda fase, se refinan las fronteras especificadas en la primera empleando modelos de locutor GMM(MAP) a partir de diferentes modelos generativos (UBM, UGM (*Universal Gender Model*) y SSM (*Sample Speaker Model*)) estimados sobre un conjunto de entrenamiento cerrado. El método comienza estimando un modelo con el primer segmento a analizar y a continuación itera para los demás segmentos del siguiente modo: evalúa la probabilidad del segmento dados los modelos precedentes y en función de que se sobrepase un umbral asume la identidad del locutor que dé máxima probabilidad para el segmento. En caso contrario determina la existencia de una frontera y estima un nuevo modelo de locutor.
- Collet et. al. proponen una aproximación a la segmentación de locutores basada en modelos denominados *anchor models* [Collet06], modelos aplicados en el reconocimiento e indexado de locutores por primera vez en [Sturim01]. El locutor no se representa de forma absoluta, sino de forma relativa a un conjunto de locutores de referencia cuyos modelos se estiman previamente sobre un conjunto de entrenamiento cerrado. En consecuencia, un locutor se caracteriza por un conjunto de vectores de probabilidades de sus pronunciaciones dados los modelos de referencia. Un vector de probabilidades se denomina Vector de Caracterización del Locutor *(Speaker*

Characterization Vector, SCV). Para la segmentación, la aproximación estima la correlación de los SCV de los dos segmentos consecutivos a analizar. Esta misma metodología se emplea tanto para el agrupamiento como para el seguimiento de locutores.

Además de estas dos estrategias de segmentación, basadas en métricas estadísticas y modelos probabilísticos generativos, algunos trabajos utilizan la energía del tramo para llevar a cabo la segmentación [Chen98] [Kemp2000]. El método consiste en detectar silencios, que determinan las fronteras, por medio de decodificadores acústico-fonéticos o mediante la estimación directa de la energía del tramo. Se trata de una aproximación sencilla pero que presenta un rendimiento inferior a las metodologías que emplean métricas estadísticas o modelos generativos estimados *offline*. Las metodologías que emplean modelos generativos tienden a obtener una gran precisión (con algunos errores en la detección de impostores). Las metodologías que aplican métricas estadísticas, en cambio, tienden a obtener muy pocos errores en la detección de impostores (a costa de una menor precisión) [Kotti08] [Kemp2000].

2.5.1.2 Clasificación de los segmentos en sistemas de diarización

Los sistemas de diarización, tras una primera fase de segmentación, y con el objetivo de detectar los segmentos pronunciados por un mismo locutor, proceden a la clasificación no supervisada de los segmentos detectados.

La mayor parte de los sistemas de diarización opera de forma *offline* sobre la señal completa siguiendo un esquema de agrupamiento de segmentos jerárquico. Se trata de un procedimiento iterativo donde los grupos (*clusters*) de segmentos definidos en la iteración anterior se agrupan o dividen, hasta que supuestamente se obtiene la agrupación óptima de segmentos. El diseño de un sistema jerárquico requiere la definición de: (1) una función distancia que determina la similitud acústica de los segmentos; y (2) un criterio de parada que detenga el proceso iterativo al obtener la agrupación óptima.

Si el agrupamiento se realiza de abajo-arriba (*bottom-up clustering*) mediante fusiones sucesivas de grupos similares, partiendo de una configuración inicial en la que cada grupo consta de un único segmento, se llama agrupamiento aglomerativo (*agglomerative clustering*). En cada iteración, se calculan todas las distancias entre los grupos presentes en ese instante, y se agrupan los más cercanos. Las primeras aproximaciones trataban de definir un árbol de segmentos (una matriz de distancias), sobre el que se buscaba la poda óptima para determinar el agrupamiento final [Solomonoff98] [Jin97] [Reynolds98]. En [Siegler97], en cambio, no se estima todo el árbol, sino que se define un umbral de agrupamiento como criterio de parada. Si la distancia mínima de un segmento respecto a los segmentos analizados previamente es inferior al umbral predeterminado, el segmento se agrupa con el de distancia mínima; en caso contrario, se define un nuevo grupo con el segmento actual. En [Chen98] se propone el empleo de la aproximación BIC para el agrupamiento de segmentos, tanto para determinar las distancias (dos segmentos se agruparán únicamente si aumenta la función BIC del

tramo analizado), como para establecer el criterio de parada (el algoritmo finaliza cuando el aumento de la función BIC no supera un umbral predeterminado).

En estas aproximaciones, el segmento más cercano a otro segmento se determina por medio de una función distancia. Entre las funciones más empleadas en la bibliografía destacan:

(1) La distancia simétrica KL (KL2) [Siegler97].
(2) La distancia o probabilidad BIC [Chen98], que es la métrica más aplicada actualmente.
(3) La aproximación GLR [Solomonoff98], descrita en el apartado anterior.
(4) La aproximación denominada *Cross Likelihood Ratio* (CLR) [Reynolds98] que representa la probabilidad de un segmento dado el modelo estimado con el otro segmento (y recíprocamente).

En [Ben04] se propone el empleo de modelos GMM(MAP) estimados a partir de un modelo denominado *Document Speech Background Model* (DSBM) para modelar los segmentos. Los GMM(MAP) se comparan mediante una función de similitud, la cual es una variación de la distancia KL2. En [Moraru05] se emplea la propuesta de [Ben04] para el agrupamiento de locutores en tareas de seguimiento, pero se incluye un criterio de parada basado en la aproximación BIC. Por otro lado, en una aproximación reciente [Barras06], se ha propuesto un procedimiento de dos fases, que de nuevo aplica modelos GMM: en la primera fase, los segmentos detectados se agrupan aplicando la aproximación BIC, tanto para calcular las distancias como para determinar el criterio de parada; en la segunda fase, se vuelven a reagrupar los segmentos definidos en la primera, empleando para ello modelos de locutor GMM(MAP) (en los que además se compensa el canal), cuyas distancias se estiman mediante CLR. En [Le07] se propone una metodología de agrupamiento muy similar, que también aplica modelos GMM(MAP) y la aproximación CLR, pero reemplaza las funciones de probabilidad del CLR por LLR. Por último, destacan las dos aproximaciones de diarización descritas en [Meignier06]. Una de las aproximaciones realiza de forma conjunta la segmentación y el agrupamiento de locutores mediante modelos HMM. La otra aproximación emplea, de nuevo, un procedimiento aglomerativo que aplica GLR para determinar las distancias entre los segmentos, representados mediante GMM(MAP), y define un criterio de parada basado en la aproximación BIC.

Por otro lado, si el agrupamiento se realiza de arriba-abajo (*top-down clustering*), es decir a partir de un único grupo que se divide iterativamente hasta que se cumple un criterio de parada, se dice que es divisivo (*divisive clustering*). Generalmente, esta aproximación presenta peor rendimiento que las aproximaciones aglomerativas, y por tanto, ha recibido menos atención. Entre los métodos de agrupamiento divisivo, destacan los siguientes:

- En [Johnson98] se propone un algoritmo de agrupamiento para el reconocimiento del habla que divide cada segmento activo en cuatro sub-segmentos de forma iterativa siguiendo una estructura de árbol, y agrupa los sub-segmentos similares. El agrupamiento trata de minimizar la distancia AHS (*Arithmetic Harmonic Sphericity*) basada en matrices de correlación; los nodos hijos del segmento activo que obtienen una distancia inferior al umbral de ocupación se agrupan con el

nodo más cercano. Estos procesos de división y agrupamiento se repiten hasta que todos los nodos derivados de un ancestro sobrepasan el umbral predeterminado de mínima ocupación. Llegados a este punto, se realiza la combinación con los nodos hijo de otros ancestros.

- Tranter y Reynolds [Tranter04] emplean prácticamente el mismo algoritmo para la diarización de locutores. Dividen cada segmento activo en cuatro sub-segmentos y agrupan los segmentos que presentan una distancia AHS inferior a la distancia a su propio nodo ancestro. Estos procedimientos de división y agrupamiento se repiten hasta que se alcanza una máxima cantidad de iteraciones o hasta que no se pueden realizar más divisiones y agrupamientos. Como alternativa a estos criterios de parada, se proponen otras dos metodologías: una de ellas emplea funciones de coste en la división; la otra, en cambio, emplea la aproximación BIC para las divisiones (el agrupamiento se realiza, de nuevo, analizando las distancias AHS).

En la bibliografía se observa que la aproximación dominante para el agrupamiento *offline* es la aglomerativa con un criterio de parada basado en la aproximación BIC.

A diferencia de los sistemas de agrupamiento *offline*, los sistemas que realizan el procesamiento de los datos de entrada *online* (de forma continua), sólo conocen los datos recogidos hasta el instante en que se realiza el agrupamiento. En ocasiones, con objeto de dar mayor robustez a la comparación de segmentos, se podrá permitir una cierta latencia en su respuesta. Pero en cualquier caso, deben asumir que no se conocen íntegramente los datos a procesar en cada momento de la clasificación; no pueden aplicar un procedimiento iterativo, ni estimar las distancias entre todos los segmentos para agrupar los más cercanos, ni tampoco se pueden comparar diferentes estrategias que determinen el agrupamiento óptimo (estrategias que determinan la cantidad óptima de grupos). Generalmente, son aproximaciones que parten de un conjunto de locutores compuesto por un único locutor (el primer locutor que habla en la grabación), y a medida que intervienen (y se detectan) más locutores en la grabación, se definen nuevos locutores en el conjunto. Tal como se ha mencionado anteriormente, los sistemas continuos han recibido poca atención en la bibliografía. Se han encontrado los siguientes trabajos:

- En [Rougui06] se propone un sistema de agrupamiento continuo que emplea modelos GMM comparados mediante la distancia KL. La segmentación de locutores se realiza también de forma continua, por lo que una vez establecida una frontera entre locutores, se calcula la distancia del GMM estimado con el segmento anterior a dicha frontera y los modelos de la base de datos. La distancia mínima se compara con un umbral dinámico para determinar si el segmento se va agrupar con el modelo correspondiente o se va a crear un nuevo modelo de locutor en la base de datos. Los autores hacen hincapié en la eficiencia de la metodología, la cual emplea un árbol de decisión para comparar los modelos, estructura que permite la reducción del coste computacional de la clasificación.
- Liu y Kubala [Liu04] proponen un conjunto de tres algoritmos para el agrupamiento de locutores continuo, basados en la adaptación de una aproximación anterior [Jin97]. Las aproximaciones propuestas requieren que los datos de entrada se hayan segmentado previamente. En este sentido,

los segmentos detectados se comparan mediante la distancia GLR y se emplea el criterio de dispersión propuesto en [Jin97] para el agrupamiento continuo, de forma que la salida del módulo de procesamiento de un segmento determine su clase o grupo definitivo.

- Mori y Nakagawa [Mori01] proponen una aproximación de agrupamiento para la adaptación al locutor de los modelos acústicos, que se puede aplicar de forma continua. Esta aproximación emplea una métrica de distorsión de cuantificación vectorial definida por Nakawaga et. al en un trabajo previo [Nakawaga93]. Comienza estimando un codebook para el primer locutor, y a continuación añade un nuevo codebook cada vez que se detecta un nuevo locutor. Para ello, se analiza la distorsión mínima de cada segmento con respecto a los codebooks de la base de datos. Si la distorsión es superior a un umbral predeterminado, se confirma que el segmento corresponde a un nuevo locutor. En caso contrario, el segmento se agrupa con el locutor de mínima distorsión.

2.5.2 Aproximaciones al seguimiento de locutores continuo

La mayoría de las aproximaciones de seguimiento de la bibliografía operan de forma *offline*, es decir, se aplican sobre un recurso de audio previamente grabado, el cual se conoce íntegramente desde el momento en que se inicia el procesamiento. Esto es debido a que los principales escenarios de aplicación (noticias, reuniones, conversaciones telefónicas, etc.) requieren el seguimiento o la diarización para el indexado de recursos de audio previamente grabados.

Generalmente, las aproximaciones de seguimiento constan de dos fases desacopladas: la segmentación y la detección de los locutores objetivo [Bonastre00][Collet05]. Estas fases se aplican de forma secuencial: primero se segmenta todo el recurso de audio a procesar, y a continuación, se realiza la detección de locutores sobre los segmentos definidos en la fase anterior. Como se ha dicho, algunos trabajos añaden una fase de detección de voz/no-voz antes de la segmentación [Moraru05] [Istrate05].

En contraste con estas aproximaciones, Magrin-Chagnolleau et. al. [Magrin-Chagnolleau99] proponen un algoritmo para la detección de uno o dos locutores objetivo en programas de televisión. Esta aproximación realiza las fases de segmentación y detección de forma conjunta. Los autores proponen un algoritmo complejo basado en cuatro umbrales y una función de suavizado de ventanas, para determinar el inicio y el final de cada segmento, descartando los segmentos de longitud inferior a 2.5 segundos. La metodología de segmentación y detección estima el LLR de cada tramo mediante modelos de locutor y un modelo de referencia. En [Dunn00] se propone un algoritmo prácticamente idéntico para el seguimiento de un locutor en señales telefónicas, donde, de nuevo, la segmentación y la detección se realizan de forma conjunta mediante LLR y modelos GMM. Con objeto de compensar el ruido del canal y de otras fuentes de distorsión, el LLR de cada tramo se suaviza mediante un filtro triangular de 2.5 segundos. Para determinar el inicio y el final de cada segmento, los autores aplican un algoritmo muy parecido al de [Magrin-Chagnolleau99]. De forma adicional, los mismos autores analizan el rendimiento que obtiene el sistema al definir segmentos de longitud fija de 2.5 segundos. Estos segmentos se solapan cada 0.25 segundos, por lo que el sistema emite una decisión de detección

del locutor cada 0.25 segundos. En experimentos realizados sobre el corpus de evaluación del NIST SRE 1999, se observa que el procedimiento basado en intervalos fijos presenta una tasa de reconocimiento mejor que la de la primera aproximación.

En la bibliografía se encuentran muy pocas aproximaciones de seguimiento y diarización orientadas a su ejecución continua (*online*) o en tiempo real:

- En [Lu05] se propone un algoritmo no supervisado de segmentación y seguimiento de locutores en tiempo real, en el que la entrada de audio se procesa una única vez. El sistema supone que la entrada contiene únicamente fragmentos hablados. Estos fragmentos se dividen en segmentos de 3 segundos, solapados a 2.5 segundos. La segmentación de locutores consta de dos fases. La primera fase consiste en comparar dos segmentos adyacentes mediante la distancia KL del modelo GMM generado a partir de cada segmento: si la distancia supera un umbral predeterminado se considera que puede haber un cambio; en caso contrario, se agrupan los dos segmentos para estimar un modelo más robusto del locutor de ese instante. El posible punto de cambio, se analiza mediante una fase de refinamiento que estima el contraste de las dos hipótesis para confirmar o rechazar el cambio. Una vez confirmado un cambio de locutor, el modelo del segmento actual se compara con todos los modelos de la base de datos para determinar el más cercano: aquél de menor distancia KL para una cierta cantidad limitada de componentes gaussianas. Si la distancia es inferior a un umbral el modelo correspondiente se adapta; en caso contrario, se considera como nuevo locutor y se introduce su GMM en la base de datos. De este modo los modelos de locutor son GMM estimados de forma incremental añadiendo datos a medida que se avanza en la señal. La cantidad de componentes del modelo depende de la cantidad de datos disponible para su estimación. Para que los modelos se puedan estimar en tiempo real, se emplea una aproximación de modelado incremental *quasi*-GMM.
- En [Liu05] se propone una metodología de adaptación al locutor de los modelos acústicos que es aplicable de forma indefinida y en tiempo real, y que trata de mejorar el rendimiento de los reconocedores de habla. La propuesta supone que el conjunto de locutores es desconocido. El sistema de seguimiento comienza con la ejecución de un módulo de segmentación de locutores que combina modelos fonotácticos empleando la aproximación GLR. De forma paralela, se aplica un algoritmo de agrupamiento que determina de forma continua la clase de cada segmento, en el mismo instante en el que se define el segmento (sin tener que esperar a segmentos futuros). Una vez etiquetados los segmentos detectados, se procede a la adaptación online de los modelos acústicos. La adaptación consiste en una transformación lineal del espacio de parámetros, con una matriz de transformación que se adapta de forma incremental por ML.
- Markov y Nakamura [Markov07] proponen un sistema de diarización continuo con una latencia inferior a 3 segundos. Es un sistema catalogado como sistema de aprendizaje *never-ending*, dado que es continuo y autosuficiente. El sistema está compuesto por varios módulos que comparten un mismo conjunto de GMM. El proceso se inicia con tres GMM que representan el silencio,

locutores de género masculino y locutores de género femenino. Para determinar los segmentos de voz se aplica un filtro por mediana a las probabilidades de los vectores de entrada, dados estos tres modelos. Los autores suponen que los cambios de locutor se dan en los puntos de silencio. Por tanto, a diferencia de las aproximaciones clásicas de diarización, este sistema no incluye un módulo de segmentación de locutores. Una vez determinado un segmento de voz, el sistema asigna al segmento el género de máxima probabilidad, y a continuación decide mediante un contraste de hipótesis (específicamente el LLR) si corresponde o no a un locutor conocido. Si se determina que el segmento corresponde a un locutor de la base de datos, el sistema re-adapta el GMM del locutor. En caso contrario, el sistema define un nuevo GMM a partir del GMM de género más probable. La adaptación de modelos se realiza mediante una variación del algoritmo EM. Para que el sistema pueda aplicarse en largos periodos de tiempo, se define un contador para cada GMM de la base de datos; si el contador indica que no se ha empleado en un largo periodo de tiempo, el modelo se elimina.

Existen otras aproximaciones de seguimiento, segmentación o agrupamiento de locutores, que aunque no se hayan diseñado para su ejecución continua o en tiempo real, presentan características adecuadas para ello. Es el caso de la aproximación de seguimiento de locutores propuesta por Collet et. al. en [Collet05] [Collet06]. Ésta se basa en los denominados *anchor models*, que emplean SVC. Para la segmentación, se estima la correlación entre los SVC de dos segmentos consecutivos, cuyo coste computacional no es muy elevado (tan sólo requiere la estimación de la covarianza entre dos vectores, y la desviación estándar de cada segmento). En [Collet05] se emplea el mismo procedimiento para la detección de locutores en un sistema de seguimiento: los segmentos previamente definidos mediante la aproximación GLR se asocian con el modelo de máxima correlación. El módulo de detección de locutores de [Collet06], en cambio, es probabilístico, pero también aplicable de forma continua. Consiste simplemente en estimar la probabilidad del SVC (estimado a partir del propio segmento) dados los modelos de locutor y los modelos de referencia (estimados offline). De todas formas, entre las fases de segmentación y detección, los autores definen un método de agrupamiento iterativo, que requiere el procesamiento del recurso de audio completo, por lo que no se puede aplicar *online*.

En cuanto a los métodos de segmentación acústica que se han descrito en el apartado anterior (BIC, GLR, CLR), hay que destacar que son aproximaciones robustas pero muy costosas computacionalmente. Por un lado, requieren la estimación de dos o tres modelos de gaussianas para evaluar cada potencial frontera del fragmento analizado, y por otro lado, se suele analizar una cantidad elevada de segmentos por segundo. Para una aplicación online parecen más apropiadas las aproximaciones de segmentación que emplean modelos generativos estimados offline. Merecen especial atención las propuestas [Wua03a] y [Wua03b], diseñadas para su ejecución continua y en tiempo real. La fase de detección de [Kwon05] es también una aproximación aprovechable en un sistema continuo. Consiste en la adaptación, mediante el algoritmo MAP, de modelos de locutor

genéricos estimados a partir de un UBM. Aunque el algoritmo no es iterativo y los cambios de locutor se pueden detectar *online*, hay que tener en cuenta que la adaptación MAP es computacionalmente costosa y, por tanto, inadecuada para sistemas que requieran una latencia baja. Por último, entre los métodos de agrupamiento, se han desarrollado también aproximaciones continuas [Rougui06] [Liu04] [Mori01]. En cuanto al resto de aproximaciones, tanto las aglomerativas como las divisivas, a excepción de la propuesta de [Siegler97], son iterativas y requieren el conocimiento íntegro del recurso de audio.

2.6 Bases de datos, métricas de rendimiento y campañas de evaluación

En este apartado se analizan los recursos y procedimientos necesarios para evaluar los sistemas de identificación, verificación y seguimiento de locutores, estudiados y desarrollados en este trabajo. El avance sostenido de las tecnologías de reconocimiento de locutores ha sido posible gracias a que a lo largo de los últimos 20 años se han creado las infraestructuras necesarias para su evaluación: bases de datos, métricas, etc. Han tenido especial importancia las campañas de evaluación de sistemas organizadas periódicamente por el NIST.

2.6.1 Bases de datos

La creación y liberación de bases de datos, diseñadas para el desarrollo y evaluación de sistemas en una tarea específica, es un elemento clave para el avance de la tecnología, puesto que establece un objetivo tecnológico y un marco de desarrollo comunes a toda la comunidad científica, y permite la comparación del rendimiento de las aproximaciones propuestas por diferentes investigadores. Así pues, es importante disponer de bases de datos que den cobertura a todos los aspectos relacionados con la tarea a resolver.

En los dos siguientes sub-apartados se describen las bases de datos de referencia en relación con las tareas de (1) identificación y verificación de locutores y (2) seguimiento de locutores. También se describen las bases de datos empleadas en las fases experimentales de este trabajo:

- Albayzín y Dihana, para el estudio de la reducción de dimensionalidad en tareas de identificación de locutores en conjuntos cerrados (Capítulo 3).
- Albayzín, para analizar el rendimiento de los sistemas de verificación e identificación sobre conjuntos abiertos de locutores, incluyendo las situaciones en las que la fuente acústica de la pronunciación a evaluar no se ha observado en entrenamiento (Capítulo 4).
- NIST SRE, para analizar, en tareas de verificación de locutores, el rendimiento de diferentes metodologías de modelado, diversas estrategias de normalización de probabilidades, y la fusión y calibración de sistemas (Capítulo 4).

- AMI, para el desarrollo y la evaluación del sistema de seguimiento de locutores continuo que se ha propuesto en esta tesis (Capítulo 5).

2.6.1.1 Identificación y verificación de locutores

En [JPCampbell99] [Godfrey94] se presentan dos listados similares de bases de datos creadas a principios-mediados de los noventa para tareas de identificación y verificación de locutores. La mayoría de ellas fueron generadas en Estados Unidos y se incluyen en el catálogo del LDC (*Linguistic and Data Consortium*) [LDC]. Asimismo, la asociación europea ELRA (*European Language Resources Association*) [ELRA], dedicada a producir, publicar y distribuir recursos orientados al procesamiento del lenguaje, principalmente de habla no inglesa, proporciona también un amplio catálogo de bases de datos relacionadas con las tecnologías del habla.

A modo de resumen, a continuación se describen las bases de datos de referencia en relación con las tareas de identificación y verificación de locutores, algunas de ellas distribuidas por el LDC (*KING*, *YOHO*, *TIMIT*, *SWITCHBOARD*, *MIXER*, *NIST SRE*), y otras por ELRA (*POLYVAR*, *SIVA*, *POLYCOST*, *ORIENTEL*). Se describen también bases de datos generadas en el ámbito estatal que, directa o indirectamente, dan cobertura a dichas tareas (*ALBAYZÍN*, *DIHANA*):

- *KING*: El corpus KING, grabado en 1987 en el instituto técnico ITT, fue publicado por el LDC a mediados de los noventa. Participan en las grabaciones 51 locutores masculinos, con señales grabadas por un canal telefónico y un micrófono de gran calidad. Por cada locutor y canal, se incluyen 10 señales cuya duración varía de 30 a 60 segundos. El periodo transcurrido de la grabación de una sesión a la grabación de la siguiente es superior a una semana. La base de datos está diseñada para tareas de identificación o verificación de locutores (independientes del texto) en conjuntos cerrados.
- *YOHO*: Al igual que el corpus KING, YOHO es un corpus grabado en el instituto técnico ITT en 1989 y se puede encontrar en el catálogo del LDC. A diferencia de KING, está diseñada para tareas de verificación de locutores dependiente del texto. El corpus está compuesto por 186 locutores (156 hombres, 30 mujeres) que aportan de 4 a 13 sesiones grabadas en una oficina mediante un terminal de teléfono. En cada sesión, el locutor pronuncia una serie de frases, donde cada frase es una secuencia de tres números de dos dígitos. El intervalo de tiempo transcurrido entre las sesiones es de unos 3 días.
- *TIMIT*: El corpus TIMIT, distribuido por el LDC, es una base de datos diseñada para la modelización acústico-fonética en sistemas de reconocimiento del habla. Dado que fue uno de los primeros corpus en incluir una cantidad elevada de locutores, es muy empleado en el ámbito del reconocimiento de locutores. Está compuesto por 630 locutores (430 hombres, 192 mujeres), grabados en una única sesión y en condiciones de laboratorio.
- *SWITCHBOARD*: El corpus SWITCHBOARD, también distribuido por el LDC, contiene varias particiones: (1) *Switchboard I*; (2) *Switchboard II*; y (3) *Switchboard Cellular*. Las dos primeras, *Switchboad I y II*, están compuestas por conversaciones telefónicas en inglés, de 5 minutos de

duración, grabadas por la línea regular de teléfono mediante diferentes terminales. Cada una de estas particiones está compuesta por unos 500-600 locutores, todos ellos de Estados Unidos y equilibrados en género (en términos generales, la mitad son hombres y la otra mitad mujeres). Las particiones difieren principalmente en las características demográficas de sus participantes. En la segunda partición, *Switchboard II,* los locutores son, en su mayoría, estudiantes del centro-oeste de Estados Unidos y por ello son más jóvenes en promedio. En *Switchboard I* los locutores provienen del noreste de Estados Unidos. Por otra parte, en *Switchboard II* se daba la opción de no seguir el tema de conversación establecido en la fase de diseño de la base de datos. *Switchboard Cellular*, la tercera partición, es de características similares; la principal diferencia es que las conversaciones se han grabado a través de teléfonos móviles y bajo diferentes condiciones acústicas (en cuanto al ruido del entorno y calidad del canal). La cantidad de locutores de esta última partición es ligeramente inferior, son unos 400 locutores, también equilibrados en género.

- *MIXER*: MIXER es un corpus grabado recientemente por el LDC [Cieri06] [JPCampbell04], y empleado en las últimas evaluaciones del NIST de reconocimiento de locutores. El corpus está compuesto por conversaciones telefónicas sobre un tema predeterminado, de 6 minutos de duración, y contiene unos 2500 locutores. Incluye varias conversaciones de un mismo locutor (puede incluir hasta treinta), generalmente en más de un idioma, por lo que la base de datos recoge una gran cantidad de locutores bilingües. De esta forma, el corpus permite analizar la dependencia de los sistemas de reconocimiento de locutores al idioma utilizado en las conversaciones. Por otro lado, ara estudiar la influencia del canal de transmisión, así como del tipo de terminal de teléfono, se solicita al locutor que, en cada conversación, indique cual es el canal de transmisión (móvil, inalámbrico o regular) y cual el tipo de terminal empleado (*head-mounted* (de cabeza), terminal estándar, *ear-bud* (auricular), teléfono personal). Además, los participantes aportan información sobre su país natal, su edad, y su situación profesional. Recientemente, se han grabado dos particiones adicionales sobre nuevos escenarios. La primera, denominada *Mixer4*, consiste en un conjunto de conversaciones grabado de forma simultánea por un canal telefónico y varios tipos de micrófonos ubicados en las salas donde se encuentran los participantes. En la grabación de *Mixer4* participan 400 locutores, procedentes de Estados Unidos. El conjunto de conversaciones reúne, por cada locutor, 10 llamadas de 10 minutos. La segunda, denominada *Mixer5*, además de conversaciones telefónicas, plantea un nuevo escenario en el que el locutor es entrevistado cara a cara. La grabación de la entrevista se realiza mediante diversos micrófonos en la misma sala. Esta nueva partición (empleada en la última evaluación del NIST) está compuesta por 300 locutores que aportan, cada uno, 10 conversaciones telefónicas y 6 sesiones de entrevistas de media hora de duración. Las entrevistas se graban en tres días, y con un intervalo mínimo de media hora entre las sesiones grabadas en un mismo día. En [Brandschain08] se ofrecen más detalles sobre estas dos últimas particiones, incluyendo la descripción de los canales de micrófono empleados.

- *NIST SRE*: En esta última década, debido a la relevancia que han adquirido las evaluaciones NIST SRE organizadas por el NIST, las bases de datos de referencia son las definidas para estas evaluaciones. En cada evaluación se plantean listados concretos de experimentos sobre un conjunto de datos específico definido para la propia evaluación, a partir de bases de datos estándares generalmente proporcionadas por el LDC. En este sentido, en [Przybocki04] se ofrece el listado de las bases de datos que se han empleado para definir las bases de datos de las NIST SRE hasta el año 2004. Entre 1996 y 2003 todas ellas están compuestas por conversaciones telefónicas adquiridas principalmente de diferentes particiones del corpus Switchboard, descrito más arriba. Las bases de datos de las últimas NIST SRE, concretamente, de los años 2004, 2005, 2006 y 2008, se han creado a partir del corpus *Mixer* también descrito más arriba, que incluye conversaciones telefónicas y conversaciones grabadas a través de diversos tipos de micrófonos. Generalmente, se incluyen entre 500-1000 locutores en cada corpus de evaluación. En la evaluación del 2008, la cantidad de locutores ha aumentado considerablemente: el corpus diseñado para la tarea principal está compuesto por alrededor de 3000 locutores.
- *POLYVAR:* POLYVAR es un corpus orientado a la verificación de locutores. Contiene habla leída y espontánea de 143 locutores franceses, tanto nativos como no-nativos.
- *SIVA:* También orientado a la verificación e identificación de locutores, el corpus SIVA está compuesto por 34 locutores objetivo y 402 impostores, todos ellos italianos.
- *POLYCOST:* Orientado a tareas de verificación de locutores, incluye 133 locutores de 13 países de Europa, cuyo idioma nativo no es el inglés. El corpus contiene, para cada locutor, algunas grabaciones en inglés y otras en su idioma nativo.
- *ORIENTEL*: Publicado por ELRA en 2004, reúne señales de 1700 locutores turcos, con frases leídas y conversaciones espontáneas grabadas por canal telefónico.
- *ALBAYZÍN*: Albayzín es una base de datos en castellano fonéticamente equilibrada y grabada a 16 kHz en condiciones de laboratorio [Casacuberta91]. Albayzín fue diseñada para el entrenamiento de modelos acústicos en el ámbito del reconocimiento del habla. La base de datos está compuesta por 204 locutores, cada uno de ellos con un mínimo de 25 señales. La longitud media de una señal es de 3.55 segundos. El corpus fue diseñado y grabado por seis grupos de investigación, interesados en las tecnologías del habla, de diferentes universidades de España.
- *DIHANA*: Dihana es un corpus en castellano de habla espontánea, orientado a una tarea específica de diálogo, y grabada a 8 kHz por canal telefónico [Alcocer04]. Contiene 900 diálogos hombre-máquina de 225 locutores, grabados siguiendo la conocida estrategia del *Mago de Oz:* una persona (el mago), ayudado por un sistema de adquisición, simula el comportamiento de un sistema automático de diálogo, ayuda al usuario a obtener respuestas a sus consultas y trata de conducir la interacción en base a una estrategia dada. Con objeto de ampliar la cobertura acústica, el corpus incluye, por cada locutor, 16 frases leídas adicionales. En consecuencia, de cada locutor se tienen 4 diálogos y un conjunto de 16 frases leídas.

2.6.1.2 Seguimiento de locutores

Generalmente, dada la naturaleza del problema, las tareas de seguimiento o de diarización de locutores se realizan sobre los mismos corpus definidos para la transcripción automática de conversaciones telefónicas, noticias o programas de radio y televisión, y reuniones. Ahora bien, para ser aplicables a esta tarea, deben estar segmentados y anotados con la identidad del locutor, de forma que se conozcan las fronteras y etiquetas de los segmentos. Cabe destacar la transcripción de reuniones como uno de los escenarios principales de aplicación, debido al interés que ha suscitado entre los grupos de investigación que trabajan en tecnologías del habla. Probablemente, todo ello es debido al potencial de aplicación que ofrecen las transcripciones automáticas de reuniones en el mercado, a las numerosas posibilidades de investigación que proporcionan las reuniones (habla espontánea, solapamientos, canales variables), y por último, a la estrecha relación que presentan los escenarios en los que se desarrollan estas reuniones con el estudio de formas de interacción no intrusivas entre personas y máquinas y la valorización de dichas interacciones.

En Estados Unidos la actividad investigadora en torno a la transcripción de reuniones comenzó alrededor de 1999. Hasta entonces, la grabación de datos en el ámbito de los sistemas de reconocimiento del habla se había concentrado en noticias de televisión y conversaciones telefónicas. A partir de entonces, se nota una migración, cada vez más clara, hacia la transcripción de reuniones, con la grabación de varias bases de datos. Entre ellas, se encuentran:

- El corpus de reuniones del ICSI (*The ICSI Meeting Corpus*) [ICSI], compuesto por 75 reuniones que proporcionan 72 horas de duración en total. Las reuniones se grabaron en una única sala, equipada con 6 micrófonos.
- El *Meeting Corpus* [MeetingRoom] de la CMU (*Carnegie Mellon University*) , que contiene grabaciones de 104 reuniones de una duración media de 60 minutos y 6.4 participantes (en promedio). Cada reunión se centra en un tema diferente.
- El corpus piloto de reuniones del NIST (*NIST Pilot Meeting Corpus*) [NIST Meeting], con 19 reuniones y una duración total de 15 horas. Las grabaciones se realizaron en una única sala equipada con 4 micrófonos.

Cabe destacar que a partir de 2002 las evaluaciones del NIST orientadas a la transcripción automática del habla incluyen un dominio específico de reuniones, creado a partir de las tres bases de datos anteriores: ICSI, CMU y el corpus piloto de reuniones del NIST. Más aún, a partir del 2004, todas las evaluaciones de sistemas de transcripción enriquecida del habla (NIST *Rich Transcription Evaluation*, RTE) tratan ya exclusivamente con reuniones, utilizando datos procedentes de ICSI, CMU y NIST, y de dos proyectos europeos AMI y CHIL que se describen a continuación:

- *AMI*: AMI es un consorcio europeo formado en 2003 e integrado por 5 universidades, 5 centros de investigación y 2 empresas de Europa, Estados Unidos y Australia [AMI]. Hasta la fecha, el consorcio ha llevado a cabo dos proyectos de investigación y numerosas actividades de formación, difusión e impulso de iniciativas empresariales. Estos dos proyectos son AMI (*Augmented Multi-*

party Interaction), entre 2004 y 2007, y AMIDA (*Augmented Multi-party Interaction with Distance Access*), entre 2006 y 2009. El objetivo de estos proyectos es, por un lado, desarrollar herramientas para explorar contenidos de reuniones o asistir a reuniones virtuales, y por otro, desarrollar las tecnologías básicas necesarias, en temas tan diversos como modelado de la interacción entre personas, reconocimiento del habla, visión por ordenador, indexado y recuperación de información multimedia, etc. Un objetivo secundario, pero muy importante para la comunidad científica, es la grabación, anotación, gestión y difusión de una gran base de datos multimodal de reuniones, denominada también AMI [AMI Corpus], con varias decenas de gigabytes de audio, video y anotaciones XML (*eXtensible Markup Language*). El corpus de reuniones AMI (de acceso público, bajo una licencia libre) es un conjunto de datos multimodal con interacciones humanas en tiempo real, en el contexto de reuniones realizadas en un entorno inteligente. Los datos se han grabado en tres salas equipadas con diversos micrófonos y cámaras, que recogen señales sincronizadas de audio y video. Las reuniones se llevan a cabo en inglés, con locutores que generalmente no son nativos.

- CHIL: El proyecto CHIL (*Computers in the Human Interaction Loop)* [CHIL], desarrollado entre 2004 y 2007, trataba de desarrollar servicios informáticos que facilitaran la interacción entre personas, y de personas con ordenadores, haciendo que éstos actuaran de forma casi transparente. El objetivo del proyecto es la definición de servicios que se puedan aplicar en cualquier situación que implique a varias personas que están frente a frente, intercambiando información, y colaborando en la resolución de problemas. Plantea el empleo de diversos medios: habla, escritura, gestos, posturas, etc. Para ello, el proyecto preparó varios entornos o espacios inteligentes equipados con sensores audio-visuales. A partir de los datos recopilados con estos sensores, se pretendía detectar, clasificar y comprender las actividades humanas desarrolladas en dichos espacios. Es decir, desarrollar tecnologías de percepción que procesan la información sensorial y posibilitan el seguimiento de los participantes, su identificación, así como el procesamiento del habla. En el marco del proyecto, con objeto de evaluar las tecnologías desarrolladas, se incluía la grabación de un corpus de estas características [Mostefa07]. El corpus se compone de 86 clases teóricas y reuniones grabadas en cinco salas diferentes, equipadas con diversos micrófonos y cámaras de video. En [Mostefa07] se describen los recursos, herramientas y metodologías del proceso de grabación. En el proyecto, coordinado conjuntamente por la Universität Karlsruhe y el Fraunhofer Institute, han participado 15 universidades, centros de investigación y empresas de Estados Unidos y Europa.

2.6.2 Evaluaciones

2.6.2.1 Evaluaciones organizadas por el NIST

El instituto NIST, fundado en 1991, es una agencia federal del Departamento de Comercio de Estados Unidos. Su misión se centra en promover la innovación tecnológica y científica de la industria

de aquel país y del resto del mundo. Para ello, el NIST cuenta con varios grupos dedicados a diferentes áreas de investigación. Entre éstos se encuentra el grupo dedicado a las tecnologías del habla que, con objeto de evaluar el estado de arte de las principales tecnologías, organiza diversas evaluaciones anuales y bianuales.

Entre ellas se encuentran las de reconocimiento de locutores, denominadas de forma abreviada como NIST SRE [NIST SRE][Martin07], que ha coordinado el NIST desde 1996. Las evaluaciones están abiertas a los grupos de investigación y entidades interesadas en el reconocimiento de locutores independiente del texto. Las últimas evaluaciones se han celebrado en 2008 y 2010. Después de organizar una evaluación por año entre 1996 y 2006, el NIST decidió realizar un paréntesis en 2007, debido al crecimiento del número de participantes de las últimas evaluaciones, tiempo que aprovechó para analizar de forma más exhaustiva el conjunto de datos a definir. A partir de entonces las evaluaciones son bianuales. Después de cada evaluación se celebra un *Workshop* en el que todos los participantes contrastan y discuten sus propuestas.

La principal tarea de estas evaluaciones es la detección (verificación) del locutor. El sistema parte de un conjunto de pronunciaciones, tales que en cada pronunciación de entrenamiento habla únicamente un locutor objetivo. A partir de esta información, el sistema debe tomar una decisión firme para cada pronunciación de test: (1) verdadero, si estima que en la pronunciación de test habla el locutor a detectar; (2) falso, en caso contrario. Junto a la decisión, se debe suministrar una tasa, la probabilidad de la hipótesis de que la pronunciación de test provenga del locutor objetivo considerado.

En cada evaluación se definen varias tareas en función de la duración de las pronunciaciones de entrenamiento y test, así como del escenario planteado. Generalmente, en las tareas definidas, la longitud las pronunciaciones de entrenamiento suele ser igual o superior a las pronunciaciones de test. Para cada tarea definida, se determina un listado de experimentos. Cada experimento consiste en verificar una pronunciación de test frente a un locutor objetivo. Se supone que cada experimento es independiente del resto: el único locutor que conoce el sistema de detección es el locutor objetivo considerado.

El NIST define un corpus para cada evaluación, generalmente a partir de conversaciones telefónicas. La definición de los conjuntos de datos ha seguido la evolución de los canales de telefonía pública:

- En el periodo 1996-2001 las bases de datos se definen a partir de conversaciones grabadas por línea regular de teléfono [Doddington00]. Los micrófonos empleados en las terminales son de carbono o electret, por lo que estas evaluaciones tratan de analizar la influencia del tipo de micrófono en el rendimiento de los sistemas.
- En 2001, se introducen señales grabadas con teléfonos móviles (señales de la base de datos *Switchboard Cellular*). Este cambio es debido a que, en esa época, comenzaron a proliferar los teléfonos móviles y a desaparecer los micrófonos de carbono. En las evaluaciones del 2002 y 2003

se sigue con la misma tendencia, ya que ambas evaluaciones emplean casi únicamente señales de telefonía móvil [Przybocki04].

- En el periodo 2004-2006 [Przybocki07], las bases de datos para las NIST SRE se definen a partir de señales de la base de datos *Mixer*. Este diseño trata de analizar la influencia que ejerce el canal de grabación y el idioma de las pronunciaciones en el rendimiento de los sistemas de detección

El escenario más común en las evaluaciones NIST SRE es la detección de un locutor en una señal extraída de uno de los canales de una conversación telefónica. Sin embargo, en la mayoría de evaluaciones NIST SRE se han definido tareas adicionales basadas en escenarios multi-locutor, como la detección de locutores en señales que resultan de la suma de ambos canales de la conversación, el seguimiento de locutores, etc. [NIST]. En concreto:

(1) La detección de locutores en señales obtenidas de la suma de ambos canales de una conversación se introduce como tarea en la NIST SRE de 1999. Es prácticamente igual a la tarea principal (detectar si un locutor objetivo aparece en una conversación de test), con la salvedad de que la pronunciación de test consiste en la suma de los dos canales de la conversación. Esta tarea se mantiene como tal hasta la evaluación de 2001. En la NIST SRE de 2002 se introduce una nueva modalidad, mantenida, de momento, hasta la última evaluación en 2010. En este caso, se proporcionan tres pronunciaciones de entrenamiento, obtenidas de la suma de ambos canales de tres conversaciones de un mismo locutor objetivo (el otro locutor es diferentes en las tres conversaciones). El sistema debe determinar si el locutor objetivo interviene en la pronunciación de test, el cual también se obtiene de la suma de los dos canales de una conversación.

(2) El seguimiento de locutores se define por primera vez como tarea en la NIST SRE de 1999, y se mantiene hasta la NIST SRE de 2001. La tarea consiste en determinar los intervalos de tiempo en los que habla un locutor objetivo en la señal de test. El modelo del locutor a seguir se entrena con una señal compuesta por uno de los dos canales de una conversación. El conjunto de test se compone de señales que contienen la suma de los dos canales de una conversación. Por tanto, partiendo de una pronunciación de entrenamiento (correspondiente al locutor objetivo) y otra de test, el sistema de seguimiento debe determinar los intervalos de tiempo de la conversación de test que corresponden al locutor objetivo. En 2002 la tarea de seguimiento fue sustituida por una tarea de diarización, que aborda un campo de investigación más amplio y un dominio de aplicación más extenso.

(3) La diarización de locutores se introduce como tarea en la NIST SRE de 2000 y se mantiene en las dos siguientes evaluaciones: las NIST SRE de 2001 y 2002. En las dos primeras evaluaciones (2000 y 2001) se distinguen dos tareas: la diarización sobre conversaciones telefónicas en las que se suman los dos canales de la conversación (por lo que, de antemano, se sabe que hay dos locutores en la señal de test); y la diarización sobre señales creadas mezclando segmentos extraídos de este tipo de conversaciones, por lo que el sistema desconoce la cantidad de locutores (aunque se sabe que no son más diez y que la duración de la señal de test no es superior a diez

minutos). En la NIST SRE de 2002 los segmentos mezclados se extraen desde otras fuentes: concretamente, desde conversaciones telefónicas, noticias de radio y televisión, y reuniones. Las señales de test se componen de segmentos extraídos de una fuente concreta (conocida de antemano), y su duración varía de uno a dos minutos. En este caso, se desconoce por completo la cantidad de locutores.

A partir de 2003, las NIST SRE excluyen la tarea de diarización, debido a que tal como se ha mencionado anteriormente se define una evaluación específica que incluye este tipo de tareas (NIST *Rich Transcripction Evaluation*, NIST RT), organizadas por el NIST desde 2002 [NIST RT]. Estas evaluaciones se centran en sistemas de transcripciones automáticas de voz a texto, pero que también producen información anexa, como marcas de sincronización, clasificación de los ítems reconocidos, identidad del locutor, etc. Las tareas definidas se engloban en dos áreas: (1) transcripción de voz a texto (*Speech-to-Text Transcription*, STT) y (2) extracción de metadatos (*Meta-Data Extraction*, MDE). Es en esta segunda área (MDE) donde se define una tarea de diarización de locutores.

2.6.2.2 Evaluaciones organizadas por otras entidades

Hasta este punto sólo se han considerado las evaluaciones organizadas por el NIST, dado que son referencia en la mayoría de las áreas relacionadas con las tecnologías del habla. Pero a lo largo de los años se han organizado otras evaluaciones, a cargo de diferentes entidades, que tratan de completar o contrastar las evaluaciones del NIST. Entre ellas, en relación con la detección y el seguimiento de locutores, se encuentran las siguientes:

- La evaluación NFI/TNO, orientada al reconocimiento forense de locutores, organizada por NFI (*Netherlands Forensic Institute*) y TNO (*Netherlands Organization for Applied Scientific Research*) en 2003. La evaluación se basa en un corpus grabado en investigaciones policiales reales. Incluye varios conjuntos de test con señales de longitud diversa e idiomas diferentes. Debido a que el corpus se ha grabado en situaciones reales, el material grabado es de duración limitada, por lo que en las tareas definidas la duración media es de 60 segundos para las pronunciaciones de entrenamiento y de 15 segundos para las del test. El corpus se compone de 22 locutores objetivo y 30 impostores. La evaluación se define como complementaria a las del NIST SRE. Véase [VanLeeuwen06] para más información.
- La campaña de Evaluación Biosecure 2007, organizada por la asociación BioSecure NoE (*BioSecure Network of Excellence*) en el marco del proyecto BioSecure del sexto Programa Marco de la comunidad Europea [Mayoue09]. La asociación BioSecure reúne a unas 30 empresas y entidades de investigación, interesadas en sistemas biométricos. La evaluación, celebrada a mediados del 2007, plantea un escenario móvil con sistemas biométricos mono-modales (basados en la voz, el rostro, la huella dactilar, la firma) y multi-modales (basados en la combinación de sistemas mono-modales).
- Las evaluaciones EVALITA, celebradas por primera vez en 2007, y organizadas, de forma voluntaria, por varias entidades (principalmente italianas) interesadas en el procesamiento de

lenguaje natural en italiano [EVALITA]. En el segundo semestre de 2009 se celebró la segunda edición de la evaluación, que incluye tareas relacionadas con el procesamiento de texto y el procesamiento del habla. En esta última categoría, se han definido una tarea de verificación de locutores para aplicaciones comerciales y forenses.

- Las evaluaciones CLEAR (*Classification of Events, Activities and Relationships*) celebradas en 2006 y 2007, con objeto de evaluar sistemas dedicados a eventos, actividades y sus interacciones [CLEAR]. Estas evaluaciones se organizan en el marco de los proyectos CHIL de la comunidad europea y VACE (*Video Analysis and Content Extraction*) de Estados Unidos. Entre las tareas definidas se encuentran: el seguimiento de personas mediante audio, video o ambas combinaciones; el seguimiento de rostros faciales; el seguimiento de vehículos; y la identificación de personas mediante audio, video o su combinación.
- La evaluación celebrada en la campaña ESTER [ESTER] organizada en el marco del proyecto EVALDA en la que participan AFCP (*Francophone Speech Communication Association*), ELDA (*Evaluations and Language Resources Distribution Agency*) y DGA/CEP (*French Defense expertise and test center for speech and language processing*). El objetivo principal de la campaña es la evaluación de sistemas de transcripción automática sobre noticias y programas difundidos por radio y televisión en lengua francesa. La primera evaluación ESTER tuvo lugar a principios de 2005 [Galliano05], e incluía diversas tareas, englobadas en las tres áreas siguientes: (1) transcripciones automáticas; (2) segmentación acústica; y (3) extracción de información. Dentro del área de segmentación acústica se definió una tarea de seguimiento de locutores. Esta evaluación trataba de completar las evaluaciones NIST-RTE, que no consideran señales en francés (en las evaluaciones NIST-RTE se analizan señales en inglés, árabe y chino). Recientemente, concretamente durante 2008, se ha definido un nuevo corpus de características similares (que también incluye noticias radio-televisivas) y una nueva evaluación denominada ESTER-2. En esta evaluación celebrada a principios de 2009, se ha añadido una nueva tarea, consistente en la detección de inicio y final de oraciones. Las bases de datos de las dos evaluaciones ESTER forman parte del catálogo de la ELRA.
- Las evaluaciones Albayzín, de ámbito estatal, celebradas en 2006 y 2008 bajo la organización de la Red Temática en Tecnologías del Habla (RTTH). Las evaluaciones se celebran dentro del marco de las Jornadas en Tecnologías del Habla que tienen lugar en España. Entre otras tareas relacionadas con la síntesis de voz y la traducción automática, en relación con la detección de locutores y verificaicón de la lengua, se organizó una evaluación de segmentación e identificación de locutores en 2006 y una evaluación centrada en la verificación de la lengua en 2008 [Rodriguez-Fuentes10].

2.6.3 Métricas de rendimiento

Las métricas que miden el rendimiento de un sistema juegan, obviamente, un papel muy importante en el proceso de evaluación de las tecnologías. Son herramientas que determinan la validez de los métodos propuestos, y permiten su comparación directa. Las métricas han de ser claras y de fácil comprensión. Generalmente, se opta por representar los errores en lugar de la tasa de aciertos. La razón de ello es que la representación de los errores resulta más intuitiva. Desde este punto de vista, si la tasa de error disminuye de un 10% a un 5% (lo que, obviamente, equivale a un incremento del 90% al 95% con respecto a la tasa de acierto) el rendimiento del sistema habrá mejorado en un 50%, y de hecho, su rendimiento real es doblemente mejor que el anterior [Doddington00].

2.6.3.1 Identificación de locutores

El rendimiento de los sistemas de identificación de locutores se representa mediante la *tasa de error de identificación*, *ER*, que indica la relación entre la cantidad de observaciones clasificadas erróneamente n_{err}, y la cantidad total de observaciones de la experimentación n_{total} [Doddington00]:

$$ER = \frac{n_{err}}{n_{total}} \tag{2.42}$$

Generalmente, se opta por representar el porcentaje de error:

$$ER\% = ER*100 \tag{2.43}$$

2.6.3.2 Verificación de locutores

En un sistema de verificación, la experimentación se divide en: (1) experimentos de los locutores objetivo (denominados experimentos objetivo (*target trials*) en adelante), que reúnen el conjunto de observaciones que el sistema debe aceptar; y (2) experimentos de impostores (denominados experimentos impostor (*impostor trials*) en adelante), que reúnen el conjunto de observaciones que el sistema debe rechazar [Przybocki06]. Los errores cometidos en experimentos objetivo se denominan *errores de detección* (*misses*), mientras que los cometidos en experimentos impostor se denominan *falsas alarmas* (*false alarms*). De ahí, el rendimiento de un sistema de detección o verificación se determina mediante las dos siguientes medidas:

(1) P_{miss}: la probabilidad de no detectar el locutor objetivo cuando éste está presente, es decir el porcentaje de experimentos objetivo que se han clasificado de forma errónea:

$$P_{miss} = \frac{n_{miss}}{n_{obj}} \tag{2.44}$$

donde n_{obj} y n_{miss} son, respectivamente, la cantidad total de experimentos objetivo y la cantidad de este tipo de experimentos en los que no se ha detectado el locutor objetivo.

(2) P_{fa}: la probabilidad de detectar de forma errónea un locutor objetivo cuando éste no está presente, es decir el porcentaje de experimentos impostor que se han clasificado de forma errónea:

$$P_{fa} = \frac{n_{fa}}{n_{imp}} \tag{2.45}$$

donde n_{imp} y n_{fa} son, respectivamente, la cantidad total de experimentos impostor y la cantidad de este tipo de experimentos en los que se ha detectado un locutor objetivo.

A partir de estas dos tasas de error, en las evaluaciones NIST SRE se define la siguiente función de coste como métrica de error de un sistema de detección. La función, denominada C_{DET}, es una combinación lineal de las probabilidades P_{miss} y P_{fa} del sistema:

$$C_{DET}=C_{miss}* P_{miss}*P_{obj}+ C_{fa}*P_{fa}*(1.0-P_{obj}) \tag{2.46}$$

donde C_{miss} y C_{fa} son el coste relativo de los errores de detección y falsas alarmas, respectivamente; y P_{obj} es la probabilidad a priori de un experimento objetivo. Son parámetros que dependen del escenario de aplicación para el que se diseña el sistema, es decir, parámetros que pretenden trasladar la evaluación del sistema al escenario real de aplicación. P_{obj} no es el porcentaje de experimentos objetivo del conjunto global de experimentos, sino el porcentaje de locutores objetivo que se prevé en los posibles escenarios de aplicación de interés. Lo mismo sucede con los costes C_{miss} y C_{fa}, que estarán relacionados con la tolerancia que se considere asumible en el sistema respecto a la comisión de cada tipo de error (p. ej. en control de acceso de alta seguridad C_{fa} tendrá un valor alto para considerar intolerable un acceso no autorizado, mientras que C_{miss} puede ser pequeño al tolerarse rechazos de personal autorizado que pueden ser "reparados" mediante reintentos).

Con objeto de mejorar la representación intuitiva de la función de coste C_{DET}, éste se normaliza con el coste mínimo de un sistema sin capacidad de discriminación, es decir con el coste mínimo entre determinar que todas las observaciones son verdaderas o que todas son falsas:

$$C_{norm}=\frac{C_{DET}}{C_{Default}} \tag{2.47}$$

donde:

$$C_{Default}=\min\begin{Bmatrix} C_{miss}*P_{obj} \\ C_{fa}*(1.0-P_{obj}) \end{Bmatrix} \tag{2.48}$$

En las NIST SRE, C_{miss} y C_{fa} se establecen a 10 y 1 respectivamente. P_{obj}, en cambio, se establece a 0.01. Por tanto, el factor de normalización, $C_{Default}$, es 0.1.

Un experimento de verificación requiere la determinación de una *decisión final*, un valor booleano que indique si la observación de test corresponde o no al locutor objetivo: *verdadero*, en caso de que se detecte el locutor en la observación; *falso*, en caso contrario. Para ello, se estima una probabilidad de detección (generalmente el LLR) que refleja la certeza del sistema en cuanto a la participación del locutor en la observación (como generalmente se estima la probabilidad del locutor dada la observación de test, valores más altos apuntan a que se tiene una certeza mayor con respecto a la participación del locutor). Esta probabilidad se compara con *un umbral de decisión* predeterminado (generalmente estimado sobre un corpus de desarrollo independiente del corpus de evaluación): si la probabilidad es superior al umbral, se determina que la observación corresponde al locutor objetivo; en caso contrario, se determina que la observación corresponde a un impostor.

Las métricas C_{DET}, P_{miss} y P_{fa} evalúan el rendimiento del sistema en base a decisiones finales y un umbral de decisión específico. No obstante, el rendimiento de un sistema se puede representar haciendo un barrido, esto es, variando el umbral por todo su rango (para un umbral determinado se habla de un "punto de operación" concreto del sistema). Esta representación trata de visualizar las métricas $P_{miss}(t)$ y $P_{fa}(t)$ de cada punto de operación *t* del sistema. Para ello, se pueden emplear las clásicas curvas ROC (*Receiver Operating Characteristics*). En [Martin97] se propone una curva alternativa, denominada DET (*Detection Error Trade-off*), que trata de mejorar la visualización de todos los puntos de operación, dibujando las probabilidades $P_{miss}(t)$ y $P_{fa}(t)$ en una escala de desviaciones estándar. Es la curva más empleada hoy en día, debido a que ofrece una distinción más clara del rendimiento de diferentes curvas. En [DET] se puede encontrar un paquete de software desarrollado por el NIST para generar curvas DET en el entorno MATLAB. En este trabajo, con objeto de disponer de un paquete ejecutable en entornos de software libre, el paquete de generación de curvas DET se ha convertido al entorno Scilab, un entorno de software dedicado a computación numérica y cuya licencia, a diferencia de MATLAB, es libre [Scilab].

En la misma línea de analizar el rendimiento del sistema en diferentes puntos de operación se encuentran otras dos métricas también muy conocidas y aplicadas, las funciones EER y C_{DET}^{min}. EER (*Equal Error Rate*) es una métrica empleada en la mayoría de los trabajos de detección de locutores, la tasa de error del punto de operación *t* en el que se igualan las probabilidades P_{miss} y P_{fa}:

$$EER = P_{miss}(t) = P_{fa}(t) \tag{2.49}$$

C_{DET}^{min}, en cambio, corresponde al punto de operación *t* de entre todos los posibles puntos, para el que $C_{DET}(t)$ es mínimo (el umbral óptimo del sistema):

$$C_{DET}^{min} = \min_t C_{DET}(t) \tag{2.50}$$

Esta función de coste es el resultado principal de las evaluaciones NIST SRE [NIST]. Obviamente, cuanto más similares son C_{DET} y C_{DET}^{min}, más apropiado es el punto de operación definido.

Una métrica independiente de la aplicación

Recientemente, se ha definido una nueva métrica, que a diferencia de C_{DET}, es independiente de la aplicación. Es la función C_{llr} propuesta y descrita en [Brummer06], que pretende evaluar el rendimiento de un sistema de forma independiente al escenario de aplicación, es decir para un amplio rango de costes y errores. Aplicando conceptos de la teoría de la información, C_{llr} trata de evaluar el rendimiento de un sistema en base a probabilidades de detección (específicamente, los LLR), y no en base a las decisiones tomadas (aceptar o rechazar al locutor objetivo) como se realiza en la función de coste C_{DET}. Así, se posibilita que el diseño de los sistemas sea independiente de la aplicación, dado que la función evalúa el sistema para un amplio rango de aplicaciones. Esta función se empleó por primera vez en la NIST SRE de 2006, y se ha vuelto a aplicar para comparar el rendimiento de los sistemas presentados en la NIST SRE de 2010. Para estimar la métrica C_{llr} se deben conocer las probabilidades de cada experimento del conjunto de evaluación:

$$C_{llr} = \frac{1}{2*(\log 2)} * \left(\frac{\sum \log\left(1+\frac{1}{s\_llr}\right)}{N_{EO}} \right) + \left(\frac{\sum \log(1+s\_llr)}{N_{EI}} \right) \quad (2.51)$$

donde el primer sumatorio se realiza sobre las probabilidades de los experimentos objetivo y el segundo sobre las probabilidades de los experimentos impostor; N_{EO} y N_{EI} son la cantidad de experimentos objetivo e impostor respectivamente; y *s_llr* representa la puntuación recibida por el locutor para un experimento.

Los sistemas de detección analizados hasta ahora, emiten una salida estricta (aceptado/rechazado) que requiere la estimación de un umbral predeterminado dependiente de la aplicación (y estimado sobre un conjunto de desarrollo). En el mismo trabajo, Brummer et. al. [Brummer06] defienden que, si un sistema de detección representa el ratio teórico entre la probabilidad de la observación para el locutor objetivo y la probabilidad de la observación para el conjunto de impostores, se puede definir un umbral teórico para la aplicación: el umbral de Bayes, definido a partir de los costes y probabilidades a priori de un escenario de aplicación en concreto. De esta forma, se posibilita que el umbral se pueda estimar de forma teórica a partir de las características de la aplicación, y no se tenga que analizar el comportamiento empírico del sistema (sobre un corpus de desarrollo) para determinar un umbral adecuado para dicha aplicación. Dados los costes C_{fa} y C_{miss}, y la probabilidad P_{obj}, el umbral de Bayes calculado en escala logarítmica es:

$$\widehat{\theta_b} = \log \frac{C_{fa}*(1-P_{obj})}{C_{miss}*P_{obj}} \quad (2.52)$$

El umbral se muestra en escala logarítmica debido a que generalmente el ratio de las dos hipótesis se calcula también en escala logarítmica. En definitiva, si $LLR > \widehat{\theta_b}$ se decide que el locutor de la señal analizada es el locutor objetivo; en caso contrario, se decide que el locutor de la señal analizada es un impostor.

En el mismo trabajo [Brummer06], se propone la transformación de las probabilidades de detección (probabilidades estimadas mediante modelos que dan por hecho que las hipótesis y los parámetros de entrada son ciertos) a ratios o probabilidades óptimas. Este proceso de optimización se denomina *calibración* de probabilidades, y consiste en maximizar la información mutua, es decir, en estimar el desplazamiento óptimo de las probabilidades estimadas con objeto de maximizar la información aportada por el sistema. El proceso de optimización es una transformación lineal estimada mediante regresión logística, cuyos parámetros se estiman sobre un corpus de desarrollo (en [Brummer06] se definen varias funciones de transformación, así como el procedimiento para estimar los parámetros óptimos de una transformación). La única precondición de las funciones definidas es que las probabilidades de detección (o scores) deben estar correlacionadas positivamente con la probabilidad del locutor objetivo, es decir que cuanto mayor sea la tasa, más probable debe ser la probabilidad del locutor objetivo, y viceversa. En esta línea, Brummer define C_{llr}^{min} como el valor de la función que determina la capacidad de discriminación máxima de un sistema, el C_{llr} de un sistema calibrado en cerrado. Asimismo, mediante el cálculo de la distancia entre C_{llr} y C_{llr}^{min}, se puede analizar si el sistema

está debidamente calibrado. Si se asume que la calibración realiza el mapeo de las probabilidades de detección a probabilidades o ratios teóricos, se puede decir que el proceso facilita la determinación del umbral óptimo (el umbral de Bayes), y en consecuencia se ha de cumplir: $C_{DET} \approx C_{DET}^{min}$. Para ello, es necesario que el corpus de desarrollo dé cobertura a los experimentos del corpus de evaluación.

2.6.3.3 Seguimiento de locutores

El rendimiento de un sistema de seguimiento de locutores se evalúa de forma muy similar al de un sistema de verificación [Przybocki99]. Al igual que en un sistema de verificación, se debe tomar una decisión final y estimar una probabilidad, pero a diferencia de los sistemas de verificación, la decisión y la probabilidad no se estiman para la pronunciación de entrada, sino para un intervalo de tiempo de la pronunciación que ha determinado el sistema de seguimiento. En consecuencia, las probabilidades P_{miss} y P_{fa}, correspondientes a los errores de detección y falsas alarmas se estiman como funciones del tiempo de los segmentos del locutor de referencia determinados en el proceso de anotación de las pronunciaciones (es decir, los segmentos reales):

$$P_{miss} = \frac{\int_{SegmentosObjetivo} f(D_t, F, l_o, l_d)\,dt}{\int_{SegmentosObjetivo} dt} \tag{2.53}$$

$$P_{fa} = \frac{\int_{SegmentosImpostor} f(D_t, T, l_o, l_d)\,dt}{\int_{SegmentosImpostor} dt} \tag{2.54}$$

donde:

- D_t es la salida del sistema de seguimiento, como función del tiempo: T (verdadero) si se ha detectado que el segmento corresponde a alguno de los locutores objetivo; F (falso) en caso contrario, si se ha detectado que el segmento es de un impostor.
- l_o es la etiqueta del locutor objetivo (el de referencia).
- l_d es la etiqueta del locutor detectado (determinado) por el sistema de seguimiento.
- $$f(x, y, o, d) = \begin{cases} 1 & \text{si } x = y \\ 1 & \text{si } D_t = T \text{ y } o \neq d \\ 0 & \text{en caso contrario} \end{cases} \tag{2.55}$$

La tarea de seguimiento definida en las evaluaciones NIST SRE consiste en detectar un único locutor en la observación de entrada, por lo que en los documentos publicados [Przybocki99][Martin01] la función *f* tiene dos únicos argumentos *x* e *y*. Dado que el sistema planteado en este trabajo pretende realizar el seguimiento de más de un locutor simultáneamente, se han añadido los parámetros *o* y *d* con objeto de detectar las situaciones en las que se determina que el segmento de un locutor objetivo pertenece a otro locutor objetivo.

Estas probabilidades se estiman como función del tiempo de los segmentos de referencia. Pero se establece un periodo neutro (denominado *collar period*), de 250 milisegundos, en cada punto de cambio de locutor. Es decir, los 250 milisegundos del inicio y final de un segmento de referencia se

ignoran en la evaluación, y en consecuencia, no se consideran los segmentos inferiores a 0.5 segundos [Przybocki99].

Para determinar la decisión final (locutor objetivo/impostor), la probabilidad de detección estimada para el segmento analizado se compara con un umbral predeterminado, al igual que en los sistemas de verificación (véase apartado 2.6.3.2). Por tanto, las funciones de coste C_{DET} y C_{DET}^{min}, y la métrica EER definidas en dicho apartado en las ecuaciones 2.46, 2.50 y 2.49, respectivamente, se pueden aplicar de forma idéntica que en la verificación para evaluar el rendimiento de un sistema de seguimiento de locutores. Lo único que cambia es el valor de los parámetros dependientes de la aplicación, que en las evaluaciones del NIST se fijan a C_{miss}=1, C_{fa}=1, y P_{obj}=0.5 (es decir, se igualan todos los costes y probabilidades a priori). Asimismo, si se quiere analizar el rendimiento del sistema con respecto a todos los puntos de operación (umbrales) posibles, se emplean las curvas DET descritas en el mismo apartado.

El proceso de calibración es también adecuado para los sistemas de seguimiento. Este proceso trata de transformar linealmente las probabilidades de detección a probabilidades o ratios, con objeto de definir un umbral para cada aplicación (el umbral de Bayes, véase la ecuación 2.52). La transformación es idéntica a la de los sistemas de verificación [Brummer][Brummer06], a excepción de que la calibración se aplica sobre los segmentos determinados por el sistema de seguimiento (y no sobre cada experimento). En cuanto a la función de coste C_{llr} (véase la ecuación 2.51), se puede definir de forma similar para el seguimiento, reemplazando el conjunto de experimentos objetivo por el conjunto de segmentos objetivo, el conjunto de experimentos impostor por el conjunto de segmentos impostor, y por último, la cantidad de experimentos de cada conjunto por el sumatorio en tiempo de los segmentos de cada conjunto. El valor C_{llr}^{min} se define de la misma forma que en verificación.

Una métrica relacionada con la Extracción de Información

Por último, otra métrica muy empleada en los sistemas de seguimiento es la conocida medida-F (*F-measure*). Esta función trata de determinar el verdadero rendimiento de los sistemas de Extracción de Información (*Information Retrieval*), entre los que se encuentran los sistemas de seguimiento (o diarización) de locutores. La evaluación se realiza para un punto de operación dado (el punto definitivo). La función se define de la siguiente manera:

$$\text{medida-F} = \frac{2.0 * \text{PRC} * \text{RCL}}{\text{PRC} + \text{RC}L} \tag{2.56}$$

donde PRC es la precisión del sistema y RCL la cobertura (*recall*), que reflejan, respectivamente, falsas alarmas y errores de detección:

$$\text{PRC} = \frac{\text{tiempo objetivo detectado correctamente}}{\text{tiempo objetivo detectado}} \tag{2.57}$$

$$\text{RCL} = \frac{\text{tiempo objetivo detectado correctamente}}{\text{tiempo objetivo total}} \tag{2.58}$$

La medida-F alcanza valores comprendidos entre 0 y 1. Obviamente, cuanto mayor es la medida-F de un sistema mejor es su rendimiento. A veces se dibuja la curva que representa la relación

precisión-cobertura de todos los puntos de operación del sistema (como con falsas alarmas y errores de detección en las curvas ROC). Está gráfica se conoce como la curva precisión-cobertura (*precisión-recall curve*).

Capítulo 3

Reducción de dimensionalidad de la parametrización acústica

El vector de parámetros acústicos es portador de información diversa, como el idioma, el locutor, el mensaje emitido, el estado emocional, etc. Es deseable contar únicamente con parámetros que son relevantes para la tarea de clasificación, ignorando características redundantes o que aporten información innecesaria. La mayoría de los sistemas actuales emplean el mismo conjunto de parámetros estándar (MFCC, la energía, y sus derivadas) tanto para el reconocimiento del habla como para la identificación y verificación de locutores, porque además de recoger la distribución frecuencial para la identificación de sonidos, transmiten información sobre el pulso glotal y la forma y longitud del tracto vocal, que son características específicas de cada locutor. En consecuencia, el vector de parámetros puede estar compuesto por un número de componentes que típicamente varía entre 20 y 50.

En una tarea de identificación o verificación de locutores, cada parámetro acústico debe contar con las siguientes características:

- Alta variación inter-locutor
- Baja variación intra-locutor
- Facilidad de cálculo
- Robustez frente al ruido
- Independencia frente a otros parámetros.

Desafortunadamente, no existe ningún parámetro que cumpla con todos estos requerimientos**.** En este apartado se estudian y aplican métodos para reducir eficazmente la dimensionalidad del espacio de los parámetros acústicos en tareas de reconocimiento de locutores.

3.1 Técnicas de Reducción de Dimensionalidad

En la bibliografía se pueden encontrar varios estudios comparativos que analizan diversos métodos de extracción de parámetros y determinan el más adecuado para una tarea concreta, es decir aquel que proporciona un mejor rendimiento (véase por ejemplo [Reynolds94]). En este trabajo se plantea un estudio diferente; en vez de analizar el rendimiento de diferentes representaciones para el reconocimiento de locutores, se determinan los parámetros acústicos más relevantes partiendo de un conjunto de parámetros acústicos estándar (concretamente MFCC, la energía y sus primeras y segundas derivadas). Se trata de obtener conjuntos reducidos de parámetros que permitan ahorrar esfuerzo computacional con una baja degradación o incluso con una mejora del rendimiento del sistema. Para ello, se analizan un gran número de subconjuntos de diferentes dimensiones: el espacio acústico d-dimensional se transforma en un sub-espacio de k dimensiones ($k<d$) tratando de minimizar la pérdida de información relevante para la discriminación. Para ello, se han seguido dos aproximaciones: (1) por un lado, se ha analizado la aplicación de transformaciones lineales en el espacio de parámetros; y (2) de forma alternativa, se ha propuesto la selección de parámetros mediante algoritmos genéticos (GA).

Una aproximación muy empleada para la reducción de dimensionalidad es la aplicación de transformaciones lineales, estimadas según un criterio de máxima variabilidad, máxima discriminación, etc. Las transformaciones lineales combinan la extracción y selección de parámetros en una única fase, y como resultado decorrelan el espacio y definen sub-espacios compuestos por las direcciones más relevantes. En este estudio se han aplicado las clásicas transformaciones lineales PCA y LDA; y una aproximación más reciente denominada HLDA que es una generalización de LDA. Así como LDA y PCA se han empleado en muchos trabajos hasta la fecha, existen muy pocos trabajos que apliquen HLDA.

La selección de parámetros consiste en la determinación de un subconjunto de k parámetros óptimos en un espacio d-dimensional. Para ello, analiza exhaustivamente las 2^d combinaciones posibles. La mayoría de los procedimientos de selección usan como función de evaluación la precisión del sistema de clasificación sobre un corpus de datos representativo. Por ello, en la práctica, la búsqueda exhaustiva resulta computacionalmente prohibitiva incluso para valores moderados de d. Una metodología más simple consistiría en evaluar de forma individual los d parámetros y seleccionar los k más discriminantes, sin tener en cuenta las dependencias entre ellos. De hecho, en la bibliografía pueden encontrarse varias técnicas que sacrifican la adecuación del conjunto seleccionado en beneficio de la eficiencia computacional [Jain00]. Una estrategia bastante razonable para realizar esta búsqueda sin establecer límites a toda la combinatoria posible es la utilización de GA. Los GA, introducidos por Holland en 1975 [Holland75], son un conjunto de técnicas de búsqueda heurística y aleatoria inspiradas en las estrategias de la evolución biológica, con tres mecanismos básicos: selección del más fuerte, mezcla y mutación. Presentan propiedades muy interesantes para el caso que nos ocupa; hacen una exploración heurística del espacio de parámetros y seleccionan aquel sub-espacio que maximiza

una cierta función de adecuación, que en este caso es el rendimiento del sistema de identificación de locutores. En este trabajo se ha propuesto el empleo de GA para la búsqueda del subconjunto óptimo de parámetros acústicos, con dos aproximaciones: (1) la selección de parámetros mediante búsqueda directa con GA; y (2) de forma alternativa, con el objetivo de reducir el coste computacional del proceso de búsqueda, la selección de parámetros mediante un ranking de pesos.

3.1.1 Transformaciones lineales

Las transformaciones lineales constituyen una aproximación muy empleada para reducir la dimensionalidad y aumentar la capacidad de discriminación. Proyectan el espacio original en un sub-espacio formado por las direcciones más relevantes. A cada vector acústico d-dimensional $\mathbf{x}$, se le aplica una matriz de transformación A de dimensión $k \times d$ ($k \leq d$), obteniendo como resultado un vector k-dimensional $\mathbf{y}$ que contiene los parámetros transformados. La matriz A se calcula de forma que maximice la capacidad de discriminación, eliminando la redundancia y manteniendo la información más relevante. De esta forma, se optimiza el rendimiento de la clasificación para un sub-espacio de k dimensiones pudiendo incluso mejorar el rendimiento obtenido en el espacio original.

En la bibliografía se encuentran varios trabajos que aplican transformaciones lineales en tareas relacionadas con la identificación y verificación de locutores:

- Errity y MckKenna analizan la aplicación de PCA y una metodología no-lineal (denominada Isomap) para la reducción de dimensionalidad en la identificación locutores [Errity07]. El sistema de identificación utiliza GMM(EM) como modelos de locutor y vectores MFCC de 19 componentes. En general, PCA ofrece mejor rendimiento que Isomap excepto para dimensiones muy pequeñas.
- Limat et. al. analizan el rendimiento de PCA en la verificación de locutores, comparando dos sistemas de clasificación; se aplica PCA sobre los datos que a continuación se clasifican mediante modelos GMM(EM) o modelos AR-Vector [Lima02]. Los autores indican que PCA aporta mejoras con GMM(EM), excepto cuando los modelos se estiman a partir de pocos datos.
- Jin y Waibel analizan la aplicación de LDA en la identificación de locutores [Jin00]. El trabajo utiliza también GMM(EM) como modelos de locutor y los autores defienden que la aplicación de LDA mejora el rendimiento de la identificación.
- Lukas Burget analiza el rendimiento de PCA, LDA y HLDA para decorrelar y reducir la dimensionalidad de vectores acústicos combinados en tareas de reconocimiento del habla [Burget04a]. La combinación consiste en la concatenación de varios conjuntos de parámetros. En los experimentos realizados, la combinación de parámetros, y su posterior decorrelación y reducción a través de transformaciones lineales, aportan muy buenos resultados.
- Yang et. al. aplican HLDA para mejorar la capacidad de discriminación de los modelos de locutor GMM(MAP) [Yang06]. La propuesta trata de aumentar la discriminación de cada distribución

gaussiana del espacio acústico, con objeto de reducir los solapamientos de las componentes gaussianas y mejorar así la robustez del sistema.

- Jang et. al., proponen otro tipo de transformación lineal denominada ICA (Independent Component Analysis) para la extracción de parámetros de un sistema de reconocimiento de locutores [Jang01]. El artículo compara tres metodologías de extracción de parámetros; ICA, PCA y FFT. En los experimentos realizados, ICA obtiene mejores resultados que las otras dos aproximaciones.

El mayor interés de las transformaciones lineales reside en su facilidad de cómputo y su resolución analítica exacta. Este apartado recoge una breve descripción de cada una de las tres metodologías que se han aplicado en este trabajo: PCA, LDA y HLDA.

3.1.1.1 PCA

El Análisis de Componentes Principales (PCA) [Pearson1901], conocida también como la transformada de Karhunen-Loève (*Karhunen-Loève Transform*, KLT), consiste en la búsqueda de las direcciones de mayor varianza de un conjunto de datos (la metodología se describe forma extensa en [Jolliffe02]). Es una técnica que permite realizar transformaciones en una base orto-normal. La matriz de transformación A se estima de forma que su primer vector base (primera fila de la matriz) define la dirección de mayor varianza del espacio original d-dimensional. El segundo vector base define una dirección perpendicular a la primera y contará con la segunda mayor varianza, y así sucesivamente.

Los vectores transformados $\mathbf{y} = A\mathbf{x}$ cumplen la siguiente condición:

$$\bar{\Sigma} = A\Sigma A^{\mathrm{T}} \tag{3.1}$$

donde Σ y $\bar{\Sigma}$ corresponden a la matriz de covarianza del espacio original y del espacio transformado, respectivamente. Por tanto, la transformación PCA consiste en la búsqueda de los auto-vectores y auto-valores de Σ. Estos auto-vectores se ordenan de acuerdo a sus auto-valores: el vector base $\mathbf{a}_i$ correspondiente a la fila i de la matriz de transformación A, es el auto-vector que obtiene el i-ésimo mayor auto-valor (o la i-ésima "componente principal").

La transformada PCA tiene dos propiedades importantes:

- Dada una matriz de transformación A, los coeficientes de los vectores transformados $\mathbf{y} = A\mathbf{x}$, no están correlados ($\bar{\Sigma}$ es diagonal). Los auto-valores indican la varianza del espacio transformado.
- Si la proyección se realiza con los p primeros vectores base $\mathbf{a}_i$, con $i \leq p < d$, PCA sirve, de hecho, para reducir la dimensionalidad.

Pero hay que tener en cuenta que aunque PCA busca una representación adecuada mediante las componentes (o direcciones) de mayor varianza, no hay razón alguna para asumir que estas componentes definen una discriminación óptima entre clases.

3.1.1.2 LDA

El Análisis Discriminante Lineal (LDA) [Fisher36][Duda00], al igual que PCA, busca una transformación lineal que posibilite la reducción de dimensionalidad. No obstante, a diferencia de PCA, que busca direcciones eficientes para representar los datos originales, LDA trata de encontrar las direcciones más relevantes para la discriminación entre clases. El espacio original puede contener direcciones que PCA descarte por su varianza y sin embargo sean necesarias para la clasificación.

Para analizar la capacidad de discriminación se debe conocer la clase a la que pertenecen los vectores con los que se estimará la matriz de transformación. LDA trata de estimar una matriz de transformación que maximice la varianza externa y que al mismo tiempo minimice la varianza interna. Parte de la hipótesis de que la distribución de los vectores de una clase es gaussiana y que todas las clases comparten la misma matriz de covarianza, es decir, que todas las clases tienen la misma varianza interna. El criterio de separabilidad consiste en maximizar el ratio entre las matrices $\bar{\Sigma}_{bc}$ y $\bar{\Sigma}_{wc}$ de los datos transformados, donde en un conjunto de datos compuesto por *N* vectores, *J* clases y N_j vectores en cada clase *j*:

- Σ_{wc} es la matriz de covarianza intra-clase (*within-class*), la media ponderada de las matrices de covarianza de todas las clases ($\hat{\Sigma}^{(j)}$ corresponde a la matriz de covarianza de la clase *j)*:

$$\hat{\Sigma}_{wc} = \frac{1}{N}\sum_{j=1}^{J} N_j \hat{\Sigma}^{(j)} \tag{3.2}$$

- Σ_{bc} es la matriz de covarianza inter-clase (*between-class*), y se estima con el vector de medias $\hat{\mu}^{(j)}$ de cada clase ($\hat{\boldsymbol{\mu}}$ es el vector de medias global del conjunto de entrenamiento):

$$\hat{\Sigma}_{bc} = \frac{1}{N}\sum_{j=1}^{J} N_j \left(\hat{\boldsymbol{\mu}}^{(j)} - \hat{\boldsymbol{\mu}}\right)\left(\hat{\boldsymbol{\mu}}^{(j)} - \hat{\boldsymbol{\mu}}\right)^{\mathrm{T}} \tag{3.3}$$

Por lo tanto, dada la matriz de transformación *A*, y las matrices Σ_{bc} y Σ_{wc} de los datos originales, se debe maximizar la siguiente función*:*

$$T(A) = \frac{\left|A^{\mathrm{T}}\Sigma_{bc}A\right|}{\left|A^{\mathrm{T}}\Sigma_{wc}A\right|} \tag{3.4}$$

La maximización de *T*(*A*) nos conduce a la búsqueda de los auto-vectores del producto $\Sigma_{bc} \times {\Sigma_{wc}}^{-1}$ que constituyen la matriz LDA. Se repite el mismo procedimiento que en PCA: los auto-valores determinan la relevancia de los auto-vectores, por lo que éstos se ordenan por auto-valor. En consecuencia, la reducción de dimensionalidad se realiza con los vectores base que tratan de conservar la máxima variabilidad relevante para la clasificación. Un espacio transformado mediante LDA puede contener como máximo *J-1* dimensiones, puesto que la matriz tendrá como máximo J-1 auto-valores no nulos.

3.1.1.3 HLDA

El Análisis Discriminante Lineal Heteroscedástico (HLDA) es una técnica propuesta por N. Kumar [Kumar97] como una generalización de LDA. Tal y como se ha mencionado anteriormente, el criterio de optimización de LDA maximiza el ratio entre la varianza externa y la varianza interna en el

conjunto de clases. La matriz LDA se puede estimar eficientemente por ML y presenta buen rendimiento para las distribuciones cuyas clases tienen diferentes medias pero comparten la misma matriz de covarianza. No obstante, LDA puede determinar transformaciones erronéas cuando la hipótesis de compartición de una misma covarianza interna es inasumible, es decir cuando la distribución de clases es heterocedástica. El término "heteroscedasticidad" deriva de la palabra *hetero* (distinto) y del verbo griego *skedasis* que significa dispersión, y por tanto hace referencia a distribuciones estadísticas cuyas matrices de covarianza difieren. Su término complementario es *homoscedásticidad* que obviamente indica que las distribuciones tienen la misma matriz de covarianza.

Por ejemplo, supóngase una distribución gaussiana con dos clases cuyas medias se distinguen suavemente en una dirección y sus varianzas notablemente en otra. En este caso, la transformación debería proyectar el espacio en esta segunda dirección que separa las clases por su varianza, siendo para ello necesario que el modelo considere las propias varianzas de las clases..

Además de tener en cuenta las distintas covarianzas internas, HLDA parte de la hipótesis de que el espacio original *d*-dimensional se puede dividir en dos sub-espacios estadísticamente independientes:

1. Un sub-espacio formado por *p* dimensiones que posibilitan la discriminación de clases.
2. Un sub-espacio formado por las *d-p* dimensiones restantes, consideradas irrelevantes, donde las distribuciones están solapadas (las medias y varianzas tienen el mismo valor para toda las clases).

Dado que *A* es la matriz de transformación HLDA de un espacio vectorial de *d* dimensiones, la proyección se puede representar de la siguiente manera:

$$\mathbf{y} = A\mathbf{x} = \begin{bmatrix} A_{[p]}\mathbf{x} \\ A_{[d-p]}\mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_{[p]} \\ \mathbf{y}_{[d-p]} \end{bmatrix} \tag{3.5}$$

donde $A_{[p]}$ es la matriz que contiene las *p* primeras filas de la matriz *A* (cada fila con *d* componentes); $A_{[d-p]}$ es la matriz que contiene las *d-p* filas restantes; $\mathbf{y}_{[p]}$ son las dimensiones del espacio transformado que aportan información a la clasificación; y, por último, $\mathbf{y}_{[d-p]}$ son las dimensiones consideradas irrelevantes.

Para la búsqueda de la transformación óptima HLDA, $\hat{A}$, se ha de definir un modelo que cumpla todas estas condiciones y permita la estimación de sus parámetros por ML. Como la distribución de clases del espacio original es gaussiana, su transformación lineal es también gaussiana. Por tanto, cada clase *j* del espacio transformado se puede modelar mediante una gaussiana y la función de probabilidad asociada es la siguiente:

$$p_j(\mathbf{x}) = \frac{|A|}{\sqrt{(2\pi)^d |\bar{\Sigma}^{(j)}|}} e^{\left(-\frac{(A\mathbf{x}-\bar{\boldsymbol{\mu}}^{(j)})^{\mathrm{T}} \bar{\Sigma}^{(j)^{-1}} (A\mathbf{x}-\bar{\boldsymbol{\mu}}^{(j)})}{2}\right)} \tag{3.6}$$

donde $\bar{\boldsymbol{\mu}}^{(j)}$ y $\bar{\Sigma}^{(j)}$ son el vector de medias y la matriz de covarianza de la clase *j* en el espacio transformado. En todos los experimentos realizados en esta tesis, como en la mayoría de los sistemas

de reconocimiento de voz de la bibliografía, se emplean matrices de covarianza diagonales. Por tanto, se asume que todos los elementos que no están en la diagonal de la matriz de covarianza tienen valor cero y se denota como $\bar{\sigma}_t^{(j)^2}$ la *t*-ésima componente de la diagonal. Asimismo, se igualan los parámetros $\bar{\mu}_t^{(j)}$ y los $\bar{\sigma}_t^{(j)^2}$ para $t > p$, que corresponden a las dimensiones irrelevantes. A partir de este modelo, se maximiza el logaritmo de la verosimilitud de los datos para los parámetros *A*, $\overline{\boldsymbol{\mu}}^{(j)}$ y $\overline{\boldsymbol{\sigma}}^{(j)^2}$, con $1 \leq j \leq J$:

$$\mathcal{L}\left(\{\mathbf{x}_i\}\middle|A, \left\{\overline{\boldsymbol{\mu}}^{(j)}, \overline{\boldsymbol{\sigma}}^{(j)^2}\right\}\right) =$$

$$= \sum_j^J \sum_i^{N_j} \log p_j\left(\mathbf{x}_i^{(j)}\right) = \sum_j^J \sum_i^{N_j} \left[\frac{1}{2}\log\left(\frac{|A|^2}{(2\pi)^d \prod_{t=1}^d \bar{\sigma}_t^{(j)^2}}\right) - \sum_{t=1}^d \frac{\left(\boldsymbol{a}_t \mathbf{x}_i^{(j)} - \mu_t^{(j)}\right)^2}{2\bar{\sigma}_t^{(j)^2}}\right] \quad (3.7)$$

Donde $\mathbf{x}_i^{(j)}$ es el vector de entrenamiento *i* de la clase *j* y $\mathbf{a}_t$ la fila *t* de la matriz de transformación *A*. Habiendo establecido la matriz *A*, se puede resolver por ML la estimación de los parámetros $\bar{\mu}_t^{(j)}$ y $\bar{\sigma}_t^{(j)^2}$. Para ello, se calculan las derivadas con respecto a $\bar{\mu}_t^{(j)}$ y $\bar{\sigma}_t^{(j)^2}$ de la ecuación 3.7 y se resuelven los puntos donde estas derivadas son iguales a cero:

$$\hat{\bar{\mu}}_t^{(j)} = \begin{cases} \mathbf{a}_t \hat{\boldsymbol{\mu}}^{(j)} & t \leq p \\ \boldsymbol{a}_t \hat{\boldsymbol{\mu}} & t > p \end{cases} \quad (3.8)$$

$$\hat{\bar{\sigma}}_t^{(j)^2} = \begin{cases} \mathbf{a}_t \hat{\Sigma}^{(j)} \mathbf{a}_t{}^{\mathrm{T}} & t \leq p \\ \mathbf{a}_t \hat{\Sigma} \mathbf{a}_t{}^{\mathrm{T}} & t > p \end{cases} \quad (3.9)$$

donde $\hat{\boldsymbol{\mu}}^{(j)}$, $\hat{\boldsymbol{\mu}}$, $\hat{\Sigma}^{(j)}$ y $\hat{\Sigma}$ son el vector de medias de la clase *j*, el vector de medias global, la matriz de covarianza de la clase *j* y la matriz de covarianza global, respectivamente. Sustituyendo las ecuaciones 3.8 y 3.9 en la ecuación 3.7, el logaritmo de la verosimilitud de los datos se puede expresar mediante las matrices de covarianza $\hat{\Sigma}^{(j)}$ y $\hat{\Sigma}$, que se conocen de antemano, y la matriz de transformación *A*:

$$\mathcal{L}(\{\mathbf{x}_i\}|A) = \sum_{j=1}^J \frac{N_j}{2} \log\left(\frac{|A|}{(2\pi)^d \prod_{t=1}^p \mathbf{a}_t \hat{\Sigma}^{(j)} \mathbf{a}_t{}^{\mathrm{T}} \prod_{t=p+1}^d \mathbf{a}_t \hat{\Sigma} \mathbf{a}_t{}^{\mathrm{T}}}\right) - \frac{Nd}{2} \quad (3.10)$$

La transformación que maximiza la ecuación 3.10 es la transformación óptima HLDA que buscamos. A diferencia de LDA, la matriz de transformación HLDA no se puede resolver analíticamente y la solución no es única. Kumar [Kumar97] emplea técnicas de optimización numéricas para solucionar el problema. Burget [Burget04b], en cambio, para calcular la matriz de transformación *A*, aplica un algoritmo de optimización iterativo propuesto por Gales [Gales99] que implementa el algoritmo EM. Mediante la ecuación 3.9 y la matriz de transformación *A* estimada en ese instante, Gales estima las varianzas $\{\hat{\bar{\boldsymbol{\sigma}}}^{(j)^2}\}$ de los datos transformados. A continuación, a diferencia de Kumar, reemplaza la estimación de las medias en la ecuación 3.7 (Kumar reemplaza las medias y las varianzas) y representa la verosimilitud de los datos con las varianzas estimadas $\{\hat{\bar{\boldsymbol{\sigma}}}^{(j)^2}\}$:

$$\mathcal{L}\left(\{\mathbf{x}_i\}\middle|A, \{\hat{\bar{\boldsymbol{\sigma}}}^{(j)^2}\}\right) = \frac{1}{2}\left(\log\left(\left(\mathbf{c}_i \mathbf{a}_i{}^{\mathrm{T}}\right)^2\right) - \sum_{j=1}^J N_j \log\left((2\pi)^d \prod_{t=1}^d \bar{\sigma}_t^{(j)^2}\right) - \sum_{t=1}^d \mathbf{a}_t \, G^{(t)} \mathbf{a}_t{}^{\mathrm{T}}\right) \quad (3.11)$$

donde $\mathbf{c_i}$ es el vector fila i de la matriz cofactor $C = |A|\,A^{-1}$ para el valor actual de A, y $G^{(t)}$ es la matriz $d \times d$, que incluye las varianzas re-estimadas, para el parámetro transformado t:

$$G^{(t)} = \begin{cases} \sum_{j=1}^{J} \frac{N_j}{\hat{\boldsymbol{\sigma}}^{(j)^2}} \hat{\Sigma}^{(j)} & t \leq p \\ \frac{N}{\hat{\boldsymbol{\sigma}}^2} \hat{\Sigma} & t > p \end{cases} \tag{3.12}$$

siendo N_j el número de vectores de la clase j y N el número total de vectores.

Si igualamos a cero la derivada parcial de la ecuación 3.11 con respecto a a_i, obtenemos la máxima verosimilitud de la fila i de A para las varianzas $\{\hat{\boldsymbol{\sigma}}^{(j)^2}\}$ ya establecidas. Por tanto, las filas de la matriz A se re-calculan iterativamente por medio de la siguiente fórmula:

$$\hat{\mathbf{a}}_t = \mathbf{c}_t G^{(t-1)} \sqrt{\frac{N}{\mathbf{c}_t G^{(t-1)} \mathbf{c}_t^{\mathrm{T}}}} \tag{3.13}$$

Así pues, el proceso de optimización consiste en la re-estimación alternativa de las varianzas, $\hat{\boldsymbol{\sigma}}^{(j)^2}$, y la matriz de transformación A.

3.1.2 Algoritmos genéticos

En términos generales, la computación evolutiva es una disciplina que imita las estrategias de la evolución biológica para dar solución a problemas de búsqueda y optimización reales. La teoría de Darwin defiende que en la naturaleza, a lo largo de las generaciones, las poblaciones evolucionan acorde a los principios de selección natural y supervivencia de los más fuertes; los miembros de una población (incluso los miembros de una misma especie) compiten entre sí continuamente en la búsqueda de recursos naturales que satisfagan sus necesidades (el agua, la comida, etc). Aquellos individuos mejor adaptados tienen mayor probabilidad de generar descendientes. Por el contrario, los individuos con menor éxito en sobrevivir producirán un menor número de descendientes. Esto significa que los genes de los individuos mejor adaptados se propagan más a lo largo de las generaciones. Además, la combinación de dos progenitores debidamente adaptados al entorno natural puede dar lugar a descendientes aún con mayor capacidad de adaptación, debido a una mejora en la combinación de buenas características provenientes de los progenitores. De esta forma, las especies evolucionan generando individuos cada vez mejor adaptados al entorno en el que viven.

Los algoritmos genéticos (GA) emplean un procedimiento similar al de la evolución natural. Parten de una población de individuos, cada uno de los cuales representa una solución al problema dado. A cada individuo se le asigna una puntuación que determina su aptitud para resolver el problema, que de forma análoga, reflejaría su capacidad de competir por los recursos en la naturaleza. El algoritmo selecciona los individuos más fuertes, o mejor adaptados al problema, y a continuación mezcla, muta o mantiene sus representaciones para generar descendientes. Se produce una nueva población de nuevas soluciones, que reemplaza la población anterior y contiene una mayor proporción de individuos mejor adaptados. De esta forma, se propagan las mejores características de los individuos de generación en generación, convergiendo hacia la solución óptima del problema.

En la **Figura 3.1** se muestra el funcionamiento del denominado Algoritmo Genético Simple (*Simple Genetic Algorithm, SGA*) [Goldberg89]. Goldberg definió la especificación del SGA y fue uno de los primeros en analizar la aplicabilidad de los GA. SGA (como cualquier otro GA) requiere una codificación adecuada de los individuos, representados por "cromosomas", que definen las posibles soluciones al problema. Generalmente los individuos consisten en un conjunto de parámetros denominados "genes". Habiendo determinado la representación de los individuos, se genera la población inicial, aleatoriamente o mediante una semilla. Después, el GA, de forma iterativa, conduce a la población hacia un punto óptimo de acuerdo a una determinada métrica (función *fitness* o de evaluación) que mide la eficiencia o adecuación de los individuos para llevar a cabo una tarea concreta. Una de las principales ventajas de los GA sobre otras técnicas de búsqueda heurísticas consiste en que no realizan ninguna hipótesis sobre las propiedades de la función de evaluación, pudiendo incluso definir y emplear funciones de evaluación multi-objetivo de forma completamente natural [Oliveira03][Yang98]. Una vez evaluados todos los candidatos, se seleccionan algunos de ellos, normalmente los más fuertes o adecuados, es decir, aquellos que ocupan los primeros puestos de acuerdo a la función de evaluación. Los genes de los padres seleccionados se combinan utilizando habitualmente operadores de cruce y mutación para producir la población de la siguiente generación. A veces, se permite que ciertos individuos (generalmente los más fuertes) se transmitan sin mezclar ni mutar a la siguiente generación, lo que se conoce en la terminología específica como "elitismo". Si el algoritmo se configura debidamente, la adecuación media de los individuos, así como la adecuación del mejor individuo, aumentará de generación en generación, convergiendo hacia el óptimo global. Los procesos de evaluación y reproducción se repiten hasta que se cumpla algún criterio de convergencia, que generalmente consiste en un número máximo de generaciones o un umbral de mejora. Una vez finalizado, el algoritmo devuelve como solución global el mejor individuo.

La evolución de un GA depende de un conjunto de operadores, denominados operadores genéticos, que se encargan de la reproducción de los individuos. Consisten en tres mecanismos básicos: la selección, el cruce y la mutación. En cuanto a la selección, hay tres métodos muy comunes:

- La selección proporcional, también denominado selección por ruleta, favorece a los individuos más fuertes, ya que la probabilidad de selección es proporcional a la adecuación del individuo. Es decir, la probabilidad de selección de cada individuo es proporcional a la magnitud de su valor de adecuación.
- La selección por rango, vuelve a dar prioridad a los individuos más fuertes, pero esta vez a través de un ranking numérico que mide la adecuación de cada individuo. La selección depende únicamente de este ranking, y no de las diferencias absolutas de adecuación.
- La *selección por torneo*, consiste en elegir aleatoriamente un subconjunto de individuos y escoger a uno de ellos, generalmente al más apto.

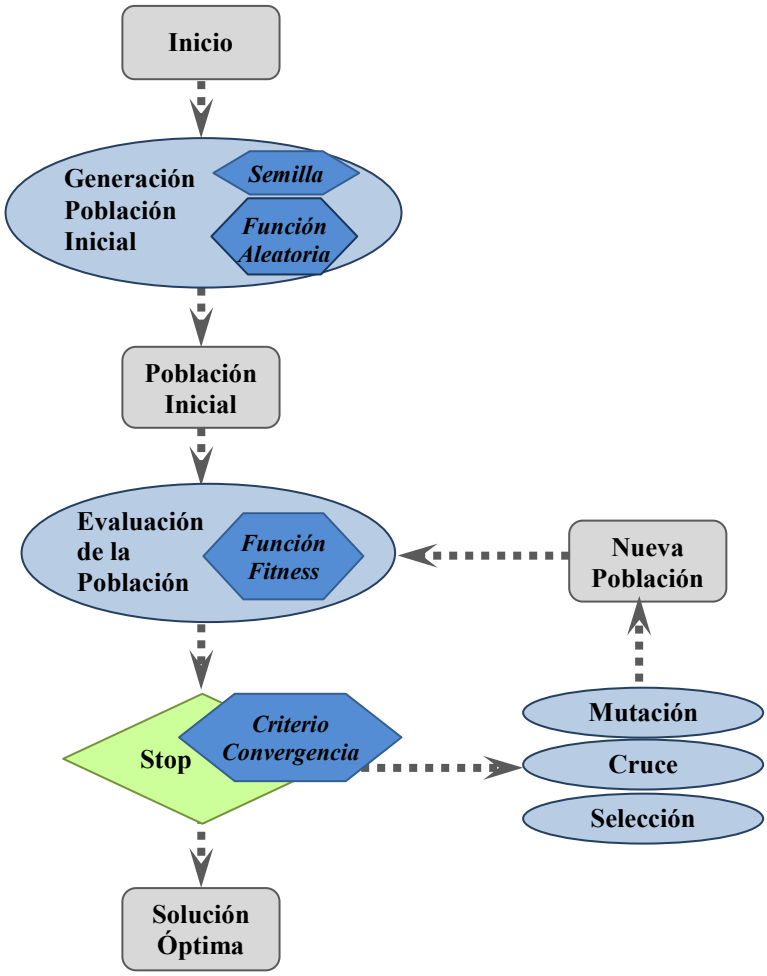

Figura 3.1. *Funcionamiento de un Algoritmo Genético Simple (SGA)*

Una vez realizada la selección, se cruzan los individuos seleccionados con objeto de intercambiar material cromosómico. El método de cruzamiento más común consiste en definir n puntos de corte, que determinan $n+1$ segmentos en cada individuo, e intercambiar segmentos alternativos para producir los individuos hijos. A veces el operador de cruce no se aplica a todos los pares de individuos que han resultado seleccionados, sino que se aplica de forma aleatoria con una probabilidad comprendida entre 0.5 y 1.0.

Por último, una mutación consiste en invertir el valor de ciertos bits tomados al azar en la representación de un individuo. La llamada tasa de mutación establece la probabilidad de que tras un cruzamiento tengan lugar mutaciones. Normalmente se le asigna una probabilidad pequeña para evitar que el algoritmo se convierta en una búsqueda aleatoria. La mutación garantiza el análisis de una región amplia del espacio de búsqueda. Se introduce para evitar que la búsqueda quede atrapada en extremos locales; ofrece diversidad y restaura la variedad genética que haya podido perderse a lo largo de las generaciones.

Por otro lado, la especificación de la función de evaluación es un elemento fundamental, puesto que debe determinar la verdadera adecuación de una solución, y de esta forma guiar la búsqueda por los caminos más adecuados. Si la función se define de forma errónea, inexacta o inconsistente, el

algoritmo podría ser incapaz de solucionar el problema o podría acabar resolviendo un problema distinto.

En la bibliografía pueden encontrarse varios trabajos que aplican GA para la extracción y selección de parámetros en tareas de reconocimiento de locutores:

- En [Charbuillet06] se aplica un GA en la extracción de parámetros, concretamente para definir el banco de filtros más adecuado para una tarea de diarización de locutores. Es decir, en vez de emplear el banco de filtros estándar (un banco de filtros triangulares en la escala de Mel, como en la mayoría de los front-end acústicos en reconocimiento de locutores), proponen el uso de un banco de filtros específico optimizado por el GA sobre un corpus de desarrollo. El algoritmo trata de optimizar el ancho de banda y la frecuencia central. En un trabajo posterior, los mismos autores han realizado un estudio parecido, pero orientado a la verificación de locutores [Charbuillet07]. En este último caso, el módulo de extracción de parámetros fusiona tres procedimientos estándares. El algoritmo determina los valores óptimos del número de filtros, la escala del banco de filtros y el número de coeficientes cepstrales de cada procedimiento de extracción. Incluye una población por procedimiento que evoluciona de forma aislada.
- Demirekler y Haydar, aplican un GA para la selección de parámetros acústicos en la identificación de locutores [Demirekler99]. Los experimentos realizados buscan la reducción del vector acústico original (compuesto por 12 LPCC y sus derivadas), obteniendo un conjunto de parámetros óptimo para cada locutor. En una primera fase, se analizan conjuntos de la misma cardinalidad para todos los locutores (conjuntos de 5, 6, 7, 8 o 10 parámetros), y a continuación el algoritmo busca la cardinalidad óptima para cada locutor. La capacidad de discriminación de un conjunto de parámetros se evalúa frente a los dos locutores más próximos (no frente a todos los locutores).
- Charlet y Douvet, proponen la optimización del vector acústico en un sistema de verificación de locutores *text-dependent* que emplea un modelo HMM para la clasificación [Charlet97]. El proceso de optimización busca el subconjunto de parámetros acústicos que proporciona el error mínimo sobre un corpus de desarrollo. Para ello, analizan varias técnicas heurísticas de selección de parámetros, entre las que se encuentra un GA. El GA busca el pesado óptimo en la emisión de probabilidades, con pesos que adquieren valores booleanos ($\alpha \in \{0,1\}$) o valores reales positivos ($\alpha \in R^+$). El pesado realizado con valores booleanos es una metodología de selección. El pesado realizado con valores reales, en cambio, es una metodología de optimización.
- Kinnunen y Fränti analizan el pesado de vectores acústicos para aumentar la capacidad de discriminación de un sistema de identificación de locutores que emplea modelos VQ [Kinnunen01]. El pesado se aplica en el proceso de emparejamiento mediante la asignación de mayores pesos a los vectores del codebook que muestran una mayor capacidad de discriminación. Concretamente, el peso depende de la distancia mínima a otros codebooks.
- En el ámbito de la aplicación de GA en el reconocimiento de locutores, Hong y Kwong han propuesto el empleo de un GA para entrenar modelos de locutor GMM(EM) [Hong2005]. La

metodología de entrenamiento es híbrida dado que combina la capacidad de búsqueda del GA y la efectividad del método de estimación de parámetros de los GMM.

3.2 Experimentos de reducción de dimensionalidad mediante transformaciones lineales

En este apartado se analiza el rendimiento de PCA, LDA y HLDA para eliminar correlaciones y reducir la dimensionalidad [Zamalloa08a] [Zamalloa08c]. El marco experimental consiste en la identificación de locutores en un conjunto cerrado.

3.2.1 Configuración experimental

La experimentación con transformaciones lineales se ha llevado a cabo sobre las bases de datos de habla castellana Albayzín y Dihana con el objetivo de comparar el rendimiento de estas metodologías con distintos tipos de habla que provienen de canales diferentes. Cada base de datos se ha dividido en dos subconjuntos disjuntos:

(1) Un conjunto de entrenamiento sobre el que se calculan las matrices de transformación y los modelos de locutor,

(2) Un conjunto de test sobre el que se evalúa el rendimiento de la metodología en cuestión.

Albayzín consta de 204 locutores y cada locutor contiene, como mínimo, 25 señales. El conjunto de entrenamiento está compuesto por 10 de estas 25 señales, y el conjunto de test por otras 8, lo que suma 2040 señales de entrenamiento y 1632 de test. Dihana en cambio, recoge el habla de 225 locutores, donde cada locutor aporta 4 diálogos y 16 señales leídas. El conjunto de entrenamiento está compuesto por 2 diálogos y 8 frases leídas por locutor, y el de test por 1 dialogo y 4 frases leídas. En total, los conjuntos de entrenamiento y test contienen 4598 y 2897 señales respectivamente. El resto de señales se ha reservado para el conjunto de desarrollo, empleado por los GA para el proceso de optimización.

El conjunto original de parámetros depende de las condiciones acústicas del corpus. El vector original *d*-dimensional está compuesto por 39 parámetros en Albayzín y 33 en Dihana:

- En Albayzín, el habla se analiza en tramos de 25 milisegundos solapados a intervalos 10 milisegundos. Después, se aplica una ventana de *Hamming* y se calcula una FFT de 512 puntos. Las amplitudes de la FFT se promedian en 24 filtros triangulares solapados, donde las frecuencias centrales y anchos de banda se definen de acuerdo a la escala Mel. A continuación, se aplica la DCT sobre el logaritmo de las amplitudes del filtro y se obtienen 12 MFCC. Para aumentar la robustez frente al ruido del canal, se aplica CMS. Por último, se estiman la energía y las primeras y segundas derivadas de los parámetros MFCC y de la energía.
- La parametrización de Dihana, con respecto Albayzín, difiere en la cantidad de puntos de la FFT (que en Dihana es de 256 puntos) y en la cantidad de filtros (que en Dihana es de 20). Se obtienen

10MFCC, sobre los que se aplica normalización CMS, y se calculan sus primeras y segundas derivadas, la energía y sus derivadas.

En la siguiente **Figura 3.2** se resumen las características principales de las particiones creadas.

Particiones para la experimentación con transformaciones lineales			
Albayzín		**Dihana**	
204 locutores		225 locutores	
25 señales/locutor		4 diálogos /locutor 16 frases leídas/locutor	
d=39 *(12MFCC+E+Δ+ΔΔ)*		d=33 *(10MFCC+E+Δ+ΔΔ)*	
Entrenamiento	**Test**	**Entrenamiento**	**Test**
10 señales/locutor	8 señales/locutor	2 diálogos/locutor 8 frases leídas/locutor	1 diálogos/locutor 4 frases leídas/locutor
2040 señales	1632 señales	4598 señales	2897 señales

Figura 3.2. *Representación esquemática de las particiones creadas sobre Albayzín y Dihana para la experimentación con transformaciones lineales*

Con el objetivo de establecer un compromiso entre el coste computacional y el rendimiento de clasificación, se ha analizado la reducción a diferentes subconjuntos de *k* parámetros, concretamente a $k = \{30, 20, 13, 12, 11, 10, 8, 6\}$. Los modelos de locutor son GMM(EM). En los experimentos con LDA y HLDA, se han utilizado también modelos GMM(MAP) estimados a partir de un UBM.

Por otro lado, PCA se ha implementado mediante la herramienta *LNKnet* [Lippmann93]. Es un paquete de software de código abierto, elaborado en el MIT Lincoln Laboratory con herramientas básicas para implementar aplicaciones de clasificación de patrones mediante métodos estadísticos, redes neuronales y aprendizaje automático. LDA y HLDA en cambio, se han implementado mediante la herramienta de software libre *FoCal* desarrollada por Niko Brummer [Brummer]. *FoCal* proporciona varios módulos en MATLAB para la fusión y calibración de los sistemas de detección automática de locutores. Estos módulos incluyen varias funciones, entre las que se encuentran LDA y HLDA.

3.2.2 Resultados

La experimentación consta de varias etapas:

(1) Empieza con la estimación de la matriz de transformación, utilizando para ello todos los datos del conjunto de entrenamiento.

(2) Una vez estimada la matriz, se transforma todo el corpus, tanto el conjunto de entrenamiento como el conjunto de test.

(3) El espacio transformado se reduce a *k* parámetros, para cada tamaño *k*.

(4) Por último, para cada tamaño *k*, se estiman los modelos de locutor y se clasifican las señales del conjunto de test.

* Debido a que PCA y LDA ordenan los parámetros en función de su relevancia, la selección de los k parámetros más relevantes consiste en la selección de los k primeros parámetros transformados. En el caso de HLDA se estiman 8 matrices de transformación, porque la transformación es distinta para cada valor de k.

Cabe señalar que se ha calculado el error promedio y el intervalo de confianza del 95% en todos los experimentos realizados, para evaluar la significación estadística de los resultados obtenidos. La necesidad de calcular estos intervalos está motivada por la inicialización aleatoria de los modelos GMM. En experimentos preliminares se ha observado que con un mismo conjunto de parámetros y datos, los parámetros de los modelos GMM, tras la convergencia, varían ligeramente debido a la inicialización aleatoria de los modelos. En consecuencia, el rendimiento de clasificación también fluctúa ligeramente. Esta incertidumbre intrínseca se puede medir por medio de la estimación de intervalos de confianza del error promedio, asumiendo que la distribución de las tasas de error es gaussiana. Para calcular el error promedio y su intervalo de confianza del 95%, el entrenamiento de los modelos y la posterior clasificación de las señales de test se han realizado 20 veces para cada conjunto de parámetros. En consecuencia, dados dos conjuntos A y B, si el rendimiento de A es mejor que el de B (porque el error promedio de A es inferior al de B) pero sus intervalos están solapados, no se puede considerar que el rendimiento de A sea superior al de B.

En este apartado se describe el rendimiento de los conjuntos reducidos mediante PCA, LDA y HLDA. Pero antes de seguir, en la **Tabla 3.1** se muestran las tasas de error obtenidas con el conjunto original de parámetros. Son de 0.2±0.03 en Albayzín, y 18.69±0.15 en Dihana.

Tabla 3.1. *Error promedio e intervalo de confianza del 95% obtenidos en experimentos de reconocimiento de locutores utilizando el conjunto original de parámetros en Albayzín y Dihana.*

Error Promedio ± Intervalo Confianza del 95 %		
Conjunto Original (MFCC+E+Δ+ΔΔ)	***Albayzín***	***Dihana***
	0.2 ± 0.03	18.69 ± 0.15

3.2.2.1 Resultados con PCA

Los resultados obtenidos con PCA se detallan en la **Tabla 3.2**. Se observa que los conjuntos PCA con $k \geq 20$ mejoran el rendimiento del conjunto original de parámetros (véase la tabla 3.1), lo que verifica la validez de PCA para la reducción de dimensionalidad. A medida que se sigue reduciendo el conjunto de parámetros, el error promedio aumenta significativamente, sobre todo para los conjuntos más pequeños, con k=6 y k=8 componentes (véase la **Figura 3.3**). En Albayzín, los subconjuntos óptimos para k=20 y k=30 producen una tasa de reconocimiento muy alta, con un error promedio por debajo del 0.2%. Además, los intervalos de confianza de los conjuntos PCA con $k \geq 20$ y el conjunto original están solapados (para facilitar la comparación, en la Figura 3.3 el rendimiento de estos subconjuntos se muestra de forma ampliada en el recuadro interior).

Tabla 3.2. *Error promedio e intervalo de confianza del 95% de los experimentos de reconocimiento de locutores con PCA sobre Albayzín y Dihana, para k= 30, 20, 13, 12, 11, 10, 8 y 6 componentes. Se incluye la tasa del conjunto transformado d-dimensional (sin reducción), con d=39 y d=33 parámetros en Albayzín y Dihana.*

Error Promedio PCA ± Intervalo Confianza del 95 %		
k	**Albayzín**	**Dihana**
6	14.37±0.15	33.23±0.12
8	5.86±0.12	24.19±0.13
10	2.73±0.12	20.67±0.12
11	1.61±0.07	20.27±0.13
12	0.94±0.06	19.75±0.16
13	0.56±0.05	19.63±0.1
20	**0.19±0.02**	17.61±0.13
30	**0.15±0.03**	15.97±0.15
d (39,33)	**0.2±0.02**	**15.92±0.13**

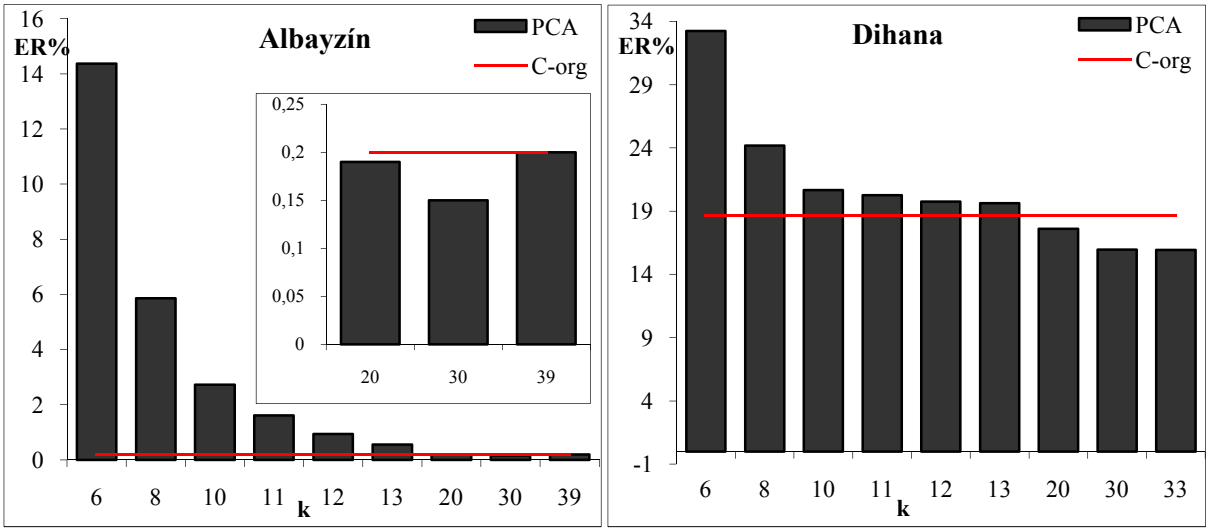

Figura 3.3. *Error promedio de los conjuntos PCA de tamaño k= d, 30, 20, 13, 12, 11,10, 8 y 6 en Albayzín (izquierda) y Dihana (derecha). Se incluye el resultado del conjunto original (C-org). Con objeto de facilitar su comparación, el rendimiento de los subconjuntos PCA de Albayzín de tamaño* $k \geq 20$ *se muestra de forma ampliada en el recuadro interior.*

Los resultados obtenidos sobre Dihana muestran un comportamiento similar. No obstante, la mejora obtenida con respecto al conjunto de datos no transformado es más notable, debido a que PCA puede compensar correlaciones debidas al ruido del canal. En el caso en que no se reduce el espacio, PCA aporta una mejora relativa del 14.82% con una tasa media del 15.92%, y la reducción a k=30 componentes una mejora del 14.55% con respecto a la tasa del espacio original. El conjunto PCA con k=20, con una tasa de error del 17.61%, sigue mejorando el rendimiento del conjunto original.

3.2.2.2 Resultados con LDA

En la **Tabla 3.3** se muestran los resultados obtenidos mediante LDA. En esta experimentación se han seguido dos aproximaciones que emplean diferentes técnicas de modelado de locutor. La primera aproximación (denominada *LDA_EM*) [Jin00], trata de separar el espacio acústico de cada locutor del resto de locutores, asumiendo que cada locutor comprende una clase compuesta por sus vectores de entrenamiento. En este caso, los modelos de locutor son GMM(EM). La segunda aproximación (denominada *LDA_MAP*) [Burget07][Yang06], en cambio, utiliza modelos de locutor GMM(MAP) estimados a partir de un UBM. El UBM contiene un gran número de componentes gaussianas debido a que trata de cobertura a todo el espacio acústico. Se asume que cada componente representa la densidad de una unidad acústica del conjunto de datos con el que se ha estimado (en este caso el UBM se calcula con todo el conjunto de entrenamiento). En la aproximación LDA_MAP cada componente constituye una clase, y al separar estas componentes gaussianas, se aumenta la capacidad de discriminación de cada unidad acústica del espacio (porque se supone que disminuyen los solapamientos). Debido a que los modelos GMM(MAP) modelan el desplazamiento de cada componente gaussiana, se asume que LDA_MAP optimiza la cuantificación de los desplazamientos. Para calcular la matriz de transformación, una vez estimado el UBM, cada vector del conjunto de entrenamiento se asocia con la gaussiana más probable.

Tabla 3.3. *Error promedio e intervalo de confianza del 95% de los experimentos de reconocimiento de locutores con LDA sobre Albayzín y Dihana, para k= 30, 20, 13, 12, 11, 10, 8 y 6 componentes. Se analizan dos aproximaciones:(1) la discriminación de locutores con modelos GMM(EM) (LDA_EM) y (2) la discriminación de las unidades acústicas del UBM mediante modelos GMM(MAP) (LDA_MAP). Se incluye la tasa del conjunto transformado d-dimensional (sin reducción), con d=39 en Albayzín y d=33 en Dihana.*

Error Promedio LDA ± Intervalo Confianza del 95 %				
k	**ALBAYZÍN**		**DIHANA**	
	LDA_EM	**LDA_MAP**	**LDA_EM**	**LDA_MAP**
6	54.48±0.26	25.95±0.2	72.68±0.21	61.31±0.13
8	37.76±0.27	15.22±0.15	60.33±0.25	49.34±0.13
10	23.3±0.28	9.78±0.14	53.63±0.2	39.51±0.12
11	17.52±0.3	6.15±0.1	51.34±0.27	34.18±0.13
12	16.02±0.3	4.64±0.08	49.22±0.34	29.96±0.11
13	13.15±0.29	3.4±0.09	47.96±0.18	26.64±0.12
20	2.69±0.12	0.45±0.04	33.71±0.23	16.71±0.09
30	0.75±0.07	0.1±0.02	18.44±0.18	13.19±0.09
d (39,33)	**0.29±0.04**	**0.06±0.0**	**15.61±0.16**	**12.43±0.09**

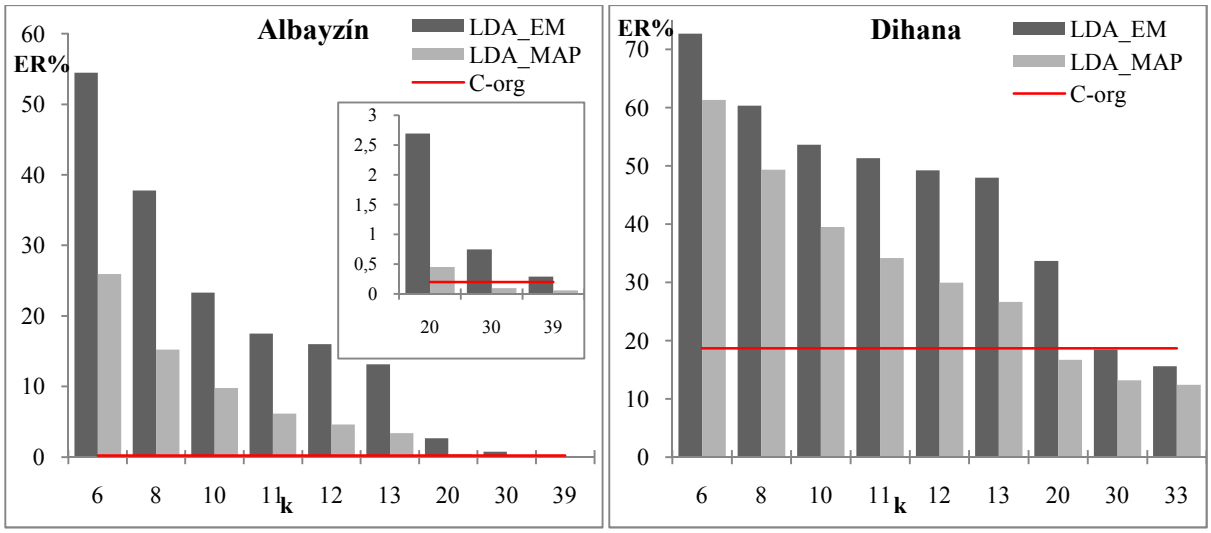

Figura 3.4. *Error promedio de los conjuntos LDA de tamaño k= d, 30, 20, 13, 12, 11,10, 8 y 6 en Albayzín (izquierda) y Dihana (derecha). Se incluye el resultado del conjunto original (C-org).Con objeto de facilitar su comparación, el rendimiento de los subconjuntos LDA de Albayzín de tamaño* $k \geq 20$ *se muestra de forma ampliada en el recuadro interior.*

Los resultados revelan que el rendimiento de LDA_MAP es superior al de LDA_EM, con mejoras superiores al 50% en la mayoría de los casos (véase la **Figura 3.4**). Como era de esperar, se observa que la reducción de parámetros conlleva el incremento de la tasa error, sobre todo cuando el conjunto de parámetros se reduce a menos de 30 componentes en LDA_EM, y a menos de 20 en LDA_MAP. Los resultados parecen indicar que la aproximación LDA_MAP es más apropiada para la reducción de dimensionalidad, puesto que muestra tasas más estables.

En Albayzín, el conjunto LDA_EM *d*-dimensional, con un error promedio de 0.29% no reduce la tasa de error del espacio original (0.2%, véase la tabla 3.1). Se cree que puede ser debido al escaso margen de mejora en los experimentos de identificación sobre Albayzín, por lo que resulta difícil realizar una comparación significativa. LDA_MAP, en cambio, mejora el resultado original, y de forma especial para el espacio *d*-dimensional transformado con el que prácticamente obtiene un error promedio del 0%.

No obstante, los resultados obtenidos sobre Dihana, de canal telefónico, muestran un comportamiento muy distinto. Se observan mejoras con las dos aproximaciones LDA_EM y LDA_MAP para valores k={30, 33} y k={20, 30, 33}, respectivamente. Se cree que estas mejoras podrían ser debidas a la compensación de las correlaciones producidas por del ruido de canal. Es especialmente notable el comportamiento de los conjuntos LDA_MAP con k=33 y k=30, que proporcionan una mejora relativa del 33.49% y el 29.42% con respecto al error promedio del espacio original. Los conjuntos LDA_EM alcanzan mejoras más moderadas, concretamente del 16.48% para k=33 y el 1.34% para k=30. Por tanto, los resultados obtenidos apuntan a que LDA_MAP, y en menor escala LDA_EM, son aproximaciones eficientes para eliminar las correlaciones y reducir de forma moderada la dimensionalidad del espacio.

Además, los resultados obtenidos ponen de manifiesto la importancia de definir adecuadamente las clases a discriminar. La aproximación LDA_EM trata de maximizar la discriminación entre los espacios acústicos de cada locutor, de forma independiente a la técnica de modelado que se aplica a continuación. LDA_MAP, sin embargo, que proporciona resultados notablemente mejores, trata de aumentar la capacidad de discriminación entre las unidades o densidades sobre las que se calcularán los modelos.

3.2.2.3 Resultados con HLDA

Los resultados obtenidos con HLDA se detallan en la **Tabla 3.4**. HLDA se define como una generalización de LDA, por lo que las aproximaciones analizadas en esta experimentación son las mismas que en LDA (véase el apartado 3.3.2.2): HLDA_EM, que equivale en diseño a LDA_EM (modelos GMM(EM), una clase por locutor) y HLDA_MAP, que sigue el mismo procedimiento que LDA_MAP (modelos GMM(MAP), una clase por cada componente del UBM).

Tabla 3.4. *Error promedio e intervalo de confianza del 95% de los experimentos HLDA sobre Albayzín y Dihana, para k= 30, 20, 13, 10 y 6 componentes. Se analizan dos aproximaciones:(1) la discriminación de locutores con modelos GMM(EM)(HLDA_EM) y (2)la discriminación de las unidades acústicas del UBM mediante modelos GMM(MAP)(HLDA_MAP). Se incluye la tasa del conjunto transformado d-dimensional (sin reducción), con d=39 en Albayzín y d=33 en Dihana.*

Error Promedio HLDA ± Intervalo Confianza del 95 %				
k	***ALBAYZÍN***		***DIHANA***	
	HLDA_EM	**HLDA_MAP**	**HLDA_EM**	**HLDA_MAP**
6	19.43±0.22	30.33±0.19	46.58±0.15	67.8±0.15
10	2.13±0.09	11.58±0.13	25.61±0.15	44.51±0.12
13	0.92±0.06	5.8±0.1	19.74±0.16	32.82±0.14
20	0.23±0.02	0.95±0.06	15.06±0.11	20.71±0.12
30	**0.17±0.03**	0.36±0.02	14.78±0.12	13.66±0.08
d (39,33)	0.24±0.04	**0.06±0.01**	**14.72±0.14**	**11.44±0.11**

Las tasas de error obtenidas sobre el conjunto de test parecen indicar que HLDA_EM es más apropiado que HLDA_MAP para los conjuntos reducidos, excepto para la reducción a 30 componentes en Dihana (véase la **Figura 3.5**). Asimismo, véase en la tabla cómo los subconjuntos HLDA_EM *k*-dimensionales de Albayzín proporcionan intervalos de confianza solapados para $k=\{d, 30, 20\}$ y el conjunto original de parámetros (0.2%, véase la tabla 3.1). En Dihana se repite la misma tendencia; los intervalos de confianza de los conjuntos HLDA_EM para $k=\{d, 30\}$ están también solapados. En este último caso, los conjuntos HLDA_EM *k*-dimensionales con $k=\{d, 30, 20\}$ mejoran el rendimiento del espacio original (18.69%, véase la tabla 3.1). La mejora obtenida con sobre Dihana es significativa (concretamente del 21.14% para $k=d$). Tal como se ha mencionado anteriormente, esta mejora podría ser debida a la compensación de canal que mediante la transformación de parámetros efectúa HLDA.

En cuanto a HLDA_MAP, al reducir la dimensionalidad a $k<d$ parámetros, se observan incrementos de error mucho mayores que en HLDA_EM (sobre todo para los valores más pequeños de k). Se cree que el empeoramiento podría ser debido a que la reducción de dimensionalidad dificulta la estimación de las densidades del UBM (que hipotéticamente definen las unidades acústicas), dando lugar a la estimación de componentes que no representan las unidades originales. De todas formas, queda pendiente analizar con detalle en trabajos futuros las razones de este comportamiento. No obstante, cabe señalar que HLDA_MAP presenta muy buen rendimiento cuando no se reduce la dimensionalidad del espacio. En Albayzín, el error promedio baja prácticamente a 0%, mientras que en Dihana pasa del 18.69% al 11.44% (con una reducción relativa del 38.79%). En ambos casos, HLDA_MAP ofrece un rendimiento mejor que HLDA_EM, especialmente en Dihana donde HLDA_MAP ha proporcionado una mejora del 17.54% respecto a HLDA_EM.

En resumen, los resultados obtenidos apuntan a que HLDA_EM es una metodología adecuada para eliminar las correlaciones y si se precisa, reducir ligeramente la dimensionalidad.

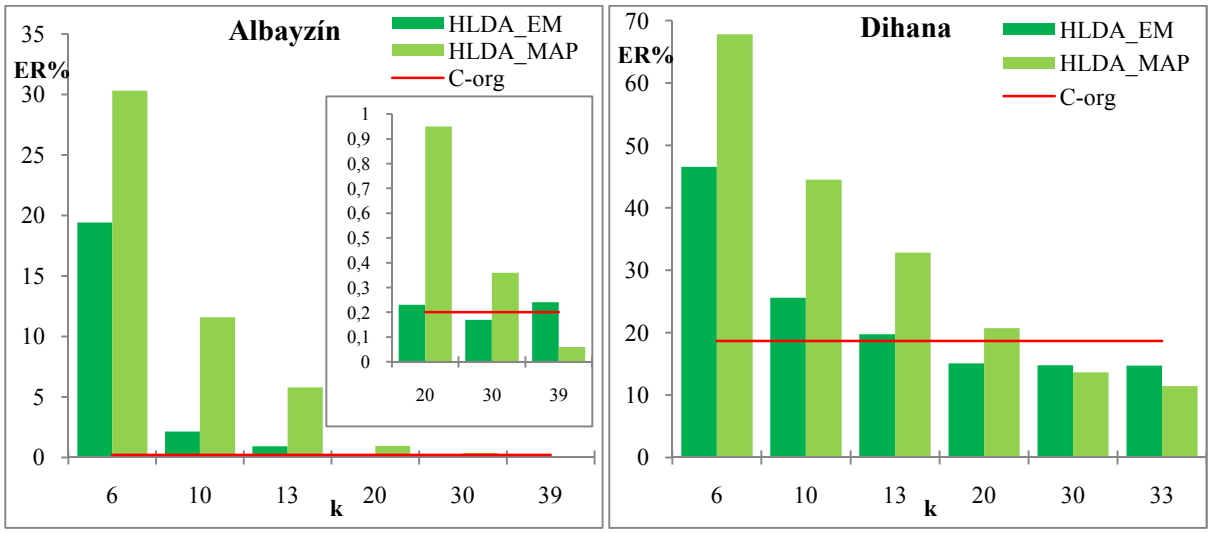

Figura 3.5. *Error promedio de los conjuntos HLDA de tamaño k= d, 30, 20, 13, 12, 11,10, 8 y 6 en Albayzín (izquierda) y Dihana (derecha). Se incluye el resultado del conjunto original (C-org).Con objeto de facilitar su comparación, el rendimiento de los subconjuntos HLDA de Albayzín de tamaño* $k \geq 20$ *se muestra de forma ampliada en el recuadro interior.*

3.2.3 Conclusiones

Los resultados obtenidos revelan que PCA, LDA y HLDA son procedimientos adecuados para eliminar correlaciones y reducir, de forma moderada, la dimensión del espacio en tareas de reconocimiento de locutores.

En la **Figura 3.6** se muestra la comparación de las metodologías citadas sobre Albayzín. Además de PCA, se han analizado dos aproximaciones, LDA_EM y LDA_MAP, para LDA, y otras dos, HLDA_EM y HLDA_MAP, para HLDA. Se observa que los conjuntos d-dimesionales obtenidos por LDA_MAP y HLDA_MAP, prácticamente reconocen correctamente el locutor en el 100% de las señales. LDA y HLDA maximizan el ratio entre la varianza externa y la varianza interna bajo la

hipótesis de que la distribución de las clases es gaussiana. LDA asume que todas las clases tienen la misma varianza interna, a diferencia de HLDA, que relaja esta condición. Por tanto, a priori, HLDA debería ser superior a LDA, tal como sucede en las aproximaciones LDA_EM y HLDA_EM. No obstante, aunque el rendimiento de LDA_MAP y HLDA_MAP es óptimo con el conjunto *d*-dimensional, a medida que se va reduciendo el conjunto de parámetros, LDA_MAP presenta mejor rendimiento que HLDA_MAP. Tal como se ha mencionado anteriormente se desconoce la razón de este comportamiento y es un tema de investigación interesante a analizar con detalle en trabajos futuros. Por otro lado, como cabía esperar, el rendimiento de HLDA_EM es superior al de LDA_EM y LDA_MAP para *k*<*d*. En cuanto a PCA, presenta el mejor rendimiento en la reducción a conjuntos de dimensión pequeña, concretamente de tamaño $k \leq 20$.

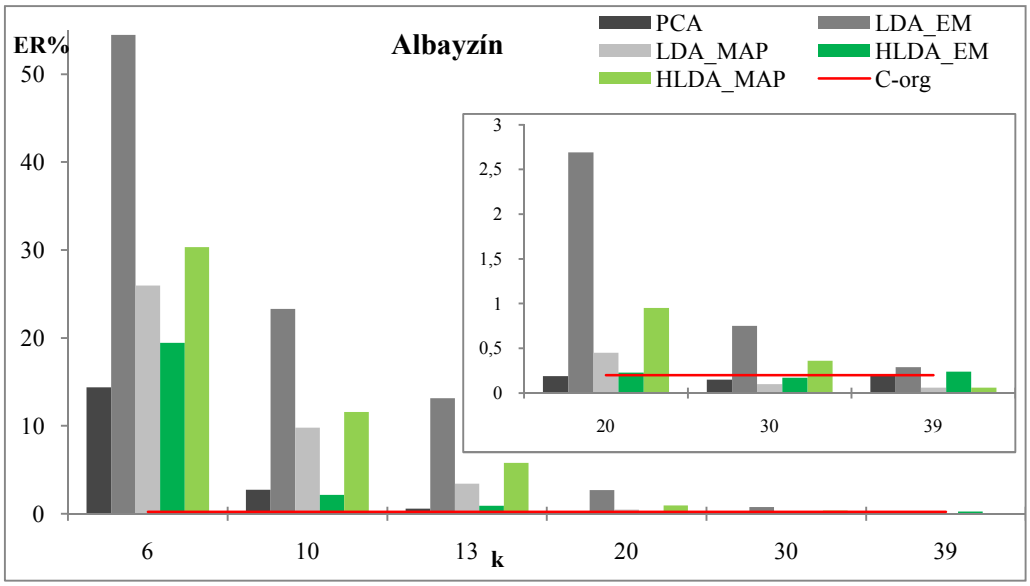

Figura 3.6. *Error promedio de los conjuntos PCA, LDA_MAP, HLDA_EM y HLDA_MAP de tamaño k= 39, 30, 20, 13,10 y 6 componentes en Albayzín. Se incluye el resultado del conjunto original (C-org). Con objeto de facilitar su comparación, el rendimiento de los subconjuntos de tamaño* $k \geq 20$ *se muestra de forma ampliada en el recuadro interior.*

En Dihana (véase **Figura 3.7**) se pueden extraer conclusiones similares: en los conjuntos *d*-dimensionales HLDA obtiene mejores resultados que LDA, pero a medida que se va reduciendo la dimensionalidad, el rendimiento de LDA_MAP es superior al de HLDA_MAP, y a su vez inferior a HLDA_EM. De nuevo, para un número limitado de componentes ($k \leq 13$) los conjuntos PCA obtienen tasas de acierto superiores a las del resto de transformaciones lineales.

Por tanto, los resultados obtenidos apuntan a que PCA y HLDA_EM son más eficientes que LDA_MAP y HLDA_MAP para reducir la dimensionalidad. Pero si la transformación se aplica con el objetivo de eliminar las correlaciones y/o mejorar el rendimiento de clasificación (prácticamente sin reducir la dimensionalidad), LDA_MAP y HLDA_MAP son más eficaces. En la Figura 3.7 se observa que todos los subconjuntos transformados *k*-dimensionales en Dihana con $k \geq 20$ mejoran el rendimiento del conjunto original de parámetros (a excepción de los conjuntos LDA_EM y

HLDA_MAP de 20 dimensiones), obteniendo una mejoría notable en la mayoría de los casos. Tal como se ha mencionado varias veces a lo largo de este apartado, se cree que puede ser debido a que estas transformaciones compensan las correlaciones producidas por el ruido del canal. Esto explica que en Albayzín (Figura 3.6), grabada en condiciones de laboratorio, las transformaciones no hayan producido mejoras tan destacadas. Hay que tener en cuenta que las tasas de error para los conjuntos de Albayzín de más de 20 componentes son inferiores al 1%, con un margen de mejora muy limitado, por lo que es difícil extraer conclusiones significativas.

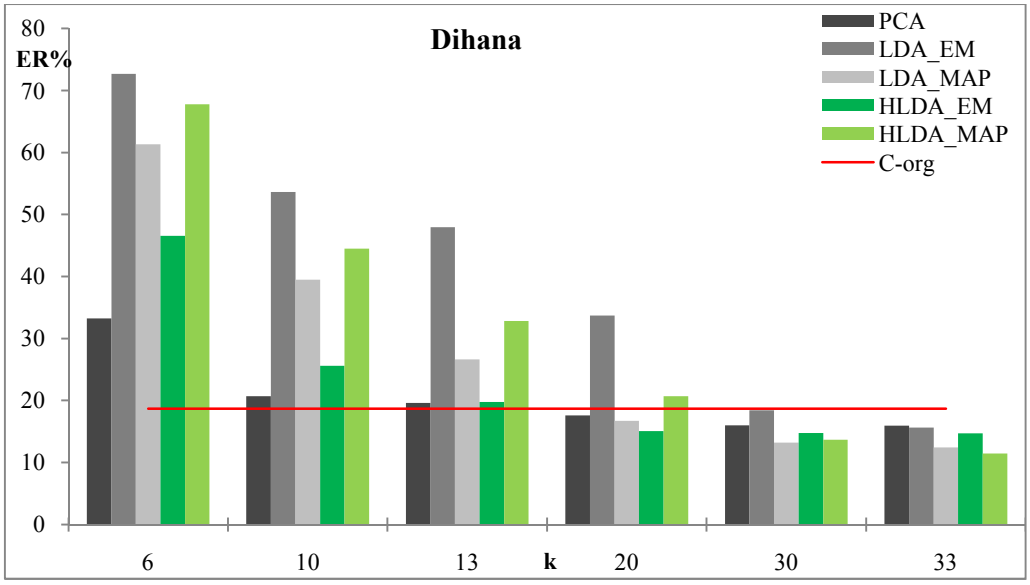

Figura 3.7. *Error promedio de los conjuntos PCA, LDA_MAP, HLDA_EM y HLDA_MAP de tamaño k= 33, 30, 20, 13, 10 y 6 en Dihana. Se incluye el resultado del conjunto original (C-org).*

3.3 Experimentos de reducción de dimensionalidad mediante GA

En este apartado se analiza el rendimiento de las aproximaciones con GA propuestas en la tesis para la búsqueda del subconjunto óptimo de parámetros acústicos. La búsqueda consiste en establecer cuáles son las $k<d$ componentes del vector acústico d-dimensional más relevantes desde el punto de vista de la clasificación de locutores. El marco experimental es el mismo que el de los experimentos con transformaciones lineales: la identificación de locutores de un conjunto cerrado.

Concretamente, se han propuesto dos aproximaciones para la selección de parámetros con:

(1) Un algoritmo puro de selección, donde el propio algoritmo determina el subconjunto óptimo de *k* parámetros acústicos [Zamalloa06b] [Zamalloa08b].

(2) Un algoritmo de selección mediante un ranking de pesos óptimos, donde el GA estima un conjunto de pesos estableciendo una modulación de los vectores acústicos (el pesado modula las dimensiones del espacio), asignando un peso mayor a los parámetros más relevantes, y un peso menor a los menos relevantes [Zamalloa06b] [Zamalloa06c]. En este caso, se seleccionan los *k* parámetros de mayor peso [Zamalloa06a]. Como el peso se puede interpretar como un ajuste de varianzas, las varianzas deben ponerse a 1 previamente.

3.3.1 Configuración experimental

La evaluación de las aproximaciones con GA se ha realizado de nuevo sobre Albayzín y Dihana, con objeto de verificar la validez de las metodologías y comparar la composición de los subconjuntos seleccionados en cada caso. Cada corpus se ha dividido en tres subconjuntos disjuntos que contienen señales del mismo conjunto de locutores:

(1) Un corpus de entrenamiento con el que se estiman los modelos de locutor (y el codebook cuando la experimentación utiliza modelos e-VQ)

(2) Un corpus de validación o desarrollo, sobre el que tiene lugar el proceso de búsqueda del GA.

(3) El conjunto de test, sobre el que se evalúa la solución óptima del GA.

En Albayzín (204 locutores, 25 señales como mínimo por locutor), de cada locutor se toman 25 señales, 10 de ellas para el corpus de entrenamiento, 7 para el de validación y 8 para el test. Así pues, el corpus de entrenamiento contiene 2040 señales, el de validación 1428 y el de test 1632. En Dihana (225 locutores, 4 diálogos y 16 frases leidas por locutor), en cambio, el conjunto de entrenamiento está compuesto por 2 diálogos y 8 frases leídas por locutor, y los corpus de validación y test, contienen, cada uno, 1 dialogo y 4 frases leídas. En total, el conjunto de entrenamiento, validación y test contienen 4598, 2379 y 2897 señales respectivamente. En la **Figura 3.8** se resumen las principales características de las particiones creadas para evaluar el rendimiento de las aproximaciones con GA.

Particiones para la experimentación con GA					
Albayzín			**Dihana**		
204 locutores			225 locutores		
25 señales/locutor			4 diálogos /locutor 16 frases leídas/locutor		
d=39 *(12MFCC+E+Δ+ΔΔ)*			d=33 *(10MFCC+E+Δ+ΔΔ)*		
Entrenamiento	**Validación**	**Test**	**Entrenamiento**	**Validación**	**Test**
10 señales/locutor	7 señales/locutor	8 señales/locutor	2 diálogos/locutor 8 frases leídas/locutor	1 diálogo/locutor 4 frases leídas/locutor	1 diálogo/locutor 4 frases leídas/locutor
2040 señales	1428 señales	1632 señales	4598 señales	2379 señales	2897 señales

Figura 3.8. *Representación esquemática de las particiones creadas sobre Albayzín y Dihana para evaluar el rendimiento de las aproximaciones con GA*

El conjunto original de d parámetros depende de las condiciones acústicas del corpus: está compuesto por 39 parámetros en Albayzín (12MFCC, la energía, y sus derivadas), y 33 en Dihana (10MFCC, la energía, y sus derivadas). La parametrización de las bases de datos se ha descrito en el apartado 3.3.1. Al igual que en la experimentación con transformaciones lineales, se han definido subconjuntos de k componentes para $k = \{30, 20, 13, 12, 11, 10, 8, 6\}$. Por otro lado, con el objetivo de determinar si el conjunto óptimo de parámetros, así como su rendimiento, dependen de la metodología de modelado se han empleado dos metodologías: modelos e-VQ sobre un codebook de

256 centroides (modelos discretos) y modelos GMM(EM) de 32 componentes con matrices de covarianza diagonales (modelos continuos).

En estos experimentos se utiliza como función de evaluación la tasa de reconocimiento de locutores sobre un corpus de desarrollo. Dicha función indica la adecuación de un individuo para la clasificación. El GA trata de obtener individuos que maximicen la tasa de acierto, y devuelve como solución el individuo que representa el subconjunto óptimo de parámetros (en la selección directa) o el individuo con los pesos óptimos para el espacio original *d*-dimensional (en la selección indirecta mediante un ranking de pesos óptimos). Con objeto de ajustar los parámetros que controlan el rendimiento y la convergencia del GA, se han llevado a cabo experimentos preliminares en los que se han fijado los siguientes valores:

- La población de algoritmo está compuesta por 120 individuos. Poblaciones demasiado extensas retrasan excesivamente la convergencia del algoritmo, mientras que poblaciones demasiado pequeñas limitan el rendimiento de búsqueda.
- Se ha comprobado también que 40 generaciones son suficientes para alcanzar la convergencia, por lo que no se ha aplicado ningún otro criterio de parada.
- A la hora de elegir qué individuos de la actual generación van a cruzarse para producir los individuos de la siguiente generación, se ha optado por la selección por torneo, con una pre-selección aleatoria de 7 y 2 individuos, entre los que se escoge el más adecuado según la función de evaluación.
- Se ha aplicado una mezcla con un punto de corte, es decir, cada uno de los dos individuos seleccionados se rompe en dos segmentos que se intercambian.
- Las tasas de mezcla y mutación se han fijado heurísticamente a aquellos valores que han producido mejores resultados: concretamente a 1.0 y 0.01, respectivamente. Es decir la mezcla se aplica sobre todos los pares de individuos seleccionados y la probabilidad de que haya una mutación es del 1%.
- Se ha aplicado el caso más básico de elitismo, manteniendo el mejor individuo de cada generación. Esto garantiza que la adecuación del mejor individuo mejore monótonamente en sucesivas generaciones.

Los GA se han implementado por medio de ECJ "*A Java-based Evolutionary Computation and Genetic Programming Research System*", una herramienta para el desarrollo de aplicaciones de computación evolutiva y programación genética, creada y mantenida en la universidad George Mason de Estados Unidos, que se ofrece bajo una licencia especial de código abierto [ECJ]. ECJ tiene una arquitectura flexible, permite definir representaciones arbitrarias, genomas de longitud fija y variable, métodos de optimización multi-objetivo y diversos operadores de selección.

A continuación se describen brevemente los dos procedimientos de selección de parámetros mediante GA: (1) la selección directa utilizando modelos e-VQ y GMM-EM; y (2) la selección indirecta mediante un ranking de pesos óptimos utilizando únicamente modelos e-VQ.

3.3.1.1 Algoritmo de selección directa

En esta aproximación el GA comienza generando aleatoriamente una población de individuos. Cada individuo es un vector de *d* enteros, **r** = (r_1, r_2, ..., r_d) y queda determinado por *k* de estas componentes. Es decir, los *k* valores más altos de **r** indican implícitamente qué componentes $\Gamma^{(k)}$ se están eligiendo:

$$\Gamma^{(k)} = \left\{ \{j_1, j_2, \dots, j_k\} \middle| r_{j_1} \geq r_{j_2} \geq \cdots \geq r_{j_k} \geq r_j, \forall j \notin \{j_1, j_2, \dots, j_k\} \right\} \quad (3.14)$$

Obviamente, un cierto conjunto de componentes puede ser representado por muchos vectores distintos, o lo que es igual, dado un cierto individuo **r**, pequeños cambios en su especificación no modifican la elección de componentes. Esta redundancia en la representación facilita que la evolución del algoritmo sea suave y converja al subconjunto óptimo $\hat{\Gamma}^{(k)}$, independientemente de la población inicial. Para reducir el tamaño de la representación y, por tanto, el coste computacional, se han utilizado tan sólo 8 bits por gen, es decir enteros en el rango [0, 255].

El GA evalúa cada candidato $\Gamma^{(k)}$, y después de unas cuantas generaciones, finalmente se queda con aquel subconjunto que recibe la mejor evaluación. Los pasos a seguir para la evaluación son (en los experimentos con GMM(EM), no se aplican los pasos 2 y 3):

(1) Cada vector de la base de datos es sustituido por el vector reducido, formado por las componentes enumeradas en $\Gamma^{(k)}$.
(2) Se genera un codebook reducido C', utilizando las componentes $\Gamma^{(k)}$ del codebook completo C
(3) Se etiquetan los corpus de entrenamiento y validación reducidos mediante C'
(4) Utilizando el corpus de entrenamiento reducido se estiman los modelos de locutor
(5) Se aplican los modelos de locutor para clasificar las señales del corpus de validación
(6) Se asigna al individuo $\Gamma^{(k)}$ la tasa de reconocimiento obtenida

Una vez se ha establecido el subconjunto óptimo de *k* componentes, $\hat{\Gamma}^{(k)} = \{\hat{j}_1, \hat{j}_2, \dots, \hat{j}_k\}$, éste se evalúa clasificando las señales del corpus de evaluación. En los experimentos con modelos e-VQ, para evaluar $\hat{\Gamma}^{(k)}$ se calcula un nuevo codebook, $C(\hat{\Gamma}^{(k)})$, con el corpus de entrenamiento reducido. Después, se etiquetan con él los vectores reducidos de los corpus de entrenamiento y test, se calculan los modelos de locutor con las etiquetas de entrenamiento y con estos modelos se clasifican las señales del corpus de test. En los experimentos con GMM(EM), con objeto de acelerar el proceso de evaluación del GA, los candidatos $\Gamma^{(k)}$ se evalúan mediante modelos GMM de 8 componentes gaussianas estimadas en 5 iteraciones. Esta configuración ha mostrado muy buen equilibrio entre coste y rendimiento en experimentos preliminares, y no se ha detectado ningún problema de convergencia. La evaluación del subconjunto óptimo $\hat{\Gamma}^{(k)}$, en cambio, se realiza estimando un GMM compuesto por un número mayor de componentes gaussianas (concretamente un GMM de 32 gaussianas estimado en 20 iteraciones).

3.3.1.2 Algoritmo de selección indirecta a través de un ranking de pesos óptimos

Esta aproximación depende del pesado de parámetros acústicos mediante GA que, tal como se ha mostrado en un trabajo anterior, puede mejorar el rendimiento de un sistema de reconocimiento de

locutores que emplea modelos discretos [Zamalloa06b]. El GA realiza la búsqueda de un conjunto adecuado de pesos $\mathbf{w} = (w_1, w_2, \dots, w_d)$ que transforma cada vector acústico $\mathbf{x} = (x_1, x_2, \dots, x_d)$ en un vector ponderado $\mathbf{x}' = (w_1 x_1, w_2 x_2, \dots, w_d x_d)$. Se asume que la mejora es debida a que los parámetros más relevantes reciben un mayor peso y viceversa, dado que todos los parámetros tienen la misma varianza. Según esta hipótesis, y suponiendo que los sub-espacios óptimos se pueden definir de forma incremental (es decir, el subconjunto óptimo de *k*+*c* parámetros ($1 \leq c < d - k$) incluye todos los parámetros del subconjunto óptimo de *k* parámetros), se puede definir un ranking de parámetros [Zamalloa06a]. Dado un conjunto de datos *d*-dimensional y un conjunto óptimo de pesos, $\hat{\mathbf{w}}$, el conjunto óptimo de *k* parámetros, $\hat{\Gamma}^{(k)}$, consiste en la selección de *k* componentes con mayor peso en $\hat{\mathbf{w}}$: $\hat{\Gamma}^{(k)} = \left\{ \{j_1, j_2, \dots, j_k\} \middle| w_{j_1} \geq w_{j_2} \geq \cdots \geq w_{j_k} \geq w_j, \forall j \notin \{j_1, j_2, \dots, j_k\} \right\}$.

El GA comienza generando aleatoriamente una población de individuos. Cada individuo *R* es un conjunto de *d* pesos representado por un vector de *d* enteros, **r**=*(r1, r2, ..., rd)*. De nuevo, se han utilizado 8 bits por gen que definen enteros en el rango [0, 255]. El gen r_i es el peso w_i. Teniendo en cuenta que los modelos de locutor son modelos e-VQ, cada conjunto de pesos candidato **w** se aplica a los vectores del conjunto de entrenamiento para calcular un codebook específico *C(***w***)*. Después, cada vector ponderado es sustituido por el índice del centroide más cercano en *C(***w***)*. Por último, utilizando las etiquetas del corpus de entrenamiento se estiman los modelos de locutor, y se asigna a **w** la tasa de acierto obtenida sobre el corpus de validación. El conjunto de pesos óptimo, $\hat{\mathbf{w}}$, es el que máximiza la tasa de reconocimiento sobre el corpus de validación, y éste se evalúa clasificando los vectores ponderados de las señales del corpus de test.

En experimentos preliminares se ha observado que el tamaño óptimo de la población del GA depende del espacio de búsqueda. Como el número de genes es proporcional al número de parámetros a pesar, se ha llegado a un tamaño de población óptimo para diferentes valores de *k* (la experimentación incluye el pesado de los subconjuntos seleccionados). Para los subconjuntos de k parámetros donde $20 \leq k \leq 30$ la población consta de 100 individuos, y de 80 individuos para *k*<20.

3.3.2 Resultados de selección mediante GA

En este apartado se describen los resultados obtenidos con ambas aproximaciones:

- La selección directa mediante GA (SD).
- La selección indirecta mediante un ranking de pesos óptimos (SPO).

Repitiendo el procedimiento seguido en los experimentos de reducción de dimensionalidad mediante transformaciones lineales, además de las tasas de error obtenidas con los conjuntos seleccionados, se muestran también los intervalos de confianza, que permiten comparar de forma significativa el rendimiento de diferentes conjuntos.

3.3.2.1 Resultados de selección directa mediante GA

En la **Tabla 3.5** se muestran los conjuntos óptimos obtenidos mediante SD al emplear modelos GMM(EM). Los términos c*XX* hacen referencia a los MFCC, d*XX* y dd*XX* a sus primeras y segundas derivadas respectivamente, E a la energía, dE a su primera derivada y dEE a su segunda derivada. Tal como muestran los conjuntos óptimos de Albayzín, el que una componente aparezca en un subconjunto óptimo *k*-dimensional no implica que esté presente en subconjuntos de mayor dimensión. Por ejemplo, se observa que el cepstral c05 aparece en el subconjunto seleccionado para *k*=6, pero no está presente en los subconjuntos seleccionados para *k*=8 y *k*=10. Estos resultados parecen indicar que los sub-espacios óptimos no se pueden definir de forma incremental. Es decir, apuntan a que la definición del subconjunto óptimo requiere una búsqueda exhaustiva que explore todas las posibles combinaciones. Hay que destacar que todos los subconjuntos de tamaño $k \leq 12$ están compuestos por algunos MFCC y la energía. Tres de estos parámetros, E, c04 y c11 han resultado seleccionados en todos casos; otros tres, c02, c06 y c09 aparecen en todos los subconjuntos excepto para *k*=6. Por tanto, los resultados obtenidos parecen indicar que los parámetros estáticos son más relevantes que los dinámicos para el reconocimiento de locutores.

Tabla 3.5. *Subconjuntos óptimos de parámetros acústicos determinados por SD sobre Albayzín y Dihana con modelos GMM(EM), para k= 30, 20, 13, 12, 11, 10, 8 y 6. Los parámetros seleccionados se indican mediante un asterisco (*).*

SD GMM(EM)		E	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12
Albayzín	30	*	*	*	*	*	*	*	*	*	*	*	*	*
	20	*	*	*	*	*	*	*	*	*	*	*	*	*
	13	*	*	*	*	*		*	*	*	*	*	*	*
	12	*	*	*		*	*	*	*	*	*	*	*	*
	11	*		*	*	*	*	*	*	*	*	*	*	
	10	*	*	*	*	*		*			*	*	*	*
	8	*		*		*		*		*	*		*	*
	6	*				*	*			*		*	*	
Dihana	30	*	*	*	*	*	*	*	*	*	*	*		
	20	*	*	*	*	*	*	*	*	*	*	*		
	13		*	*	*	*	*	*	*	*	*	*		
	12		*	*	*	*	*	*	*	*	*	*		
	11		*	*	*	*	*	*	*	*	*	*		
	10		*	*	*	*	*	*	*	*	*	*		
	8		*	*	*		*	*	*	*		*		
	6		*				*	*	*	*		*		

dE	d01	d02	d03	d04	d05	d06	d07	d08	d09	d10	d11	d12
		*	*		*			*	*	*	*	*
					*				*	*		
	*	*	*	*	*	*	*	*	*	*		
		*	*	*	*	*	*		*	*		
					*		*	*				
			*				*					
							*					

ddE	dd01	dd02	dd03	dd04	dd05	dd06	dd07	dd08	dd09	dd10	dd11	dd12
		*		*	*		*	*	*	*	*	*
				*		*	*				*	
					*							
*			*	*	*	*	*	*	*	*		
								*				

El comportamiento de los subconjuntos óptimos en Dihana es prácticamente secuencial, sobre todo si lo comparamos con los resultados obtenidos en Albayzín. Se han detectado dos únicos casos que muestran un comportamiento no-secuencial: d08, el cual desaparece de *k*=20 a *k*=13; y d03, que desaparece de *k*=13 a *k*=12. Por otro lado, se observa que todos los subconjuntos óptimos de dimensiones pequeñas (con $k \leq 10$) están compuestos exclusivamente por MFCC (no aparecen ni la energía, ni los parámetros dinámicos). De nuevo, parece que los MFCC aportan más información relativa al locutor que sus derivadas. Pero, a diferencia de los subconjuntos de Albayzín, la energía no

aparece en ningún subconjunto óptimo con $k \leq 13$ parámetros, lo que sugiere que este parámetro es menos robusto para habla espontánea adquirida por canal telefónico que para habla leída grabada en condiciones de laboratorio.

En las tablas 3.6 y 3.7 se comparan los subconjuntos óptimos obtenidos con la aproximación SD utilizando diferentes estrategias de modelado (e-VQ y GMM(EM)). El análisis se ha realizado en Albayzín (**Tabla 3.6**) y Dihana (**Tabla 3.7**), con objeto de validar la metodología de selección y analizar si los subconjuntos óptimos, y su rendimiento, dependen de la metodología de modelado.

Tabla 3.6. *Subconjuntos óptimos de parámetros acústicos determinados por SD sobre Albayzín, para k= 30, 20, 13, 12, 11, 10, 8 y 6, con modelos discretos (e-VQ) y continuos (GMM(EM)). Los parámetros seleccionados se indican mediante un asterisco (*).*

k	Albayzín SD	E	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	dE	d01	d02	d03	d04	d05	d06	d07	d08	d09	d10	d11	d12	ddE	dd01	dd02	dd03	dd04	dd05	dd06	dd07	dd08	dd09	dd10	dd11	dd12
30	GMM-EM	*	*	*	*	*	*	*	*	*	*	*	*	*			*	*		*			*	*	*	*	*			*		*	*		*		*	*	*	*
	e-VQ	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*			*		*		*		*	*		*	*	*	*	*		*		*		*	*
20	GMM-EM	*	*	*	*	*	*	*	*	*	*	*	*	*						*				*	*								*		*	*				*
	e-VQ	*	*	*	*	*	*	*	*	*	*	*	*	*								*			*				*		*	*								*
13	GMM-EM	*	*	*	*	*		*	*	*	*	*	*	*																			*							
	e-VQ	*	*	*	*	*	*	*	*	*	*	*		*																			*							
12	GMM-EM	*	*	*		*	*	*	*	*	*	*	*	*																										
	e-VQ	*	*	*	*		*	*	*	*	*	*	*	*																										
11	GMM-EM	*		*	*	*	*	*	*	*	*	*	*																											
	e-VQ	*	*	*	*	*	*	*		*	*	*		*																										
10	GMM-EM	*	*	*	*	*		*			*	*	*	*																										
	e-VQ		*		*	*		*	*	*	*	*	*	*																										
8	GMM-EM	*		*		*		*		*	*		*	*																										
	e-VQ		*		*	*		*	*	*	*			*																										
6	GMM-EM	*				*	*			*		*	*																										
	e-VQ		*		*	*		*			*			*																										

En Albayzín, aunque las componentes de los subconjuntos óptimos para k={6, 8, 10, 11, 12} varían, revelando la aleatoriedad del procedimiento de búsqueda, todos consisten en un número de MFCC y la energía del tramo. Para k=13 el conjunto óptimo está compuesto por parámetros estáticos y la segunda derivada dd05. Los parámetros dinámicos no resultan seleccionados para conjuntos de pequeñas dimensiones (excepto para k=13), y más aún, los subconjuntos de tamaño k={20, 30} contienen todos los parámetros estáticos (MFCCs y la energía).

Los resultados obtenidos sobre Dihana indican una tendencia similar. De nuevo, todos los subconjuntos de tamaño $k \leq 10$ no incluyen ningún parámetro dinámico y están todos formados por MFCC. Los subconjuntos óptimos para $k \leq 10$ no contienen la energía del tramo, volviendo a indicar la posible falta de robustez de la energía al clasificar señales grabadas por canal telefónico. Más aún, el conjunto óptimo de k=10 componentes está formado únicamente por las MFCC con ambas técnicas de modelado. En cuanto a los subconjuntos de tamaño k={20, 30}, se repite la misma tendencia que en Albayzín: los subconjuntos de dimensión alta incluyen todos los parámetros estáticos.

Tabla 3.7. *Subconjuntos óptimos de parámetros acústicos determinados por SD sobre Dihana, para k= 30, 20, 13, 12, 11, 10, 8 y 6, con modelos discretos (VQ) y continuos (GMM). Los parámetros que han resultado seleccionados se indican mediante un asterisco (*).*

K	Dihana	E	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	dE	d01	d02	d03	d04	d05	d06	d07	d08	d09	d10	ddE	dd01	dd02	dd03	dd04	dd05	dd06	dd07	dd08	dd09	dd10
30	GMM-EM	*	*	*	*	*	*	*	*	*	*	*		*	*	*	*	*	*	*	*	*	*	*			*	*	*	*	*	*	*	*
	e-VQ	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*	*	*	*			*	*
20	GMM-EM	*	*	*	*	*	*	*	*	*	*	*			*	*	*	*	*	*		*	*									*		
	e-VQ	*	*	*	*	*	*	*	*	*	*	*		*	*	*					*	*	*	*	*				*					
13	GMM-EM		*	*	*	*	*	*	*	*	*	*						*		*	*													
	e-VQ	*	*	*	*	*	*	*	*	*	*	*										*			*									
12	GMM-EM		*	*	*	*	*	*	*	*	*	*				*				*														
	e-VQ		*	*	*	*	*	*	*	*	*	*					*							*										
11	GMM-EM		*	*	*	*	*	*	*	*	*	*								*														
	e-VQ		*	*	*	*	*	*	*	*	*	*												*										
10	GMM-EM		*	*	*	*	*	*	*	*	*	*																						
	e-VQ		*	*	*	*	*	*	*	*	*	*																						
8	GMM-EM		*	*	*		*	*	*	*		*																						
	e-VQ		*		*		*	*	*	*	*	*																						
6	GMM-EM		*				*	*	*	*		*																						
	e-VQ		*	*		*		*	*			*																						

El error promedio y el intervalo de confianza del 95% de los subconjuntos obtenidos en los experimentos SD se muestran en la **Tabla 3.8**. Como referencia, se incluyen las tasas de error obtenidas con el vector original *d*-dimensional y los subconjuntos compuestos por las componentes MFCC y MFCC+E. La evolución del GA depende del conjunto de validación empleado para evaluar los subconjuntos candidatos *k*-dimensionales. El conjunto de test, independiente del conjunto de entrenamiento y validación, sólo se utiliza para medir el rendimiento de la solución óptima y así evaluar la convergencia del GA. Por tanto, con objeto de analizar la consistencia de los subconjuntos obtenidos mediante la aproximación SD, en la Tabla 3.8 se presentan las tasas de error sobre ambos corpus. En las **Figuras 3.9 y 3.10** se visualiza el rendimiento de todos los subconjuntos obtenidos mediante SD en Albayzín y Dihana, respectivamente. En todos los casos hay una estrecha correlación entre las tasas obtenidas sobre las dos particiones, lo que pone de manifiesto la validez y robustez de los GA para este tipo de búsquedas heurísticas.

Como cabía esperar, los GMM(EM) han proporcionado mejores resultados que los modelos e-VQ. Pero este trabajo no trata de comparar el rendimiento de las diferentes estrategias de modelado, sino determinar el subconjunto óptimo de parámetros acústicos para el reconocimiento de locutores, y analizar en qué medida depende de la estrategia de modelado que se aplica para clasificar los locutores. El incremento absoluto del error promedio para los subconjuntos óptimos de tamaño $k \geq 11$ de Albayzín es muy pequeño (véase la Figura 3.9). De nuevo, estos resultados sugieren que los parámetros estáticos contienen más información referente al locutor que los parámetros dinámicos (primeras y segundas derivadas). Esto no significa que los parámetros dinámicos no aporten información, sino que al definir conjuntos reducidos los parámetros estáticos son más relevantes que los dinámicos (en la bibliografía se pueden encontrar diversos estudios que demuestran mejoras al incluir derivadas). Los subconjuntos de tamaño k>10 de Dihana reflejan esta misma tendencia, con un

aumento absoluto del error promedio de nuevo muy pequeño (véase la Figura 3.10). En este caso, al incluir la energía en el conjunto formado por 10MFCC (que es el conjunto óptimo para *k*=10 mediante SD), se observa un incremento absoluto del error promedio de 1.85 puntos al emplear modelos GMM(EM) y de 0.89 puntos al emplear modelos e-VQ. Este resultado refleja la falta de robustez de la energía al procesar señales telefónicas, comportamiento ya manifestado por el GA.

Tabla 3.8. *Error promedio e intervalo de confianza del 95% de los subconjuntos obtenidos en los experimentos SD sobre Albayzín y Dihana, para k= 6, 8, 10, 11, 12, 13, 20 y 30. Se añaden como referncia las tasas de los conjuntos MFCC, MFCC+E y el vector original d-dimensional.*

k	Error Promedio ± Intervalo Confianza del 95%							
	Albayzín SD				Dihana SD			
	e-VQ		*GMM-EM*		*e-VQ*		*GMM-EM*	
	Val	Eval	Val	Eval	Val	Eval	Val	Eval
6	22.9±0.3	16.69±0.2	7.64±0.12	5.71±0.09	44.72±0.23	47.36±0.24	31.76±0.16	34.23±0.12
8	14.34±0.22	10.06±0.2	2.86±0.12	1.81±0.09	39.38±0.21	42.45±0.16	21.99±0.13	23.90±0.14
10	14.01±0.19	9.8±0.16	2.24±0.11	0.94±0.04	36.27±0.21	38.49±0.22	17.91±0.16	19.70±0.12
11	11.59±0.2	7.91±0.16	0.81±0.06	0.35±0.04	35.23±0.2	37.45±0.17	17.64±0.11	19.32±0.14
12	11.96±0.29	8.69±0.18	1.23±0.07	0.3±0.04	35.47±0.26	37.31±0.21	17.37±0.09	19.27±0.14
13	10.63±0.23	7.07±0.22	1.05±0.06	0.36±0.03	35.28±0.21	37.34±0.2	17.30±0.12	19.12±0.14
20	10.47±0.23	7.14±0.15	0.67±0.09	0.16±0.02	35.1±0.23	37.54±0.23	17.59±0.09	19.99±0.11
30	10.71±0.26	7.17±0.22	**0.57±0.05**	**0.13±0.02**	35.26±0.26	37.31±0.22	16.05±0.14	19.10±0.14
MFCC	10.78±0.21	7.07±0.21	1.27±0.08	0.4±0.06	36.27±0.21	38.49±0.22	17.91±0.16	19.70±0.12
MFCC+E	**10.42±0.24**	**6.92±0.18**	0.90±0.05	0.22±0.04	36.81±0.2	39.38±0.15	19.76±0.14	22.34±0.1
d(39,33)	10.49±0.26	7.21±0.13	0.77±0.09	0.2±0.03	**35.07±0.23**	**37.23±0.21**	**15.66±0.16**	**18.69±0.15**

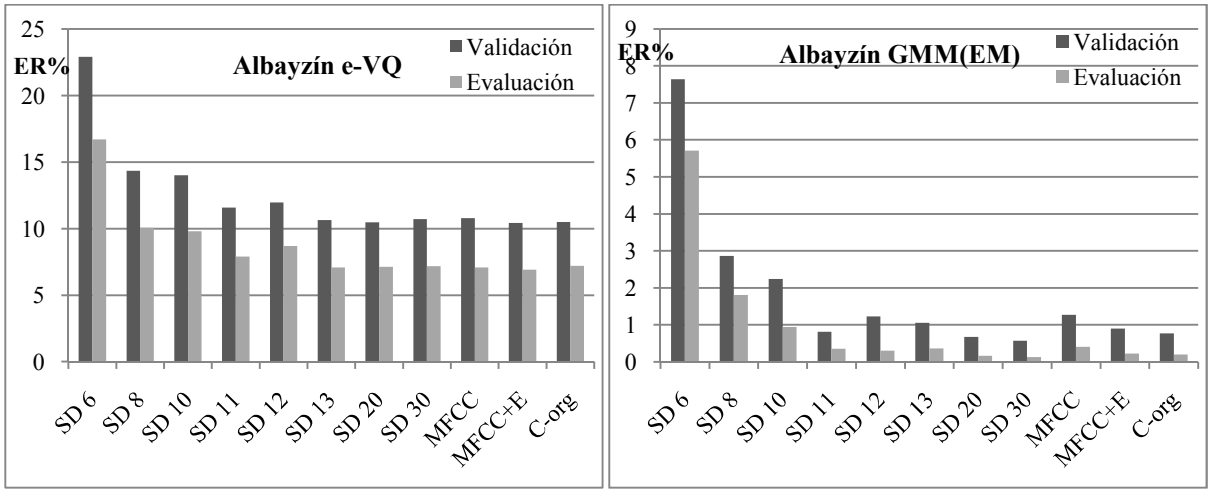

Figura 3.9. *Error promedio de los subconjuntos óptimos de k parámetros obtenidos mediante SD en Albayzín con modelos e-VQ y GMM(EM), para k= 30, 20, 13, 10, 8 y 6. Se incluye el resultado de los conjuntos MFCC, MFCC+E y el conjunto original de parámetros.*

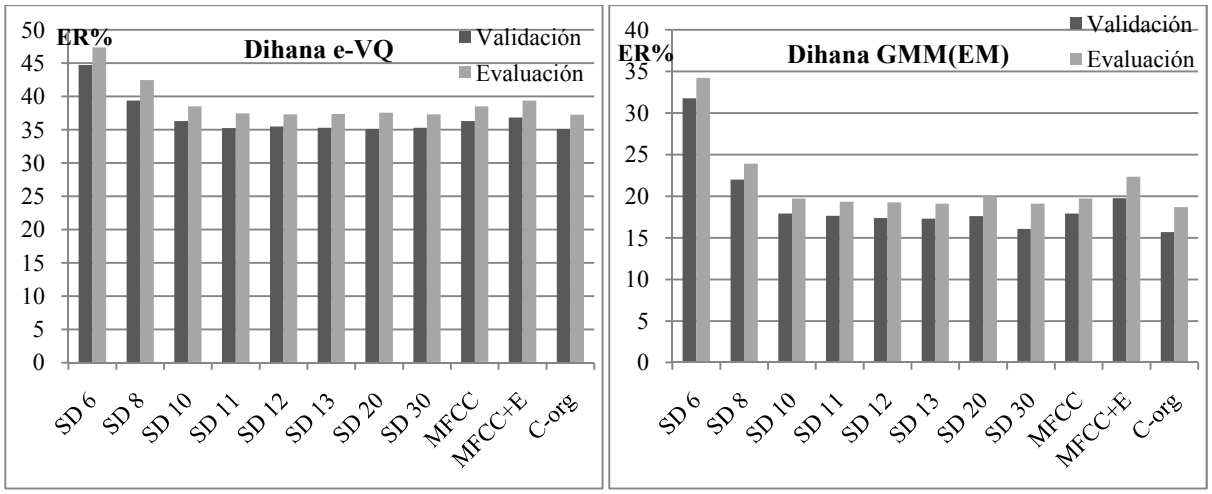

Figura 3.10. *Error promedio de los subconjuntos óptimos de k parámetros obtenidos mediante SD en Dihana con modelos e- VQ y GMM(EM), para k= 30, 20, 13, 10, 8 y 6. Se incluye el resultado de los conjuntos MFCC, MFCC+E y el conjunto original de parámetros.*

La búsqueda SD es óptima en Dihana, puesto que el subconjunto óptimo para k=11 mejora el conjunto de referencia MFCC+E. Lo mismo sucede en Albayzín con los modelos GMM(EM) y k=12, donde el error promedio del subconjunto óptimo (compuesto por 11MFCC más la energía, véase la tabla 3.5) es inferior al del subconjunto de referencia compuesto por 12MFCC. Se repite la misma tendencia con el conjunto k=13 en Albayzín y e-VQ cuyo intervalo de confianza está solapado con el del conjunto de referencia 12MFCC+E. No obstante, los conjuntos de referencia MFCC+E en Albayzín con GMM(EM), y 12MFCC en Albayzín con e-VQ, obtienen mejores resultados que los subconjuntos óptimos determinados por el algoritmo (para k=13 y k=12, respectivamente).

En un principio, se creyó que este comportamiento sería debido a una convergencia inadecuada del GA, por lo que se iniciaron algunos experimentos de monitorización del GA. Se observó que el problema no estaba relacionado con la convergencia del GA, sino con la inicialización y convergencia de los modelos de locutor. El GA, durante el proceso de búsqueda, evalúa los subconjuntos candidatos y promueve la reproducción de los más adecuados, de aquellos individuos que reciben una mejor evaluación en un experimento de clasificación *específico* (escribimos el término específico en cursiva porque es una razón crítica). Supongamos que el conjunto de parámetros A es mejor que el conjunto de parámetros B en promedio (como en el caso del conjunto de referencia 12MFCC+E y el conjunto óptimo para k=13 en Albayzín con GMM(EM)). Si la diferencia entre las tasas que produce cada conjunto es pequeña, el conjunto B podría obtener una tasa mejor que A en un *experimento específico* (debido a la inicialización aleatoria de los modelos). Por tanto, es suficiente que el conjunto B obtenga una tasa mejor en un experimento específico para que el GA lo seleccione como óptimo. Para solucionar este problema, se podría redefinir la función de evaluación del GA: en vez de realizar un único experimento de reconocimiento, se podría realizar un mayor número de experimentos (por ejemplo 20) por individuo, y asignarle al individuo el promedio de las tasas obtenidas. No obstante, esta solución aumentaría considerablemente el coste computacional del proceso de búsqueda hasta un punto inadmisible en la práctica.

3.3.2.2 Resultados de selección mediante un ranking de pesos

En la **Tabla 3.9** se muestran los subconjuntos óptimos *k*-dimensionales obtenidos mediante SPO sobre Albayzín y Dihana. Los resultados obtenidos revelan la misma tendencia observada en los conjuntos obtenidos mediante SD: los parámetros estáticos son más relevantes que los dinámicos. Los subconjuntos óptimos de k=20 y k=30 componentes incluyen todos los parámetros estáticos y ciertas derivadas. En Dihana esta condición se cumple para todos los subconjuntos de tamaño $k \geq 11$. Para valores de *k* más pequeños, se descarta la energía y algún coeficiente cepstral, lo que vuelve a poner de manifiesto la posible falta de robustez de la energía al procesar señales telefónicas. Asimismo, en los subconjuntos de dimensión reducida de Albayzín ($k \leq 13$), las derivadas sólo aparecen en los subconjuntos de tamaño k={13, 12}. En resumen, estos resultados vuelven a poner de manifiesto que los parámetros estáticos aportan más información referente al locutor que los parámetros dinámicos. Además, se comprueba la validez de los GA para resolver este tipo de búsquedas heurísticas, bien mediante búsqueda de los pesos óptimos, bien mediante búsqueda directa de los subconjuntos óptimos de parámetros.

Tabla 3.9. *Subconjuntos óptimos de parámetros acústicos determinados por SPO sobre Albayzín y Dihana, para k = 30, 20, 13, 12, 11, 10, 8 y 6 componentes. Los parámetros que han resultado seleccionados se indican mediante un asterisco (*).*

K	SPO	E	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	c11	c12	dE	d01	d02	d03	d04	d05	d06	d07	d08	d09	d10	d11	d12	ddE	dd01	dd02	dd03	dd04	dd05	dd06	dd07	dd08	dd09	dd10	dd11	dd12
30	Albayzín	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*	*	*			*		*		*	*			*	*	*	*	*	*			*	*	*
	Dihana	*	*	*	*	*	*	*	*	*	*	*			*	*	*	*	*	*			*	*				*	*	*	*	*	*	*	*	*	*	*		
20	Albayzín	*	*	*	*	*	*	*	*	*	*	*	*	*					*							*					*	*	*	*					*	
	Dihana	*	*	*	*	*	*	*	*	*	*	*											*	*				*	*	*	*		*		*				*	
13	Albayzín	*	*		*	*	*	*	*	*	*	*	*	*																	*									
	Dihana	*	*	*	*	*	*	*	*	*	*	*																*			*									
12	Albayzín	*	*		*	*	*	*	*	*	*	*	*	*																										
	Dihana	*	*	*	*	*	*	*	*	*	*	*																			*									
11	Albayzín	*	*		*	*	*	*	*	*	*	*		*																										
	Dihana	*	*	*	*	*	*	*	*	*	*	*																												
10	Albayzín	*	*		*	*	*	*		*	*	*		*																										
	Dihana		*	*	*	*	*	*	*	*	*	*																												
8	Albayzín	*	*		*	*	*	*		*				*																										
	Dihana		*	*	*	*		*	*	*	*																													
6	Albayzín	*	*		*	*	*							*																										
	Dihana		*		*	*			*	*	*																												

En las **Tablas 3.10 y 3.11** se detallan el error medio y el intervalo de confianza del 95% de los subconjuntos obtenidos con SD y SPO sobre los corpus de validación y test de Albayzín y Dihana. En el corpus de validación de Albayzín el error promedio de los subconjuntos SPO es superior al de los subconjuntos SD. Aunque ambas aproximaciones emplean GA para la búsqueda, SD realiza la optimización sobre el mismo corpus de validación y evalúa la propia selección. Por tanto, como cabe esperar, los resultados SD sobre el corpus de validación son, en la mayoría de los casos, ligeramente mejores que los resultados SPO (para los subconjuntos de tamaño $k \leq 13$ la superioridad de SD es

más notable; obsérvese el rendimiento del subconjunto SD de tamaño *k*=13). Esta tendencia se repite sobre el corpus de test, donde el error promedio de los subconjuntos SD es inferior al de los subconjuntos SPO, excepto para *k*=20, cuyos intervalos de confianza están solapados.

Tabla 3.10. *Error promedio e intervalo de confianza del 95% de los subconjuntos de parámetros obtenidos en los experimentos SPO y SD en Albayzín para k= 6, 8, 10, 11, 12, 13, 20 y 30.*

k	**Error Promedio ALBAYZÍN e-VQ ± Intervalo Confianza del 95%**			
	ER% Val		***ER% Eval***	
	SD	**SPO**	**SD**	**SPO**
6	22.9±0.3	30.71±0.23	16.69±0.16	24.99±0.24
8	14.34±0.22	16.25±0.3	10.06±0.2	12.03±0.26
10	14.01±0.19	14.07±0.25	9.8±0.16	10.28±0.26
11	11.59±0.2	12.71±0.28	7.91±0.16	8.87±0.13
12	11.96±0.29	12.45±0.27	8.69±0.18	8.84±0.22
13	10.63±0.23	12.24±0.25	**7.07±0.22**	8.89±0.14
20	**10.47±0.23**	**10.62±0.26**	7.14±0.15	**7±0.2**
30	10.71±0.26	10.75±0.27	7.17±0.22	7.34±0.18

Tabla 3.11. *Error promedio e intervalo de confianza del 95% de los subconjuntos obtenidos en los experimentos SPO y SD en Dihana, para k= 6, 8, 10, 11, 12, 13, 20 y 30.*

k	**Error Promedio DIHANA e-VQ ± Intervalo Confianza del 95%**			
	ER% Val		***ER%Eval***	
	SD	**SPO**	**SD**	**SPO**
6	44.72±0.23	48.49±0.18	47.36±0.24	51.6±0.22
8	39.38±0.21	39.68±0.25	42.45±0.16	43.65±0.18
10	36.27±0.21	36.27±0.21	38.49±0.22	38.49±0.22
11	35.23±0.2	36.81±0.22	37.45±0.17	39.38±0.15
12	35.47±0.26	35.21±0.22	**37.31±0.21**	37.54±0.21
13	35.28±0.21	35.25±0.18	37.34±0.2	**37.38±0.14**
20	**35.1±0.23**	**35.04±0.23**	37.54±0.23	37.39±0.27
30	35.26±0.26	35.35±0.19	37.31±0.22	37.48±0.19

En cuanto a los resultados sobre Dihana, se observa un rendimiento ligeramente mejor de la aproximación SPO sobre ambos corpus (validación y test). Esto podría deberse a que, tal como se mostraba en la tabla 3.7, el comportamiento de los subconjuntos SD de Dihana es casi secuencial. En este caso, en validación, el error promedio de los subconjuntos SPO de tamaño *k*={20, 13, 12} es ligeramente inferior al de los conjuntos SD correspondientes, aunque los intervalos están solapados y por tanto no se puede considerar que uno es mejor que otro. En cuanto al corpus de test, los intervalos

de confianza de los subconjuntos SD y SPO de tamaño *k*={30, 20, 13, 12} están solapados, por lo que tampoco puede decirse que un método sea mejor que el otro.

En ambas tablas (tabla 3.10 y tabla 3.11) se observa que para las dimensiones más pequeñas, especialmente para *k*=6, el rendimiento de los subconjuntos óptimos SD es notablemente superior al de los subconjuntos SPO. Es un comportamiento esperado, debido a la búsqueda global que realiza la aproximación SD. SPO, en cambio, depende de que se cumpla la hipótesis de que la relevancia del parámetro sea proporcional al peso que se le ha asignado. El rendimiento de los subconjuntos SD y SPO para $k \geq 10$ es muy similar, por lo que se puede concluir que SPO es también una buena aproximación para definir conjuntos de dimensión media a grande. Es una conclusión interesante, debido a que el coste computacional de SPO es considerablemente inferior al de SD.

3.3.2.3 Resultados utilizando parámetros acústicos pesados

En este apartado se presentan los resultados obtenidos en la búsqueda y aplicación de un conjunto de pesos óptimo $\widehat{\mathbf{w}} = (\widehat{w}_1, \widehat{w}_2, \dots, \widehat{w}_d)$ que transforma cada vector acústico $\mathbf{x} = (x_1, x_2, \dots, x_d)$ en un vector ponderado $\mathbf{x}' = (\widehat{w}_1 x_1, \widehat{w}_2 x_2, \dots, \widehat{w}_d x_d)$, con el objetivo objeto de mejorar el rendimiento un sistema de reconocimiento de locutores que utiliza modelos e-VQ. En la búsqueda de los pesos óptimos debe aplicarse un peso mayor a las dimensiones con mayor capacidad de discriminación y viceversa. Así, el pesado modula las dimensiones del espacio parametrizado, y trata de aumentar las distancias que favorecen la clasificación de locutores y de disminuir las que lo dificultan. Los parámetros originales se normalizan a media cero y varianza unitaria.

El peso asignado a cada parámetro es un entero en el rango [0, 255]. La búsqueda de pesos se realiza mediante GA, sobre los subconjuntos óptimos SD, SPO, y los conjuntos de referencia MFCC, MFCC+E y el conjunto original de parámetros. En la **Tabla 3.12** se presentan los pesos óptimos obtenidos para los (sub)conjuntos de Albayzín, donde se puede observar que, en general, los parámetros estáticos del conjunto original y de los subconjuntos SD y SPO de tamaño *k*={30, 20}, adquieren pesos más altos que las derivadas. Aunque hay ciertas excepciones, las derivadas adquieren pesos inferiores a 150 y los parámetros estáticos pesos superiores a 150. En cuanto a los subconjuntos de menor dimensión (MFCC, MFCC+E y los subconjuntos SD y SPO con $k \leq 13$) se observa que los pesos asignados están más equilibrados. Estos resultados sugieren de nuevo la superioridad de los parámetros estáticos frente a los dinámicos para la clasificación de locutores.

En la **Tabla 3.13** se muestra el rendimiento de los conjuntos SD y SPO pesados (denominados SD_W y SPO_W, respectivamente) de Albayzín. Se incluyen los resultados obtenidos sobre el corpus de validación y test, que presentan una gran correlación. La aplicación de los pesos óptimos produce una gran mejora de rendimiento: los resultados obtenidos sobre el corpus de validación presentan mejoras relativas del 28.26% y el 29.21% en promedio con SD_W y SPO_W, respectivamente; sobre el corpus de test, se obtienen mejoras relativas (en promedio) del 26.33% y 27.58%, respectivamente (véase la **Figura 3.11**).

Tabla 3.12. *Pesos óptimos asignados en los experimentos de pesado sobre Albayzín a los conjuntos óptimos SD, SPO, y conjuntos de referencia MFCC, MFCC+E y el vector original d-dimensional. El peso óptimo es un entero en el rango [0, 255].*

Pesos Óptimos Albayzín																			
k	**Vector Original**	**12MFCCE**	**12MFCC**	**SD**								**SPO**							
				30	20	13	12	11	10	8	6	30	20	13	12	11	10	8	6
E	220	240	-	209	221	205	193	249	-	-	-	181	223	225	236	186	167	251	242
c01	225	168	181	244	212	177	171	183	182	204	158	246	205	169	194	173	176	178	202
c02	152	174	200	217	251	208	239	226	-	-	-	242	218	-	-	-	-	-	-
c03	228	136	145	246	230	200	229	236	185	190	156	220	194	207	179	161	175	200	163
c04	223	197	165	220	193	200	-	227	232	225	160	217	226	203	249	233	170	239	168
c05	238	189	149	226	237	182	253	187	-	-	-	236	194	164	205	211	171	198	187
c06	218	149	140	198	201	150	182	212	167	157	-	245	200	192	189	159	140	192	-
c07	189	148	140	231	232	183	201	-	175	201	-	189	214	187	163	143	-	-	-
c08	209	149	150	232	212	191	208	197	161	189	-	225	179	209	181	144	131	183	-
c09	207	160	161	233	243	161	222	171	192	158	118	225	206	177	181	137	157	-	-
c10	204	137	145	211	202	146	172	195	150	-	-	205	211	162	169	129	121	-	-
c11	184	166	141	191	173	-	190	-	162	-	-	218	161	161	202	-	-	-	-
c12	237	138	144	210	199	139	170	169	145	147	114	213	225	151	169	158	139	158	133
dE	25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
d01	58	-	-	8	-	-	-	-	-	-	-	62	-	-	-	-	-	-	-
d02	93	-	-	121	-	-	-	-	-	-	-	73	-	-	-	-	-	-	-
d03	62	-	-	-	-	-	-	-	-	-	-	37	-	-	-	-	-	-	-
d04	118	-	-	-	-	-	-	-	-	-	-	73	14	-	-	-	-	-	-
d05	33	-	-	36	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
d06	11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
d07	65	-	-	39	32	-	-	-	-	-	-	71	-	-	-	-	-	-	-
d08	39	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
d09	45	-	-	47	-	-	-	-	-	-	-	44	-	-	-	-	-	-	-
d10	36	-	-	-	54	-	-	-	-	-	-	-	-	-	-	-	-	-	-
d11	124	-	-	72	-	-	-	-	-	-	-	58	115	-	-	-	-	-	-
d12	73	-	-	90	-	-	-	-	-	-	-	46	-		-	-	-	-	-
dd01	14	-	-	127	117	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dd02	61	-	-	124	-	-	-	-	-	-	-	76	-	-	-	-	-	-	-
dd03	147	-	-	76	13	-	-	-	-	-	-	100	67	-	-	-	-	-	-
dd04	161	-	-	63	101	-	-	-	-	-	-	140	92	123	-	-	-	-	-
dd05	104	-	-	125	-	46	-	-	-	-	-	127	125	-	-	-	-	-	-
dd06	114	-	-	-	-	-	-	-	-	-	-	23	21	-	-	-	-	-	-
dd07	61	-	-	36	-	-	-	-	-	-	-	86	-	-	-	-	-	-	-
dd08	14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dd09	43	-	-	114	60	-	-	-	-	-	-	-	-	-	-	-	-	-	-
dd10	60	-	-	133	-	-	-	-	-	-	-	31	-	-	-	-	-	-	-
dd11	112	-	-	24	-	-	-	-	-	-	-	77	19	-	-	-	-	-	-
dd12	84	-	-	38	35	-	-	-	-	-	-	39	-	-	-	-	-	-	-

En los pesos óptimos asignados a los subconjuntos de parámetros óptimos para Dihana se observa la misma tendencia que en Albayzín (véase la **Tabla 3.14)**. De nuevo, los parámetros estáticos reciben un peso mayor que los parámetros dinámicos, a excepción de la energía, que aunque en la mayoría de los casos recibe un peso superior al de las derivadas, no supera los valores asignados a los coeficientes cepstrales. Por tanto, se manifiesta de nuevo la mayor robustez de los parámetros estáticos frente a los dinámicos y la falta de robustez de la energía al clasificar las señales grabadas por canal telefónico.

Por último, en la **Tabla 3.15** se detallan el error promedio y el intervalo de confianza del 95% de los conjuntos SD, SD_W, SPO, SPO_W de Dihana. Se incluyen las tasas sobre los conjuntos de validación y test, que de nuevo presentan una gran correlación (véase la **Figura 3.12**). Los conjuntos SD_W presentan mejoras relativas del 11.74% y el 10.55% con respecto a SD sobre los corpus de

validación y test, respectivamente. Para los conjuntos SPO_W, estas mejoras son del 7.44% en validación y del 10.31% en test.

Cabe destacar que probablemente la aproximación SPO proporcionaría mejores resultados si la reducción de componentes se realizara en base a los pesos re-calculados para cada subconjunto. Es decir, si la reducción de k=30 a k=20 componentes se llevara a cabo realizando el ranking con el conjunto de pesos óptimo 30-dimensional, y no con el conjunto de pesos d-dimensional, y así sucesivamente.

Tabla 3.13. *Error promedio e intervalo de confianza del 95% de los subconjuntos SD, SPO, SD_W (SD con pesos óptimos) y SPO_W (SPO con pesos óptimos) en Albayzín. Se incluyen los resultados de los conjuntos de referencia MFCC, MFCC+E y el conjunto original d-dimensional.*

k	Error Promedio ALBAYZÍN ± Intervalo Confianza del 95%							
	SD				SPO			
	ER% Val		*ER% Eval*		*ER% Val*		*ER% Eval*	
	SD	SD_W	SD	SD_W	SPO	SPO_W	SPO	SPO_W
6	22.9±0.3	18.8±0.27	16.69±0.16	15.19±0.21	30.71±0.23	25.99±0.38	24.99±0.24	23.23±0.3
8	14.34±0.22	11.87±0.28	10.06±0.2	8.86±0.18	16.25±0.3	13.89±0.28	12.03±0.26	11.27±0.21
10	14.01±0.19	8.7±0.22	9.8±0.16	6.27±0.2	14.07±0.25	8.95±0.24	10.28±0.26	6.66±0.16
11	11.59±0.2	8.44±0.23	7.91±0.16	5.98±0.29	12.71±0.28	8.66±0.19	8.87±0.13	6.29±0.2
12	11.96±0.29	7.65±0.25	8.69±0.18	5.86±0.18	12.45±0.27	7.64±0.31	8.84±0.22	5.4±0.19
13	10.63±0.23	7.73±0.17	7.07±0.22	5.48±0.15	12.24±0.25	8.1±0.19	8.89±0.14	5.81±0.17
20	10.47±0.23	7.12±0.21	7.14±0.15	5.01±0.2	10.62±0.26	7.31±0.26	7±0.2	4.89±0.21
30	10.71±0.26	7.53±0.24	7.17±0.22	5.21±0.2	10.75±0.27	7.11±0.19	7.34±0.18	4.77±0.17
MFCC	10.78±0.21	7.56±0.25	7.07±0.21	5.06±0.15	10.78±0.21	7.56±0.25	7.07±0.21	5.06±0.15
MFCCE	**10.42±0.24**	**7.09±0.15**	**6.92±0.18**	**4.55±0.15**	**10.42±0.24**	**7.09±0.15**	**6.92±0.18**	**4.55±0.15**
d(39,33)	10.49±0.26	7.99±0.25	7.21±0.13	5.45±0.22	10.49±0.26	7.99±0.25	7.21±0.13	5.45±0.22

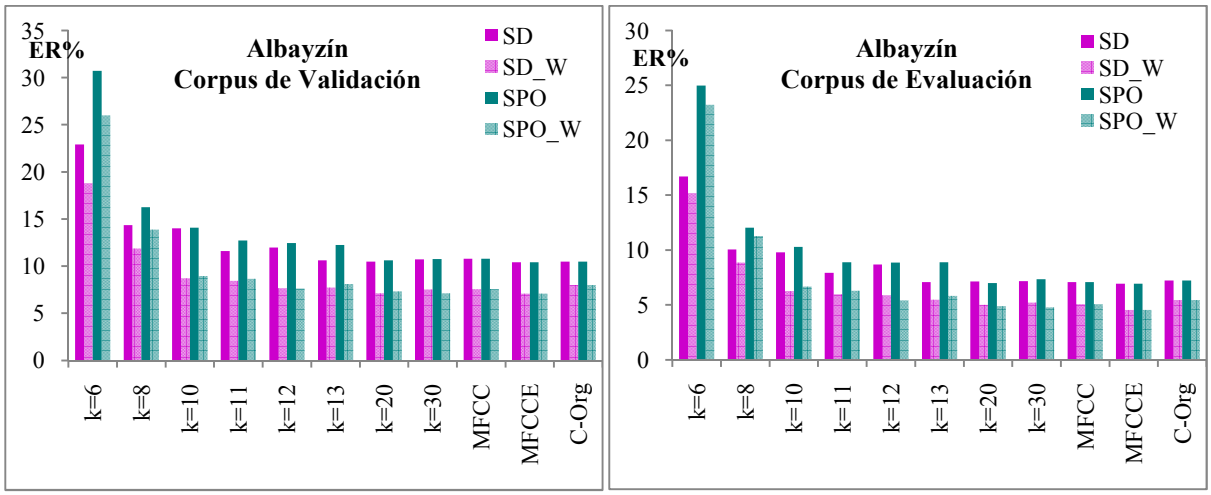

Figura 3.11. *Error promedio de los subconjuntos SD, SPO, SD_W (SD con pesos óptimos) y SPO_W (SPO con pesos óptimos) de tamaño k = 6, 8, 10, 11, 12, 13, 20, 30 sobre los corpus de validación (izquierda) y test (derecha) de Albayzín. Se incluyen los resultados de los conjuntos de referencia MFCC, MFCC+E y el conjunto original de parámetros.*

Tabla 3.14. *Pesos óptimos asignados en los experimentos de pesado sobre Dihana a los conjuntos óptimos SD y SPO, y a los conjuntos de referencia MFCC, MFCC+E y el vector original d-dimensional. El peso óptimo es un entero en el rango [0, 255].*

Pesos Óptimos Dihana VQ																			
K	**Vector Original**	**10MFCCE**	**10MFCC**	**SD**								**SPO**							
				30	20	13	12	11	10	8	6	30	20	13	12	11	10	8	6
E	190	173	-	219	187	177	-	-	-	-	-	186	182	119	188	173	-	-	-
c01	241	194	243	251	221	214	234	194	170	225	221	204	222	244	223	194	250	202	191
c02	232	198	197	226	231	213	250	198	173	-	188	251	247	202	231	198	209	193	-
c03	238	167	212	227	217	180	238	167	179	175	-	185	202	212	226	167	188	190	164
c04	237	158	192	195	227	190	215	158	138	-	174	164	169	213	174	158	190	164	154
c05	218	168	187	173	200	189	216	168	174	192	-	217	175	196	165	168	168	-	-
c06	221	179	182	237	220	200	238	179	165	206	223	214	229	249	210	179	246	153	-
c07	238	173	196	197	198	188	210	173	169	196	181	241	218	203	171	173	178	139	171
c08	239	140	151	200	199	176	215	140	159	192	-	196	175	221	156	140	188	152	128
c09	237	154	151	226	223	191	198	154	171	207	-	212	215	210	169	154	203	154	134
c10	213	180	182	220	196	165	197	180	137	178	145	209	230	209	218	180	203	-	-
dE	51	-	-	42	-	-	-	-	-	-	-	67	-	-	-	-	-	-	-
d01	23	-	-	67	10	-	-	-	-	-	-	8	-	-	-	-	-	-	-
d02	15	-	-	11	33	-	-	-	-	-	-	0	-	-	-	-	-	-	-
d03	47	-	-	64	62	-	-	-	-	-	-	98	-	-	-	-	-	-	-
d04	33	-	-	13	-	-	95	-	-	-	-	10	-	-	-	-	-	-	-
d05	43	-	-	84	-	-	-	-	-	-	-	10	-	-	-	-	-	-	-
d06	5	-	-	99	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
d07	9	-	-	11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
d08	65	-	-	78	14	-	-	-	-	-	-	80	26	-	-	-	-	-	-
d09	70	-	-	65	7	31	-	-	-	-	-	6	75	-	-	-	-	-	-
d10	12	-	-	47	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ddE	104	-	-	25	119	-	55	68	-	-	-	87	70	83	-	-	-	-	-
dd01	62	-	-	-	74	15	-	-	-	-	-	47	103	-	-	-	-	-	-
dd02	91	-	-	109	-	-	-	-	-	-	-	142	108	-	-	-	-	-	-
dd03	120	-	-	38	-	-	-	-	-	-	-	58	60	62	99	-	-	-	-
dd04	49	-	-	48	-	-	-	-	-	-	-	62	-	-	-	-	-	-	-
dd05	60	-	-	66	96	-	-	-	-	-	-	67	72	-	-	-	-	-	-
dd06	22	-	-	11	-	-	-	-	-	-	-	30	-	-	-	-	-	-	-
dd07	80	-	-	-	-	-	-	-	-	-	-	109	82	-	-	-	-	-	-
dd08	43	-	-	-	-	-	-	-	-	-	-	17	-	-	-	-	-	-	-
dd09	15	-	-	26	-	-	-	-	-	-	-	25	-	-	-	-	-	-	-
dd10	52	-	-	77	-	-	-	-	-	-	-	56	47	-	-	-	-	-	-

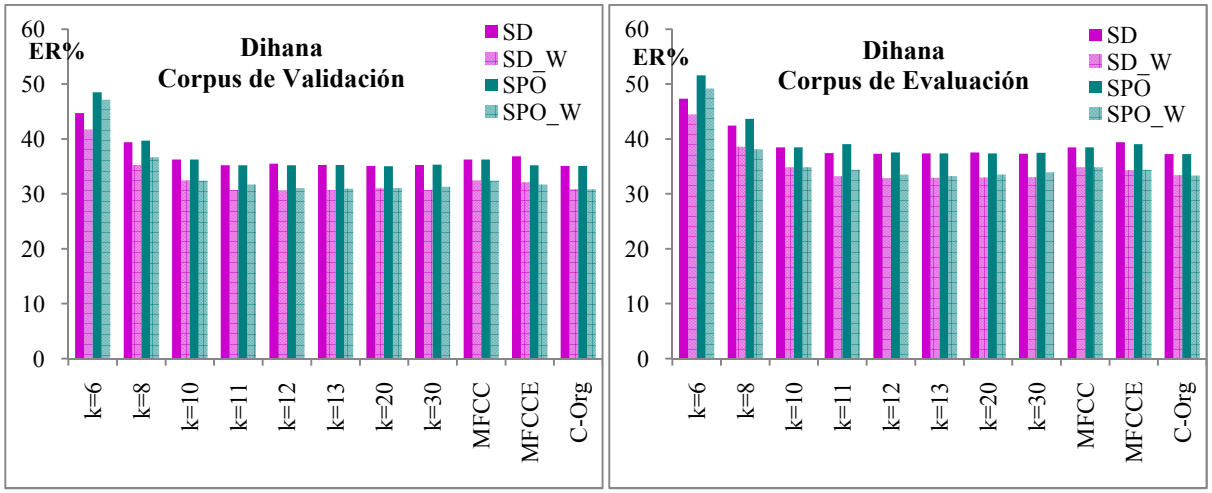

Figura 3.12. *Error Promedio de los subconjuntos SD, SPO, SD_W (SD con pesos óptimos) y SPO_W (SPO con pesos óptimos) de tamaño k=6, 8, 10, 11, 12, 13, 20, 30 sobre los corpus de validación (izquierda) y test (derecha) de Dihana. Se incluyen los resultados de los conjuntos de referencia MFCC, MFCC+E y el conjunto original de parámetros.*

Tabla 3.15. *Error promedio e intervalo de confianza del 95% de los subconjuntos SD, SPO, SD_W (SD con pesos óptimos) y SPO_W (SPO con pesos óptimos) en Dihana. Se incluyen los resultados de los conjuntos de referencia MFCC, MFCC+E y el conjunto original d-dimensional.*

k	Error Promedio DIHANA ± Intervalo Confianza del 95%							
	SD				SPO			
	ER% Val		ER% Eval		ER% Val		ER% Eval	
	SD	SD_W	SD	SD_W	SPO	SPO_W	SPO	SPO_W
6	44.72±0.23	41.72±0.18	47.36±0.24	44.47±0.23	48.49±0.18	47.15±0.21	51.6±0.22	49.19±0.15
8	39.38±0.21	35.26±0.28	42.45±0.16	38.61±0.2	39.68±0.25	36.66±0.16	43.65±0.18	38.05±0.14
10	36.27±0.21	32.43±0.24	38.49±0.22	34.83±0.24	36.27±0.21	32.43±0.24	38.49±0.22	34.83±0.24
11	35.23±0.2	30.69±0.17	37.45±0.17	33.2±0.2	35.23±0.2	31.69±0.17	39.08±0.15	34.4±0.2
12	35.47±0.26	**30.62±0.23**	37.31±0.21	**32.9±0.25**	35.21±0.22	31.05±0.21	37.54±0.21	33.53±0.21
13	35.28±0.21	30.63±0.19	37.34±0.2	32.93±0.2	35.25±0.18	30.94±0.24	37.38±0.14	**33.2±0.2**
20	35.1±0.23	30.99±0.2	37.54±0.23	32.98±0.19	**35.04±0.23**	31.05±0.19	37.39±0.27	33.54±0.21
30	35.26±0.26	30.71±0.17	37.31±0.22	33.03±0.18	35.35±0.19	31.31±0.27	37.48±0.19	33.93±0.28
MFCC	36.27±0.21	32.43±0.24	38.49±0.22	34.83±0.24	36.27±0.21	32.43±0.24	38.49±0.22	34.83±0.24
MFCCE	36.81±0.2	32.13±0.24	39.38±0.15	34.35±0.26	35.23±0.2	31.69±0.17	39.08±0.15	34.4±0.2
d	**35.07±0.23**	30.83±0.19	**37.23±0.21**	33.31±0.23	35.07±0.23	**30.83±0.19**	**37.23±0.21**	33.31±0.23

3.3.3 Conclusiones

El estudio presentado en este apartado trata de determinar un subconjunto óptimo de parámetros acústicos para el reconocimiento de locutores, mediante técnicas heurísticas de búsqueda, en concreto, GA. Se han propuesto y evaluado dos aproximaciones: (1) SD, donde el propio GA trata de identificar los parámetros más relevantes; y (2) SPO, una aproximación subóptima donde la selección depende de un ranking de pesos obtenido mediante un GA.

Los resultados obtenidos sobre la bases de datos Albayzín (habla grabada en condiciones de laboratorio) y Dihana (habla adquirida por canal telefónico), determinan que los parámetros estáticos (los coeficientes MFCC y la energía del tramo) son más relevantes que los parámetros dinámicos (sus derivadas) para el reconocimiento de locutores. Además, los resultados obtenidos apuntan a que la energía parece no ser robusta para clasificar las señales telefónicas dado que desaparece en los subconjuntos óptimos de pequeña dimensión en Dihana. Parece que el ruido del canal está distorsionando la información que transmite la energía del tramo. Este comportamiento, la prevalencia de los parámetros estáticos y falta de robustez de la energía sobre habla telefónica, se observa tanto en los experimentos SD, como en los experimentos SPO. En resumen, si, debido a las limitaciones computacionales o de almacenamiento, tuviéramos que seleccionar un conjunto de parámetros reducido, los coeficientes MFCC constituirían la mejor solución (conjunto que deberíamos aumentar con la energía si el habla se adquiere en condiciones de laboratorio). Además, estos experimentos validan el empleo de GA para resolver este tipo de búsquedas heurísticas, dado que todos los experimentos realizados llevan a las mismas conclusiones.

En cuanto a las aproximaciones SD y SPO se observa que el rendimiento de la metodología SD es superior al de SPO. Es un comportamiento esperado, debido a que el procedimiento SD dirige la evolución del algoritmo hacia la optimización de la función objetivo: reducir el número de parámetros acústicos con el menor impacto posible en el rendimiento de la clasificación de locutores. La aproximación SPO se propuso como sistema alternativo, menos costoso computacionalmente, para comprobar el rendimiento del procedimiento de pesado de parámetros acústicos, que se aplica con objeto de mejorar el rendimiento del sistema.

3.4 Comparación de metodologías

En este capítulo se han analizado diferentes estrategias de reducción de dimensionalidad. Como metodología principal, se ha propuesto la búsqueda heurística con GA del subconjunto óptimo de parámetros acústicos (aproximación denominada SD). Previamente, y a modo de contraste, se ha planteado el empleo de las transformaciones lineales PCA, LDA y HLDA. En cuanto a LDA y HLDA, se han analizado dos variantes: (1) LDA_EM y HLDA_EM en las que la transformación trata de maximizar la discriminación del espacio acústico de cada locutor; y (2) LDA_MAP y HLDA_MAP, que tratan de aumentar la capacidad de discriminación de cada unidad acústica del UBM. Esencialmente, partiendo un conjunto de parámetros original, estándar, *d*-dimensional, compuesto por los coeficientes MFCC, la energía y sus derivadas, se trata de definir el conjunto de *k* parámetros más discriminante para el reconocimiento de locutores, donde $k<d$. Con objeto de obtener el mejor equilibrio entre el rendimiento del sistema y el coste computacional, se ha analizado el rendimiento de las metodologías para diferentes valores de *k*.

En la **Tabla 3.16** y la **Figura 3.13** se ofrece una comparativa del rendimiento de dichas metodologías sobre Albayzín. Se observa que los conjuntos *d*-dimensionales óptimos son los obtenidos por LDA_MAP y HLDA_MAP. No obstante, para reducir la dimensionalidad, la mejor aproximación es SD, dado que, a diferencia de las transformaciones lineales, realiza la reducción optimizando la función objetivo real: la tasa de acierto en la clasificación de locutores.

De la comparación sobre Dihana podemos extraer conclusiones similares. Los resultados se presentan en la **Tabla 3.17** y en la **Figura 3.14**. En este caso, sobre los conjuntos de dimensión alta (concretamente para $k \geq 30$), las transformaciones lineales presentan un rendimiento mejor, superando el rendimiento de SD, con un error promedio considerablemente inferior, probablemente debido a la compensación del ruido de canal que las transformaciones lineales aplican implícitamente. Es una ventaja de las transformaciones lineales que no sólo seleccionan parámetros, sino que los combinan, por lo que pueden eliminar correlaciones y, por ejemplo, compensar el ruido del canal. Sin embargo, el criterio de estimación de las transformaciones no se corresponde con la función que se trata de optimizar: la tasa de acierto en la identificación. Sí lo hace la aproximación SD, y de hecho, para los conjuntos más reducidos SD sigue mostrando los mejores resultados (a excepción del subconjunto PCA de tamaño k=6).

Tabla 3.16. *Error promedio e intervalo de confianza del 95% de los experimentos de reducción de dimensionalidad con SD, PCA, LDA y HLDA sobre Albayzín, para k= 30, 20, 13, 12, 11, 10, 8 y 6 componentes. Para las transformaciones LDA y HLDA se analizan dos aproximaciones: (1) la discriminación del espacio acústico de cada locutor (LDA_EM y HLDA_EM); y (2) la discriminación de las unidades acústicas del UBM (LDA_MAP y HLDA_MAP). Se incluyen la tasas de los conjuntos d-dimensionales (sin reducción, d=39), y los conjuntos de referencia MFCC y MFCC+E.*

Error Promedio ± Intervalo de Confianza del 95 % en ALBAYZÍN						
k	***SD***	***PCA***	***LDA_EM***	***LDA_MAP***	***HLDA_EM***	***HLDA_MAP***
6	**5.71±0.09**	14.37±0.15	54.47±0.26	25.95±0.2	19.43±0.21	30.33±0.19
8	**1.81±0.09**	5.86±0.12	37.76±0.27	15.22±0.15	--	--
10	**0.94±0.04**	2.73±0.12	23.3±0.28	9.78±0.14	2.13±0.09	11.58±0.13
11	**0.35±0.04**	1.61±0.07	17.52±0.3	6.15±0.1	--	--
12	**0.3±0.04**	0.94±0.06	16.02±0.3	4.63±0.08	--	--
13	**0.33±0.05**	0.56±0.05	13.15±0.29	3.4±0.09	0.92±0.06	5.8±0.1
20	**0.16±0.02**	0.19±0.02	2.69±0.12	0.45±0.04	0.23±0.02	0.95±0.06
30	0.13±0.02	0.15±0.03	0.75±0.07	**0.1±0.02**	0.17±0.03	0.36±0.02
d(39)	0.2±0.03	0.2±0.02	0.29±0.04	**0.06±0.0**	0.24±0.03	**0.06±0.0**

MFCC	0.4±0.06	**MFCC+E**	0.22±0.04

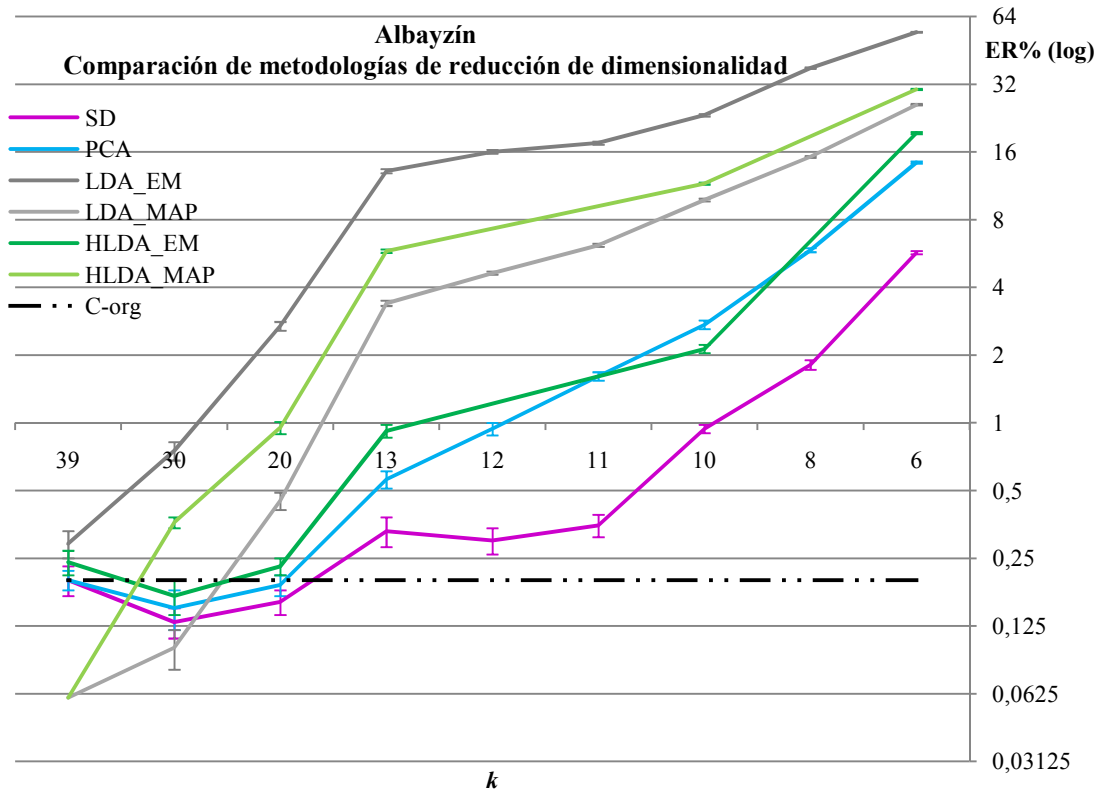

Figura 3.13. *Error promedio (en escala logarítmica) de los experimentos de reducción de dimensionalidad sobre Albayzín a k= 6, 10, 13, 20, 30 y 39 (sin reducción) componentes mediante SD, PCA, LDA y HLDA. Para las transformaciones LDA y HLDA se analizan dos aproximaciones: (1) la discriminación del espacio acústico de cada locutor (LDA_EM y HLDA_EM) y (2) la discriminación de las unidades acústicas del UBM (LDA_MAP y HLDA_MAP). Como referencia, se incluye el resultado del conjunto de datos original (C-org).*

Tabla 3.17. *Error promedio e intervalo de confianza del 95% de los experimentos de reducción de dimensionalidad con SD, PCA, LDA y HLDA sobre Dihana, para k= 30, 20, 13, 12, 11, 10, 8 y 6 componentes. Para las transformaciones LDA y HLDA se analizan dos aproximaciones: (1) la discriminación del espacio acústico de cada locutor (LDA_EM y HLDA_EM); y (2) la discriminación de las unidades acústicas del UBM (LDA_MAP y HLDA_MAP). Se incluyen la tasas de los conjuntos d-dimensionales (sin reducción, d=33), y los conjuntos de referencia MFCC y MFCC+E.*

Error Promedio ± Intervalo de Confianza del 95 % en DIHANA						
k	**SD**	**PCA**	**LDA_EM**	**LDA_MAP**	**HLDA_EM**	**HLDA_MAP**
6	34.23±0.16	**33.23±0.12**	72.68±0.21	61.3±0.13	46.58±0.15	67.8±0.15
8	**23.90±0.14**	24.19±0.13	60.33±0.25	49.34±0.13	--	--
10	**19.70±0.12**	20.67±0.12	53.63±0.28	39.51±0.12	25.61±0.15	44.5±0.12
11	**19.32±0.14**	20.27±0.13	51.33±0.27	34.18±0.13	--	--
12	**19.27±0.14**	19.75±0.16	49.22±0.34	29.96±0.11	--	--
13	**19.12±0.11**	19.63±0.1	47.96±0.18	26.64±0.12	19.74±0.16	32.82±0.14
20	19.99±0.11	17.61±0.13	33.71±0.23	16.71±0.09	**15.06±0.11**	20.71±0.11
30	19.10±0.14	15.97±0.15	18.44±0.18	**13.19±0.09**	14.78±0.12	13.66±0.08
d(33)	18.69±0.15	15.92±0.13	15.61±0.16	12.43±0.09	14.72±0.14	**11.44±0.11**

MFCC	19.70±0.12

MFCC+E	22.34±0.1

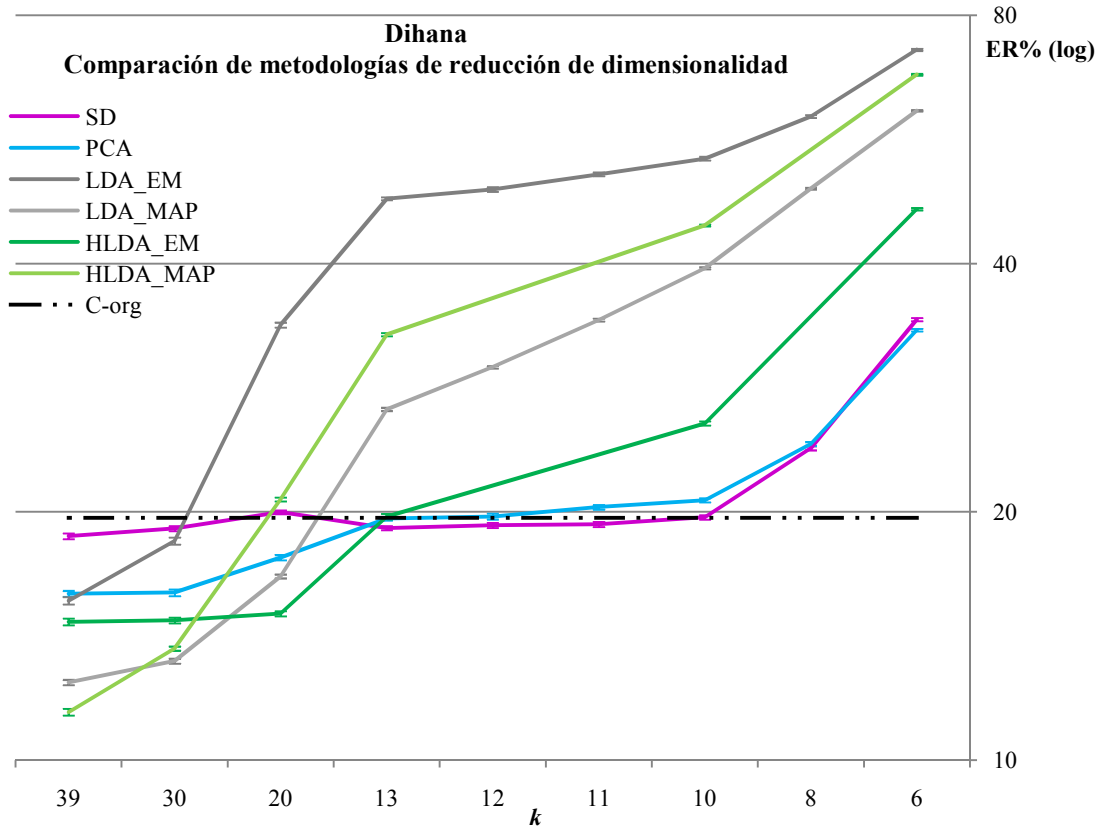

Figura 3.14. *Error promedio (en escala logarítmica) de los experimentos de reducción de dimensionalidad sobre Dihana a k= 6, 10, 13, 20, 30 y 39 (sin reducción) componentes mediante SD, PCA, LDA y HLDA. Para las transformaciones LDA y HLDA se analizan dos aproximaciones: (1) la discriminación del espacio acústico de cada locutor (LDA_EM y HLDA_EM) y (2) la discriminación de las unidades acústicas del UBM (LDA_MAP y HLDA_MAP). Como referencia, se incluye el resultado del conjunto de datos original (C-org).*

En conclusión, en los experimentos realizados sobre las señales de audio grabadas en condiciones de laboratorio, la aproximación SD ha mostrado mejor rendimiento que las transformaciones lineales para la reducción de dimensionalidad. Al tratar con habla espontánea grabada por canal telefónico, siguen mostrando buenos resultados para conjuntos de dimensión pequeña. No obstante, en este caso, para los conjuntos de mayor dimensión las transformaciones lineales han mostrado mejor rendimiento, probablemente debido a la compensación del ruido del canal. Por último, hay que tener en cuenta que las transformaciones lineales son más costosas que la simple selección de un subconjunto de parámetros, debido a que se debe reestimar la matriz de transformación para cada conjunto de datos de entrada. En consecuencia, si el subconjunto a definir es de dimensión pequeña, la selección del subconjunto óptimo puede ser preferible a la aplicación una transformación lineal.

Este estudio parece sugerir la combinación de transformaciones lineales y GA con objeto de aprovechar los puntos fuertes de ambas tecnologías. Se podría diseñar un GA orientado a la búsqueda de la matriz de transformación que maximice la tasa de identificación. Sin embargo, el coste computacional de esta aproximación no es abordable en la práctica, debido a que requiere la determinación de $k \cdot d$ coeficientes de transformación de coma flotante (en vez de k índices) y una gran cantidad de datos de entrenamiento y validación para que el GA converja y obtenga una transformación robusta. Otra posible aproximación basada en ambas tecnologías, y abordable en la práctica, sería la selección de parámetros mediante GA de un espacio transformado mediante LDA_MAP o HLDA_MAP que presentan muy buen rendimiento al no reducir la dimensionalidad. En este sentido, sería interesante explorar la interacción entre las técnicas de modelado y las técnicas de reducción de dimensionalidad, ya el rendimiento de LDA_MAP y HLDA_MAP decrece de forma considerable al reducir la dimensionalidad, mientras que HLDA_EM, el cual parte de una tasa de error superior, presenta buen comportamiento ante la reducción (véanse las figuras 3.13 y 3.14). Por último, resultaría interesante evaluar estas mismas estrategias (transformaciones lineales y selección del subconjunto óptimo de parámetros acústicos mediante GA) sobre una base de datos estándar más amplia, como las del NIST SRE.

Capítulo 4

Sistemas de identificación y verificación de locutores

Este capítulo incluye, en primer lugar, la evaluación de sistemas de reconocimiento de locutores sobre la base de datos en castellano Albayzín, de 204 locutores y grabada en condiciones de laboratorio. Se han desarrollado dos sistemas básicos con objeto de cubrir las tareas de: (1) identificación de locutores en conjuntos abiertos; (2) verificación de locutores; y (3) clasificación (identificación y verificación) de señales cuyas fuentes no se han modelado en entrenamiento, lo que revela un desajuste (*mismatch*) entre las señales de entrenamiento y test. Dentro de la línea de investigación relacionada con esta tercera tarea, se ha propuesto una nueva metodología de modelado para dar cobertura a fuentes no modeladas. Su objetivo principal es aumentar la robustez de los sistemas al tratar con señales que contienen fuentes no modeladas.

En la segunda parte del capítulo, se detalla la experimentación realizada sobre el corpus NIST SRE06. Se trata de una tarea de detección o verificación, concretamente la tarea principal de las evaluaciones NIST. Partiendo de una señal de entrenamiento y otra de test extraídas de sendas conversaciones telefónicas, cada una de unos 2 minutos y medio de duración, la tarea consiste en determinar si el locutor que aparece en la señal de entrenamiento habla en la señal de test. En la evaluación se analizan sistemas basados en las dos técnicas de modelado que han demostrado un mejor rendimiento en los últimos años: (1) modelos de locutor GMM(MAP); y (2)modelos de locutor GMM-SVM.

Asimismo, se estudia el rendimiento de diversas metodologías propuestas recientemente para la normalización de probabilidades y la calibración y fusión de sistemas.

4.1 Experimentos de reconocimiento de locutores en Albayzín

La experimentación llevada a cabo sobre Albayzín, enfocada tanto a tareas de identificación en un conjunto abierto, como a tareas de verificación, analiza el rendimiento de dos aproximaciones clásicas que emplean GMM como metodología de modelado: GMM(EM) y GMM(MAP). Ambas aproximaciones estiman el LLR entre la probabilidad condicional del modelo de locutor y la probabilidad condicional del UBM, y sólo difieren en la estimación del modelo de locutor.

En estos sistemas, el UBM consta de 1024 gaussianas estimadas mediante el algoritmo EM. El criterio de convergencia es un número máximo de iteraciones, que en esta experimentación se ha fijado a 20. En el sistema GMM(EM) el modelo de locutor es un GMM con 64 gaussianas estimadas en 20 iteraciones. En el sistema GMM(MAP), el modelo de locutor es un GMM de 1024 gaussianas adaptado a partir del UBM, donde únicamente se adaptan las medias de las componentes. El valor del coeficiente de adaptación es 16. Estos valores se han fijado empíricamente a aquéllos con el mejor rendimiento en experimentos preliminares.

En cuanto al proceso de parametrización,el habla se analiza en tramos de 25 milisegundos solapados a intervalos 10 milisegundos. Después, se aplica una ventana de *Hamming* y se calcula una FFT de 512 puntos. Las amplitudes de la FFT se promedian en 24 filtros triangulares solapados, donde las frecuencias centrales y anchos de banda se definen de acuerdo a la escala Mel. A continuación, se aplica la DCT sobre el logaritmo de las amplitudes del filtro y se obtienen 12 MFCC. Para aumentar la robustez frente al ruido del canal, se aplica CMS. Por último, se estiman la energía y las primeras y segundas derivadas de los parámetros MFCC y de la energía.

4.1.1 Identificación de locutores sobre un conjunto abierto

Con objeto de analizar el rendimiento de los sistemas GMM(EM) y GMM(MAP) en tareas de identificación de locutores en conjuntos abiertos, se ha definido una partición específica sobre la base de datos Albayzín, denominada "**AlbID 34/102/68**". La partición consta de cuatro subconjuntos disjuntos (su representación esquemática se muestra en la **Figura 4.1 (a)**):

- Un conjunto de entrenamiento con 34 locutores objetivo y 15 pronunciaciones por locutor. Es el conjunto empleado para estimar los modelos de locutor.
- Un conjunto de referencia (*background*) compuesto por 102 locutores y 25 pronunciaciones por locutor, mediante el cual se estima el UBM. En la medida de lo posible, los locutores que forman este conjunto están equilibrados en género.
- Un conjunto de test de locutores objetivo compuesto por 10 pronunciaciones de cada locutor objetivo (los locutores objetivo son los de entrenamiento, 34 en total).
- Un conjunto de test de impostores constituido por 68 locutores, para el que cada locutor aporta 10 pronunciaciones.

La partición da lugar 340 experimentos objetivo y 680 experimentos impostor, 1020 en total.

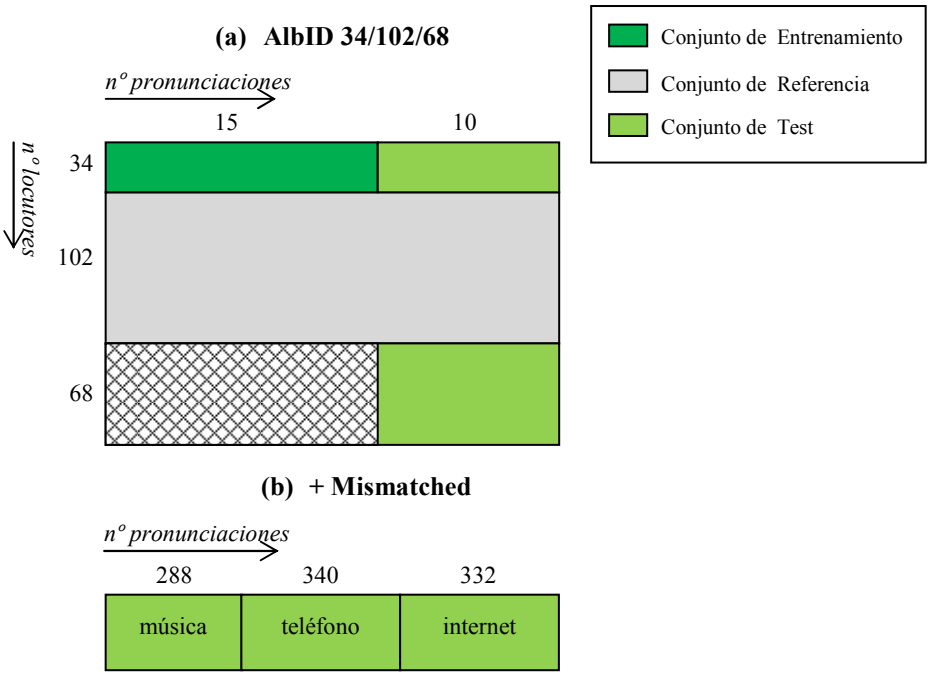

Figura 4.1 (a) *Representación esquemática de los conjuntos que componen la partición "AlbID 34/102/68" de Albayzín, definida para tareas de identificación de locutores en conjuntos abiertos.* **(a+b)** *Representación esquemática de la partición "AlbID 34/102/68+Mismatched", que además de los conjuntos de la partición anterior, contiene un conjunto de test compuesto por señales no modeladas en entrenamiento (señales de música, grabadas por teléfono y descargadas aleatoriamente desde Internet).*

En la **Figura 4.2** se muestra la curva DET de los dos sistemas sobre esta partición. Aunque el sistema GMM(EM) obtiene un rendimiento superior, hay que tener en cuenta que los dos sistemas presentan tasas de error muy pequeñas. El EER se sitúa en torno al 1.4% para el sistema GMM(EM), y en torno al 3.5% para el sistema GMM(MAP). El coste mínimo de todos los puntos de operación, C_{DET}^{min}, es de 0.01029 para el sistema GMM(EM), y de 0.03309 para el GMM(MAP) – los parámetros C_{fa}, C_{miss} y P_{obj} de la función de coste C_{DET} se han fijado a valores estándares para la identificación, concretamente a $C_{fa}=C_{miss}=1$ y $P_{obj}=0.5$.

Con objeto de analizar el rendimiento de los sistemas al tratar con señales que contienen fuentes acústicas no modeladas en entrenamiento, se ha definido una nueva partición denominada "**AlbID 34/102/68+Mismatched**". Es una extensión de la partición "AlbID 34/102/68" que, además de los cuatro subconjuntos de dicha partición, contiene un quinto subconjunto compuesto por señales que provienen de fuentes acústicas no observadas en entrenamiento que denominamos "señales desajustadas" *(Mismatched)*. Se trata de 960 pronunciaciones, de las cuales: (1) 288 son fragmentos de música seleccionados al azar sobre un conjunto de canciones; (2) 340 son fragmentos de conversaciones telefónicas escogidos de forma aleatoria de la base de datos Dihana; y (3) 332 son fragmentos de audio (la mayoría de voz) descargados aleatoriamente desde Internet. En la **Figura 4.1 (a+b)** se representa esquemáticamente la composición de esta nueva partición que produce 340 experimentos objetivo y 1640 experimentos impostor, lo que suma 1980 experimentos.

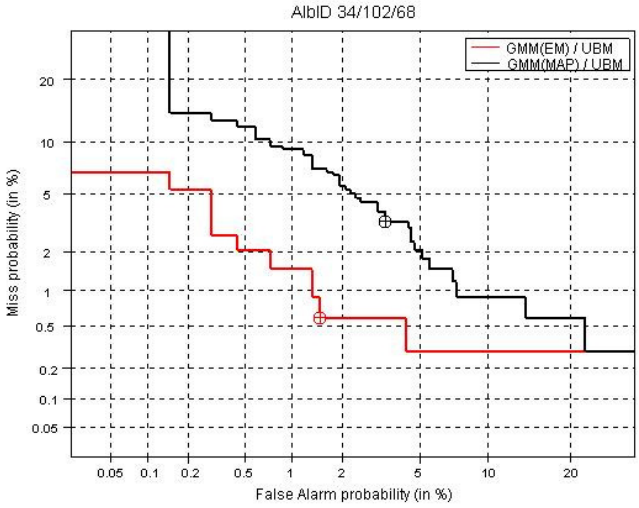

Figura 4.2. *Curvas DET de los sistemas GMM(EM) y GMM(MAP) sobre la partición "AlbID 34/102/68" de Albayzín definida para tareas de identificación de locutores en conjuntos abiertos.*

Al definir el conjunto de señales desajustadas se ha pretendido que su volumen sea parecido al volumen del conjunto de test de fuentes modeladas (es decir, a la suma del conjunto de test de locutores objetivo y el conjunto de test de impostores). Al igual que las pronunciaciones de Albayzín, las señales son como mínimo de 3 segundos de duración y, en caso necesario, se han re-muestreado a 16 kHz. El comportamiento de los modelos estas señales es prácticamente impredecible. Nótese que el conjunto de señales que se puede encontrar el sistema en una situación real no está definido a priori, y puede suceder que la fuente de la señal a evaluar no esté presente en las señales de entrenamiento.

El rendimiento de los sistemas sobre esta partición se muestra en la **Figura 4.3**

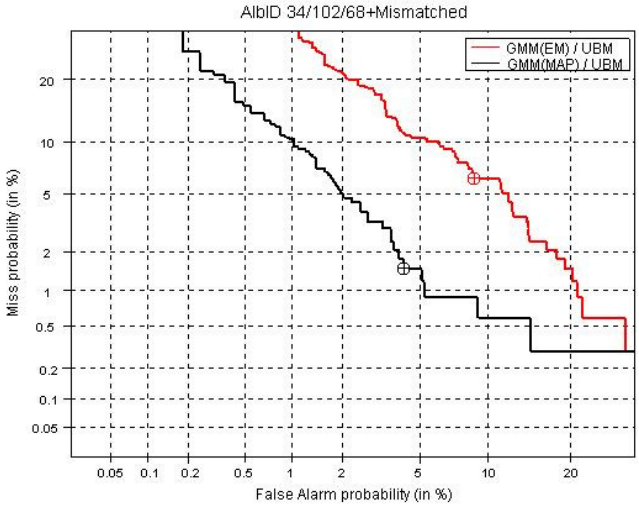

Figura 4.3. *Curvas DET de los sistemas GMM(EM) y GMM(MAP) sobre la partición "AlbID 34/102/68+Mismatched" de Albayzín, definida para tareas de identificación de locutor sobre conjuntos abiertos. Esta partición incluye, en el conjunto de test, un conjunto adicional con señales que contienen fuentes acústicas no modeladas en entrenamiento.*

Se observa que, al contrario de lo sucedido sobre la partición que no contiene señales desajustadas, en este caso el sistema GMM(MAP), con EER de 3.5%, supera el rendimiento del sistema GMM(EM), que presenta un EER del 8.1%. Lo mismo sucede con el coste mínimo C_{DET}^{min}, el cual es de 0.02808 para GMM(MAP), y de 0.07448 para GMM(EM). Al comparar los resultados obtenidos sobre esta partición y sobre la partición que no contiene señales desajustadas (Figura 4.2), el rendimiento del sistema GMM(EM) empeora considerablemente (el EER sube del 1.4% al 8.1%), mientras que el rendimiento del sistema GMM(MAP) es prácticamente el mismo (el EER se mantiene en 3.5%). Estos resultados indican que los modelos GMM(MAP) son más robustos que los modelos GMM(EM) para tareas de identificación sobre señales que contienen fuentes no modeladas.

4.1.2 Verificación de locutores

Para tareas de verificación, se ha definido una nueva partición de la base de datos, denominada "AlbVER 34/102/68" que consta también de cuatro subconjuntos disjuntos:

- Un conjunto de entrenamiento con 34 locutores objetivo y 10 pronunciaciones por locutor, empleados para estimar los modelos de locutor.
- Un conjunto de referencia (*background*) compuesto por 102 locutores y 25 pronunciaciones por locutor, empleado para la estimación del UBM. Este subconjunto es el mismo definido para tareas de identificación (véase 4.2.1).
- Un conjunto de test de locutores objetivo que consta de 15 pronunciaciones por locutor (estos locutores son los mismos que los de entrenamiento, 34 en total).
- Un conjunto de test de impostores constituido por 68 locutores y 15 pronunciaciones por locutor. Son locutores de Albayzín, pero distintos a los locutores de referencia y a los locutores objetivo.

En la **Figura 4.4 (a)** se muestra la representación esquemática de esta partición. Para cada locutor se evalúan 15 pronunciaciones propias y 1515 de impostores. En total, suman 510 experimentos objetivo y 51510 experimentos impostor.

En la **Figura 4.5** se muestran las curvas DET que producen los sistemas GMM(EM) y GMM(MAP) sobre este conjunto experimental. Los dos sistemas presentan un alto rendimiento: las tasas de error P_{miss} y P_{fa} son muy bajas en la mayor parte del barrido que realizan las curvas. Aunque prácticamente no se puede considerar que un sistema es mejor que otro, se observa un rendimiento ligeramente superior del sistema GMM(EM), con un 0.7% EER. El sistema GMM(MAP) obtiene una curva muy similar, con un EER también por debajo del 1% (concretamente, 0.85%). En cuanto al coste mínimo de los sistemas, C_{DET}^{min}, se obtiene el valor 0.004148 para GMM(EM), y 0.005463 para GMM(MAP) – los parámetros C_{fa}, C_{miss} y P_{obj} de la función C_{DET} (véase el apartado 2.6.3.2) se han fijado a valores estándares para tareas de verificación (según las últimas evaluaciones del NIST), concretamente a $C_{fa}=10$, $C_{miss}=1$ y $P_{obj}=0.01$.

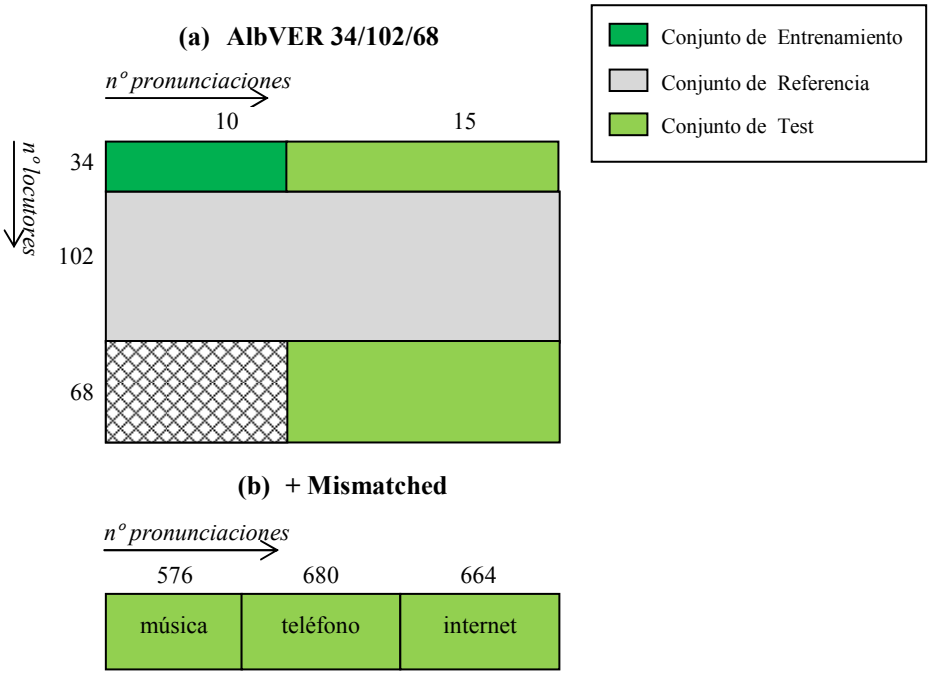

Figura 4.4 (a) *Representación esquemática de los conjuntos que componen la partición "AlbVER 34/102/68" de Albayzín, definida para tareas de verificación de locutores.* **(a+b)** *Representación esquemática de la partición "AlbVER 34/102/68+Mismatched", que además de los conjuntos de la partición anterior, contiene un conjunto de test compuesto por señales no modeladas en entrenamiento (señales de música, grabadas por teléfono y descargadas aleatoriamente desde Internet).*

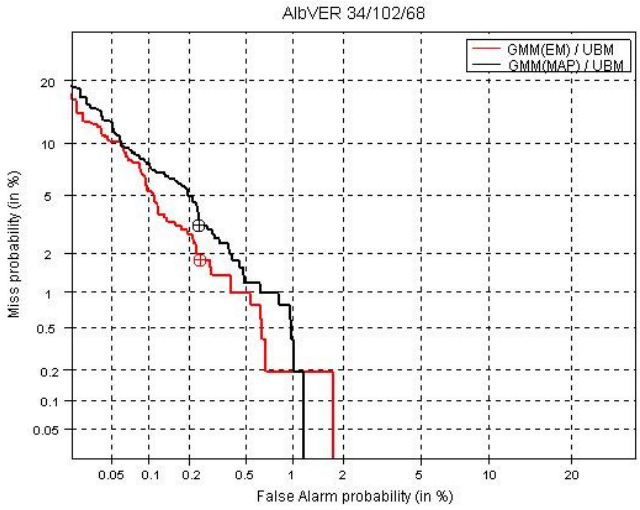

Figura 4.5. *Curvas DET de los sistemas GMM(EM) y GMM(MAP) evaluados sobre la partición "AlbVER 34/102/68" de Albayzín, definida para tareas de verificación de locutores.*

Siguiendo el mismo procedimiento que en la experimentación dedicada a tareas de identificación, se analiza el rendimiento de estos sistemas sobre señales que contienen fuentes acústicas no modeladas en entrenamiento. Para ello, se vuelve a definir una nueva partición, denominada "AlbVER 34/102/68+Mismatched", dividida en cinco subconjuntos disjuntos: 4 de estos subconjuntos son los de la partición "AlbVER 34/102/68". El quinto subconjunto está compuesto por "señales desajustadas" *(Mismatched)*. Contiene 1920 pronunciaciones, de las cuales: (1) 576 son fragmentos de

música seleccionados al azar sobre un conjunto de canciones; (2) 680 son fragmentos de conversaciones telefónicas escogidos de forma aleatoria de la base de datos Dihana; y (3) 664 son fragmentos de audio (la mayoría de voz) descargados aleatoriamente desde la web. Al igual que las pronunciaciones de Albayzín, estas señales son como mínimo de 3 segundos de duración y, en caso necesario, se han re-muestreado a 16 kHz. La representación esquemática de la composición de esta partición se muestra en la **Figura 4.4 (a+b).** En total, la partición da lugar a 510 experimentos objetivo y 116970 experimentos impostor, debido a que por cada locutor se evalúan 15 pronunciaciones suyas y 3435 de impostores.

En la **Figura 4.6** se muestra el rendimiento de los sistemas sobre esta última partición. Se observa un deterioro considerable del sistema GMM(EM). Su EER aumenta de 0.7% a 4.7% al incluir las señales no modeladas en el conjunto de test. El sistema GMM(MAP), en cambio, presenta prácticamente la misma curva que sobre la partición anterior y un rendimiento considerablemente superior a GMM(EM) (con un EER del 0.7%). El coste mínimo C_{DET}^{min} es de 0.02333 para GMM(EM) y de 0.004436 para GMM(MAP).

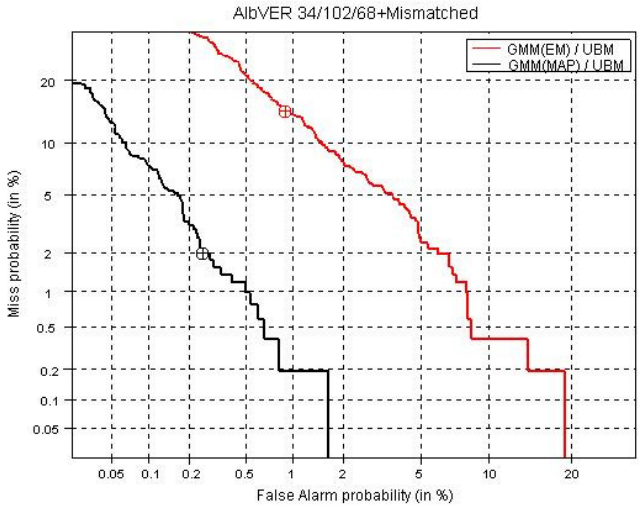

Figura 4.6. *Curvas DET de los sistemas GMM(EM) y GMM(MAP) sobre la partición "AlbVER 34/102/68+Mismatched" de Albayzín, definida para tareas de verificación. Esta partición incluye, en el conjunto de test, señales cuya fuente acústica no se ha modelado en entrenamiento.*

Estos experimentos revelan que tanto los modelos GMM(EM) como los modelos GMM(MAP) son apropiados para verificar señales que contienen fuentes acústicas modeladas. Sin embargo, el rendimiento de los modelos GMM(EM) se deteriora de forma significativa al incluir señales no modeladas en el conjunto de test. A diferencia de los modelos GMM(EM), el rendimiento de los modelos GMM(MAP) apenas se deteriora.

4.1.3 Aproximaciones para mejorar la identificación y verificación de locutores en señales procedentes de fuentes no modeladas

4.1.3.1 Consideraciones sobre criterios de verificación e identificación de locutores

Tal como se ha mencionado en el estado de arte (véase el apartado 2.4) los sistemas de identificación del locutor toman sus decisiones a partir de las probabilidades $p(l|X)$ a posteriori. El criterio es escoger el locutor con máxima probabilidad a posteriori. Aplicando la regla de Bayes, la probabilidad a posteriori del locutor *l* dada la pronunciación *X* se puede expresar como sigue:

$$p(l|X) = \frac{p(X|l)P(l)}{p(X)} \tag{4.1}$$

Para calcular la probabilidad condicional, se estima un modelo acústico λ_l para cada locutor l, de modo que $p(X|l) \cong p(X|\lambda_l)$. Por otra parte, si se supone que todos los locutores son igualmente probables, $P(l)$ es constante y la maximización de $p(l|X)$ no requiere de $P(l)$. Así, la maximización de $p(l|X)$ depende de un ratio entre dos probabilidades: (1) la probabilidad condicional de la pronunciación dado del modelo de locutor $p(X|\lambda_l)$; y (2) la probabilidad de la pronunciación $p(X)$. Si el conjunto de locutores es cerrado, $p(X)$ se calcula sumando las probabilidades de la pronunciación para los locutores objetivo (véase la ecuación 2.3). Sin embargo, si el conjunto de locutores es abierto, $p(X)$ es la probabilidad de la pronunciación para el espacio global de locutores, tanto locutores objetivo como impostores (más adelante se discuten distintas formas de estimar esta probabilidad). En este caso, $p(l|X)$ se compara con un umbral predeterminado θ. Si $p(l|X) > \theta$ se decide que *X* ha sido pronunciado ciertamente por *l*; en caso contrario, se decide que es un impostor. En escala logarítmica esta condición quedaría como sigue:

$$\Lambda(X) = \ln p(X|\lambda_l) - \ln p(X) > \theta \tag{4.2}$$

Los sistemas de verificación aplican un contraste de hipótesis (véase la ecuación 2.7) entre la hipótesis de que la pronunciación de entrada *X* provenga del locutor *l* y la hipótesis de que la pronunciación de entrada *X* no provenga de *l*. De nuevo, si el ratio es superior a un umbral predeterminado, se decide que *X* proviene de *l*; en caso contrario, se decide que proviene de un impostor:

$$\Lambda(X, l) = \frac{p(l|X)}{p(\bar{l}|X)} = \frac{p(X|\lambda_l)P(l)}{p(X|\lambda_{\bar{l}})P(\bar{l})} > \theta \tag{4.3}$$

Si se supone que $P(l) = P(\bar{l}) = 0.5$, la condición (4.3) se convierte en un ratio entre la probabilidad condicional de la pronunciación dado el modelo de locutor y la probabilidad condicional de la pronunciación dado el modelo alternativo.

Existen diferentes metodologías para representar la probabilidad condicional $p(X|\lambda_l)$, entre las que destacan los GMM (véase el apartado 2.4.2). λ_l se estima a partir de pronunciaciones del locutor que forman parte de un corpus de entrenamiento. Por tanto, $p(X|\lambda_l)$ está bien definido y representado.

Por otro lado, para calcular $p(X)$ se necesita un modelo λ_f de la fuente de *X*, que represente el espacio de todas las posibles pronunciaciones de entrada, de modo que $p(X) = p(X|\lambda_f)$. Si de

antemano se conoce la fuente de las señales a evaluar, se puede recopilar un conjunto de datos amplio, diverso y equilibrado con señales de este tipo y estimar un modelo acústico λ_f a partir de dicho conjunto. El modelo resultante recibe el nombre de modelo de referencia (*background model*). Desde un punto de vista práctico, el modelo de referencia proporciona un factor de normalización de la probabilidad de modelo de locutor, que trata de minimizar o eliminar la variabilidad específica de la pronunciación (tales como el canal, el ruido del entorno, etc.) no relacionada con el locutor. De esta forma, se puede establecer un único umbral para todos los locutores.

Así como los modelos de locutor y de referencia están bien definidos y representados, no está claro cómo se ha de estimar el modelo alternativo o de impostores, $\lambda_{\bar{l}}$, dado que los verdaderos impostores no se conocen de antemano. Se distinguen dos aproximaciones principales (descritas en el apartado 2.4.1): (1) definir una cohorte de locutores; o (2) definir un UBM. Estas dos aproximaciones tratan de modelar fuentes desconocidas (locutores desconocidos) a partir del conocimiento adquirido a partir de fuentes conocidas (locutores conocidos).

En el primer caso, $p(X|\lambda_{\bar{l}})$ se calcula mediante una función (generalmente la media aritmética) de las probabilidades condicionales de *X* para los locutores de la cohorte. Hay dos cuestiones a tener en cuenta al emplear esta aproximación: (1) se debe seleccionar una cohorte distinta para cada locutor objetivo; y (2) los locutores que forman la cohorte deben representar todo el espacio de locutores alternativos, lo cual resulta complicado de realizar con una cantidad limitada de locutores.

En el segundo caso (que es la aproximación más común), se estima un único modelo, independiente del locutor (idéntico al modelo de referencia descrito más arriba), a partir de un conjunto de locutores amplio, diverso y equilibrado: es el denominado modelo universal o UBM, citado en numerosas ocasiones a lo largo de este trabajo. El UBM presenta abundantes ventajas: (1) todas las probabilidades de los locutores objetivo se normalizan mediante este modelo, por lo que el sistema debe estimar un único modelo y calcular un único factor de normalización para cada señal; (2) proporciona una amplia cobertura acústica, siempre que para su estimación se emplee un conjunto amplio y diverso de datos; y (3) se puede emplear como modelo base en la adaptación bayesiana de los modelos de locutor. Hay que tener en cuenta, no obstante, que no siempre será posible reunir un conjunto de datos diverso para el UBM, lo que conllevaría el deterioro de la cobertura acústica que proporciona este modelo. En este caso, como en cualquier otra situación en la que la fuente que ha generado la pronunciación de entrada no esté debidamente modelada en el UBM (porque por ejemplo, no es voz sino ruido, música, etc.), la fuente acústica de la señal no estará representada ni en el modelo de locutor ni en el UBM, por lo que ninguno de los modelos dará cobertura a la señal. En consecuencia, la decisión que proporcionan los sistemas para estas entradas no es fiable.

Supóngase que el locutor de la señal de entrada *X* es l. En este caso la probabilidad $p(X|\lambda_l)$ será considerablemente superior a la probabilidad de referencia $p(X|\lambda_{UBM})$, dado que el UBM modela todas las posibles fuentes (locutores objetivo e impostores) y λ_l se concentra en una fuente específica: el locutor l. Si *X* es de un impostor similar pero diferente a l, $p(X|\lambda_l)$ será sólo ligeramente superior (o

incluso inferior) a la probabilidad de referencia. Pero si el UBM no representa la fuente acústica de la señal de entrada, $p(X|\lambda_{UBM})$ será impredecible, pudiendo ser arbitrariamente superior o inferior a $p(X|\lambda_l)$, y la pronunciación puede resultar erróneamente clasificada.

Para evitar estos errores, en esta tesis se ha propuesto una nueva metodología que trata de estimar la fuente de la señal de entrada, tanto en tareas de identificación en conjuntos abiertos [Zamalloa08e], como en tareas de verificación [Zamalloa08f]. El objetivo principal de la metodología propuesta es mejorar la cobertura acústica proporcionada por el UBM. Para ello, se estima un modelo acústico de la fuente que ha generado la pronunciación de entrada, denominado "modelo superficial de fuente". En los siguientes apartados se describe con más detalle esta nueva metodología, y se analiza su rendimiento en tareas de identificación y verificación de locutores.

4.1.3.2 Modelo superficial de fuente

El modelo superficial de fuente, denominado *Shallow Source Model* (SSM) [Zamalloa08e] trata de modelar la fuente acústica de la señal de entrada, con objeto de aumentar la robustez de los sistemas al tratar con señales no representadas de forma adecuada en el UBM. En definitiva, SSM es un GMM, que se denota como λ_X, y se estima a partir de la propia pronunciación de entrada *X* empleando el algoritmo EM. Como generalmente *X* suele ser de corta duración (de unos 2-10 segundos), la cantidad de componentes de λ_X debe ser reducida, de forma que modele debidamente la señal de entrada y no haya sobre-entrenamiento. Debe tenerse en cuenta que el SSM no pretende modelar específicamente la señal de entrada, sino la fuente acústica de la señal (por ejemplo, el locutor, o cualquier otro tipo de fuente). Por lo tanto, si λ_X se definiera con una cantidad elevada de componentes, modelaría de forma específica las características de la pronunciación, y no las de la fuente.

Supóngase que λ_X es un modelo de fuente perfecto. En este caso, se cumple:

$$p(X|\lambda_X) > p(X|\lambda_l) \tag{4.4}$$

donde λ_l es el GMM estimado a partir de las pronunciaciones del locutor, es decir el modelo de locutor. En esta situación, el ratio de probabilidades en escala logarítmica (LLR):

$$\Lambda(X) = \ln p(X|\lambda_l) - \ln p(X|\lambda_X) \tag{4.5}$$

es negativo, o cero en caso de que el modelo de locutor λ_l se ajuste de forma perfecta al SSM. Obviamente, en este último caso (o incluso cuando el ratio es muy próximo a cero), el sistema debería clasificar positivamente la pronunciación de entrada. De hecho, al normalizar la probabilidad del modelo de locutor con la probabilidad del modelo de fuente, el ratio obtenido indica el grado de aproximación del modelo de locutor a la fuente de la señal.

Sin embargo, en la práctica, la ecuación 4.4 no se cumple, dado que λ_X no es un modelo perfecto, sino un modelo superficial y aproximado. Los modelos de locutor contienen mucha información acústica, dado que se estiman a partir de muchos más datos que el SSM, por lo que muchos modelos de locutor representan de forma más adecuada la fuente de la señal de entrada que el SSM. En todo caso, siguiendo la misma interpretación anterior, el SSM se puede emplear para estimar el grado de

aproximación del modelo de locutor a la fuente de la señal, y la probabilidad condicional aportada por el SSM se puede emplear como información adicional en la normalización. Además, para las situaciones en las que ni el modelo de locutor ni el UBM representan adecuadamente la fuente de la señal de entrada (porque procede de un locutor raro o no contiene voz), el SSM garantiza una mínima cobertura: en este caso, la probabilidad del SSM será mayor y más robusta que la del modelo de locutor, y en consecuencia, el sistema decidirá que *X* proviene de un impostor, como debe ser.

Combinación del UBM y el modelo superficial de fuente

La aproximación SSM soluciona el problema de la cobertura acústica y no requiere tantos datos como el UBM para su estimación, pero hay que tener en cuenta que presenta una gran sensibilidad a las variaciones específicas de la señal de entrada, y en consecuencia, puede no representar de forma robusta la fuente de la señal. En los experimentos llevados a cabo en [Zamalloa08e] se observa como el rendimiento del UBM es claramente superior al del SSM.

Por tanto, para resolver los problemas de cobertura del UBM y de robustez del SSM, se podría estimar un modelo de referencia a partir de la señal de entrada, adaptando el UBM por medio del algoritmo MAP (véase [Aronowitz05]). El coste computacional de esta aproximación es elevada; al igual que en la aproximación SSM, se requiere la estimación de un nuevo modelo para cada pronunciación a evaluar, y además el modelo adaptado es un modelo complejo con un número elevado de componentes gaussianas estimadas por medio del algoritmo MAP. Por tanto, el coste de esta aproximación es considerablemente superior al coste de la aproximación SSM.

De forma alternativa, la probabilidad de la pronunciación de entrada se puede estimar mediante una combinación lineal de las probabilidades del UBM y el SSM como sigue:

$$p(X) = \alpha p(X|\lambda_{UBM}) + (1-\alpha)p(X|\lambda_X) \tag{4.6}$$

donde $\alpha > 0$ es un coeficiente de ponderación establecido de forma heurística. La ecuación 4.6 formula la hipótesis de que la fuente de *X* está representada, bien mediante el UBM (que da cobertura a un amplio rango de locutores), o bien mediante el SSM (que da una ligera cobertura a cualquier otra posible fuente de la señal de entrada).

Antecedentes

En la bibliografía se han encontrado muy pocos trabajos orientados a modelar el espacio de impostores y definir una alternativa al UBM:

- En [Siohan99] se estima un modelo simple, de baja resolución acústica como factor de normalización. El trabajo está orientado a la verificación de locutores con un sistema dependiente del texto. Los autores tratan de definir una alternativa que no requiera un volumen de datos elevado para estimar el modelo de referencia, de forma los sistemas se puedan integrar en dispositivos portátiles e inalámbricos. El modelo de locutor, así como el modelo de referencia (ambos son HMM), se estiman a partir de de los datos de entrenamiento del locutor objetivo, pero la resolución acústica del modelo de referencia es considerablemente inferior a la del modelo de locutor (el

modelo de referencia se define con 5 estados mientras que el modelo de locutor se define con 25). De esta forma, los autores evitan que haya algún desajuste (de canal, de entorno, etc.) entre el modelo de locutor y el modelo de referencia. Una aproximación muy similar es la propuesta por Tran en [Tran04]. La propuesta es prácticamente la misma, pero orientada a la verificación de locutores en sistemas independientes del texto y empleando GMM como metodología de modelado. De nuevo, el modelo de locutor y el de referencia se estiman con las mismas pronunciaciones (correspondientes al locutor objetivo), y se define un modelo de menor resolución (específicamente, un modelo que contiene menos componentes gaussianas) para representar el modelo de referencia.

- En [Hsu03], en cambio, se propone un sistema de verificación que no emplea modelos de referencia, por lo que, como en [Siohan99] y [Tran04], el sistema se define únicamente a partir de las pronunciaciones de los locutores objetivo. En este caso, al igual que en el SSM, se estima un modelo acústico (un GMM) a partir de la señal de entrada. El sistema utiliza este modelo para calcular la probabilidad de una pronunciación de entrenamiento del locutor hipotético. La pronunciación a evaluar se selecciona cuidadosamente, siguiendo un algoritmo específico que determina cuál es la pronunciación de entrenamiento con menor desviación estándar para los modelos generados con el resto de pronunciaciones de ese locutor (véase [Hsu03] para más información). Si la probabilidad estimada no supera un umbral específico del locutor (estimado a partir de las señales de entrenamiento), se decide que el locutor hipotético es un impostor, y viceversa.
- Tran y Wagner [Tran01] proponen una metodología muy simple para reducir los errores de falsa alarma en sistemas de verificación. Consiste en añadir una constante de pertenencia al factor de normalización, $\varepsilon > 0$, factor que establece un mínimo grado de pertenencia de las señales de evaluación al grupo de impostores:

$$p(X) = p(X|\lambda_{UBM}) + \varepsilon \tag{4.7}$$

  Esta constante desempeña la misma función que el SSM (al combinarlo con el UBM), con la diferencia de que el SSM no es una constante establecida a ciegas o empíricamente, sino extraída de la señal a evaluar.
- En [Tsai01] se propone la aplicación de las aproximaciones CLR y BIC en la identificación de locutores sobre conjuntos cerrados. Estas metodologías, descritas en el apartado 2.5.1, son muy empleadas en tareas de seguimiento y diarización de locutores. CLR es una aproximación bilateral, que calcula el LLR de forma recíproca: estima la probabilidad del modelo generado con la pronunciación de entrenamiento para la pronunciación a evaluar, y viceversa. Los autores defienden que esta aproximación puede compensar el rendimiento de los sistemas basados en LLR, especialmente cuando el volumen de los datos de entrenamiento es escaso y en consecuencia, la estimación de los parámetros del modelo no se puede realizar de forma robusta. Por otro lado, de forma alternativa, se propone la aplicación de la aproximación BIC con objeto de analizar si la

probabilidad del modelo generado con la unión de las pronunciaciones de entrenamiento y test, es superior a estimar un modelo para cada pronunciación.

- En [Aronowitz05] los autores proponen una aproximación alternativa al LLR. Al igual que en la aproximación SSM, se estima un GMM(MAP) a partir del UBM por medio de la señal de test. A continuación se calcula el ratio entre: (1) la distancia del GMM(MAP) de la señal de test y el GMM(MAP) de la pronunciación de entrenamiento (es decir, el modelo de locutor); y (2) la distancia entre el GMM(MAP) de la señal de test y el modelo UBM. La métrica que definen para calcular la distancia entre dos GMM es similar a la distancia KL.

Por último, hay que destacar que una versión preliminar del SSM, y empleada como único factor de normalización en tareas de seguimiento de locutores sobre noticias de radio y televisión, se propuso por primera vez en el trabajo [Rodriguez07].

4.1.3.3 Configuración del modelo superficial de fuente como factor de normalización

Con el objetivo de determinar la resolución óptima del SSM (la cantidad de componentes gaussianas) como factor de normalización del modelo de locutor, se han llevado a cabo experimentos preliminares. En base a las hipótesis realizadas más arriba, cabe esperar que la cantidad de componentes del modelo será reducida, dado que un SSM con una cantidad elevada de componentes gaussianas representaría características específicas de la señal de entrada, y no su fuente.

En la **Figura 4.7** se muestran las curvas DET de los sistemas de identificación de locutores que emplean sólo el modelo SSM como factor de normalización.

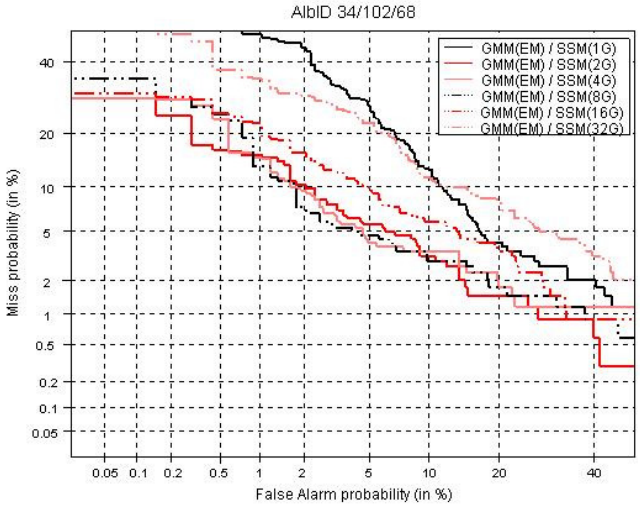

Figura 4.7. *Curvas DET de los 6 sistemas de identificación de locutores que emplean, respectivamente, modelos SSM con 1, 2, 4, 8, 16 y 32 componentes como factor de normalización. Los modelos de locutor son GMM(EM). La experimentación se ha llevado a cabo sobre la partición "AlbID 34/102/68" de Albayzín, definido para tareas de identificación en conjuntos abiertos.*

La experimentación se ha llevado a cabo sobre la partición "AlbID 34/102/68" (Figura 4.1(a)). Los modelos SSM son GMM estimados en 10 iteraciones (este número de iteraciones se ha fijado

empíricamente a aquél que ha mostrado el mejor equilibrio entre coste computacional y rendimiento de clasificación). Con objeto de analizar la influencia que ejerce la resolución del SSM en el rendimiento del sistema, se han definido 6 sistemas, con modelos SSM que contienen, respectivamente, 1, 2, 4, 8, 16 y 32 componentes. En la Figura 4.7 se presentan sus curvas DET. Se observa que el mejor rendimiento se obtiene para los modelos SSM de 4 componentes: el EER de este sistema, que no emplea ningún conjunto adicional al de entrenamiento y test, es del 4.9%. En cuanto al comportamiento general de los 6 sistemas, se observa que se cumple la hipótesis realizada más arriba: el modelo de fuente debe ser un modelo simple y de baja resolución, de forma que no modele la características específicas de la señal. De ahí el adjetivo "superficial" (*Shallow*, en ingles) que aparece en el nombre. El EER de los sistemas SSM con 2, 4 y 8 componentes se sitúa en torno al 5%. A medida que se aumenta la cantidad de componentes, el rendimiento de los sistemas se deteriora de forma considerable. El SSM con 16 componentes produce un EER del 7.5%, y con 32 componentes un EER del 10%, el mismo rendimiento que se obtiene con un SSM de una única gaussiana.

En la **Figura 4.8** se muestra la comparación de los mismos sistemas para la verificación de locutores.

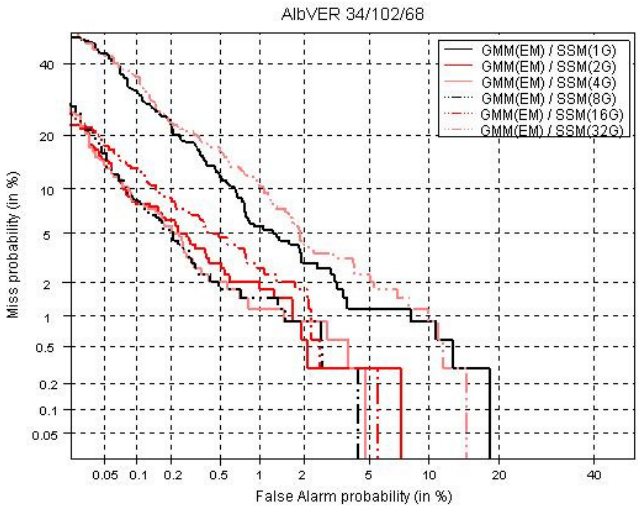

Figura 4.8. *Curvas DET de los seis sistemas de verificación de locutores que emplean, respectivamente, modelos SSM con 1, 2, 4, 8, 16 y 32 componentes como factor de normalización. Los modelos de locutor son GMM(EM). La experimentación se ha llevado a cabo sobre la partición "AlbVER 34/102/68" de Albayzín, definida para tareas de verificación.*

En este caso, la experimentación se ha llevado a cabo sobre la partición "AlbVER 34/102/68" (Figura 4.4(a)). De nuevo, el mejor rendimiento se obtiene con el sistema SSM de 4 componentes, que presenta un EER del 1%. Al comparar el rendimiento de los 6 sistemas, se vuelve a observar la misma tendencia que en tareas de identificación: al aumentar la resolución del SSM el rendimiento de los sistemas empeora. El sistema que emplea SSM de 16 componentes produce un EER del 2%, que es del 3.5% para el que emplea 32 componentes. La curva de este último sistema es muy similar al del SSM con una única gaussiana. Este comportamiento resalta de nuevo el hecho de que el modelo de fuente

debe ser un modelo simple, cuya resolución no ha de ser ni demasiado pequeña, ni demasiado grande. También es posible que la resolución óptima del SSM dependa de la longitud de la señal entrada.

4.1.3.4 Combinación del UBM y el modelo superficial de fuente en la identificación de locutores

En este apartado, se analiza el rendimiento de diferentes aproximaciones para la normalización de la probabilidad de un modelo de locutor, entre ellas, la combinación de UBM y SSM. Tal como se ha mencionado anteriormente, el UBM y el SSM se pueden combinar linealmente, para, de esta forma, resolver sus respectivas limitaciones. El análisis se realiza para: (1) modelos de locutor GMM(EM), en la primera parte del apartado; y (2) modelos de locutor GMM(MAP), posteriormente.

Modelos de locutor GMM(EM)

Para los modelos de locutor GMM(EM), los cuatro sistemas analizados son:

(1) *GMM(EM)*: sistema que emplea el UBM como factor de normalización.

(2) *GMM(EM)/SSM*: sistema que sólo emplea el SSM como factor de normalización (con su resolución óptima, es decir, con 4 gaussianas).

(3) *GMM(EM)/UBM+SSM*: sistema que analiza la combinación lineal del UBM y el SMM como factor de normalización (véase la ecuación 4.6). En experimentos preliminares, se ha analizado el rendimiento de este sistema para diferentes coeficientes en la combinación, concretamente para $\alpha=[0.1, 0.2,..., 0.9]$. Se ha observado que el mejor rendimiento se obtiene cuando el coeficiente α es alto, concretamente cuando $\alpha\cong 0.8, 0.9$. Por tanto, en una segunda batería de experimentos se han analizado coeficientes α en el rango $\alpha=[0.81, 0.99]$. El mejor rendimiento se obtiene con $\alpha = 0.97$, valor que se ha establecido como definitivo.

(4) *GMM(EM)/UBM+Uniform*: sistema que se basa en la propuesta de Tran y Wagner [Tran01], que consiste en añadir una constante $\varepsilon>0$, un factor mínimo de pertenencia, al modelo de referencia (véase la ecuación 4.7). En el artículo se utiliza $\varepsilon=0.01$. La propuesta trata de compensar el ratio para situaciones en las que las dos probabilidades toman valores muy pequeños (sobre todo, la probabilidad del modelo de referencia), y en consecuencia, los errores de falsa alarma aumentan. Nótese, por ejemplo, que resultaría posible distinguir los dos siguientes ratios, 0.01/0.02 y 0.0000001/0000002, que dan como resultado el mismo valor: 0.5. Se han llevado a cabo experimentos preliminares para determinar el valor óptimo de la constante en el rango $\varepsilon=[0.1,0.09,..., 0.001]$, en los que el mejor rendimiento se ha obtenido para $\varepsilon=0.005$. Este sistema se incluye como sistema de contraste para la aproximación que combina linealmente el UBM y el SSM.

En la **Figura 4.9** se muestran las curvas DET de estos 4 sistemas sobre la partición "AlbID 34/102/68*"* (Figura 4.1(a)).

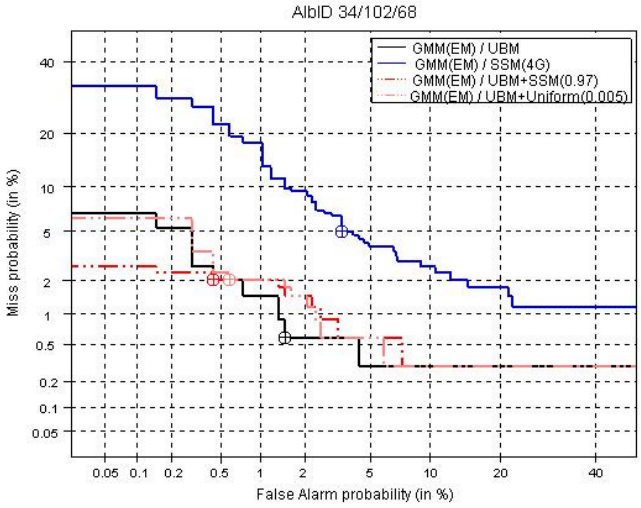

Figura 4.9. *Curvas DET sobre la partición "AlbID 34/102/68" de Albayzín, definida para tareas de identificación en conjuntos abiertos, de los cuatro siguientes sistemas: (1) GMM(EM), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(EM)/SSM, sistema que emplea el SSM como factor de normalización; (3) GMM(EM)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (4) GMM(EM)/UBM+Uniform, sistema que añade una constante de pertenencia al modelo de referencia para compensar el ratio y disminuir las falsas alarmas. Los puntos circulares indican el C_{DET}^{min} de cada sistema.*

Tabla 4.1. *EER y C_{DET}^{min} sobre la partición AlbID 34/102/68 de Albayzín, definida para tareas de identificación en conjuntos abiertos, de los siguientes cuatro sistemas: (1) GMM(EM), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(EM)/SSM, sistema que emplea el SSM como factor de normalización; (3) GMM(EM)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (4) GMM(EM)/UBM+Uniform, sistema que añade una constante al modelo de referencia para compensar el ratio y disminuir las falsas alarmas.*

AlbID 34/102/68		
Sistema	**EER**	C_{DET}^{min}
GMM(EM)	1.4	0.01029
GMM(EM)/SSM	4.86	0.04191
GMM(EM)/UBM+SSM	1.54	0.0125
GMM(EM)/UBM+Uniform	1.55	0.01323

La tasa EER y el coste mínimo C_{DET}^{min} de cada uno de estos sistemas se presentan en la **Tabla 4.1**. En los resultados obtenidos se observa que todos los sistemas, a excepción del sistema que emplea únicamente SSM como factor de normalización (sistema SSM, en adelante), obtienen un alto rendimiento (con EER en torno al 1.5%). De estos tres sistemas, es el sistema GMM(EM) (denominado sistema UBM, en adelante), el de mejor rendimiento. De todas formas, los sistemas que como factor de normalización aplican, respectivamente, la combinación lineal de UBM y SSM (denominado UBM+SSM) y el UBM suavizado con una constante (UBM+Uniform, en adelante) obtienen una curva DET muy similar. Resulta difícil realizar una comparación cualitativa con tasas tan pequeñas, y por tanto, no se puede considerar que uno de estos tres sistemas sea mejor que otro (en las curvas DET de la Figura 4.9, algunos puntos de operación de los sistemas UBM+SSM y

UBM+Uniform están por debajo de la curva correspondiente a UBM). El sistema SSM, en cambio, presenta un rendimiento inferior al resto (con EER del 5%). Es una tasa aceptable si se tiene en cuenta que este sistema, a diferencia del resto, no requiere ningún conjunto adicional para estimar los modelos, únicamente pronunciaciones del locutor objetivo y la propia señal a evaluar.

Por otro lado, en la **Figura 4.10** se presenta el rendimiento de los mismos sistemas sobre la partición "AlbID+Mismatched 34/102/68", que incluye un conjunto de señales no modeladas (Figura 4.1(a+b)). Las tasas EER y los costes C_{DET}^{min} de los sistemas se muestran en la **Tabla 4.2.** En los resultados obtenidos, destaca el alto rendimiento que obtiene, frente al resto, el sistema UBM+SSM (con un EER del 1.1%). Este resultado destaca la validez del SSM para representar de forma adecuada las señales con fuentes acústicas no modeladas en entrenamiento, ya que aumenta considerablemente la robustez del UBM. El rendimiento de los sistemas UBM o UBM+Uniform se deteriora al insertar señales no modeladas en el conjunto de test. En el caso del GMM(EM), el EER ha aumentado del 1.4% (Figura 4.9) al 8.1% (Figura 4.10). Se observa la misma tendencia con el UBM+Uniform, cuyo EER ha subido del 1.55% (Figura 4.9) al 8.05% (Figura 4.10).

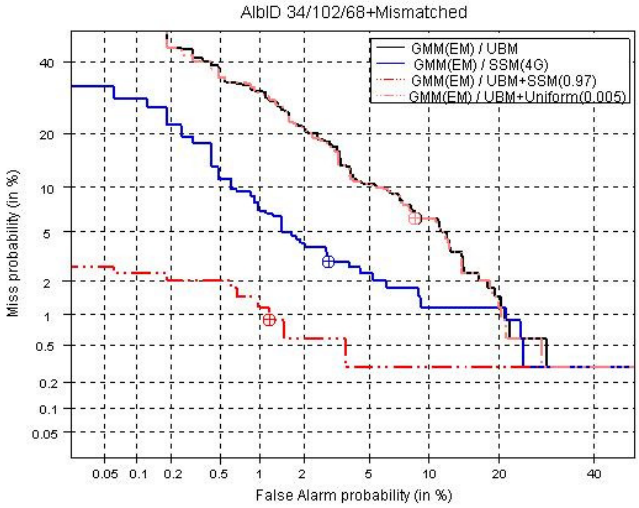

Figura 4.10. *Curvas DET sobre la partición "AlbID+Mismatched 34/102/68", que incluye señales con fuentes no modeladas y se ha definido para tareas de identificación sobre conjuntos abiertos, de los siguientes cuatro sistemas: (1) GMM(EM), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(EM)/SSM, sistema que emplea un SSM para normalizar la probabilidad del locutor; (3) GMM(EM)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (4) GMM(EM)/UBM+Uniform, sistema que añade una constante de pertenencia al modelo de referencia para compensar el ratio y disminuir las falsas alarmas. Los puntos circulares indican el* C_{DET}^{min} *de cada sistema.*

Por tanto, a diferencia del UBM, o el UBM suavizado, la combinación de UBM y SSM da muy buena cobertura a las señales no modeladas y las rechaza debidamente. En la partición "AlbID 34/102/68", que sólo contiene señales modeladas (Figura 4.9), la tasa P_{fa} de los tres sistemas basados en el UBM se sitúa en torno al 0.5% cuando $P_{miss}\cong 2\%$. Sin embargo, en la partición "AlbID+Mismatched 34/102/68", la cual incluye señales no modeladas (Figura 4.10), los sistemas UBM y UBM+Uniform, obtienen una tasa $P_{fa}\cong 15\%$ en el mismo punto de operación $P_{miss}\cong 2\%$. El

sistema UBM+SSM, en cambio, se mantiene en un 0.5%. Por último, destaca el buen comportamiento que presenta el SSM como factor de normalización (con un EER del 3.2%). En este caso, su rendimiento es considerablemente superior al de UBM.

Tabla 4.2. *EER y C_{DET}^{min} sobre la partición "AlbID+Mismatched 34/102/68", que incluye señales con fuentes no modeladas y se ha definido para tareas de identificación sobre un conjunto abierto, de los siguientes cuatro sistemas: (1) GMM(EM), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(EM)/SSM, sistema que emplea el SSM como factor de normalización; (3) GMM(EM)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (4) GMM(EM)/UBM+Uniform, sistema que añade una constante al modelo de referencia para compensar el ratio y disminuir las falsas alarmas.*

AlbID+Mismatched 34/102/68		
Sistema	**EER**	C_{DET}^{min}
GMM(EM)	8.1	0.07448
GMM(EM)/SSM	3.2	0.02904
GMM(EM)/UBM+SSM	1.1	0.0102
GMM(EM)/UBM+Uniform	8.05	0.7387

Modelos de locutor GMM(MAP)

Para los modelos de locutor GMM(MAP) se analizan tres tipos de normalización:

(1) *GMM(MAP)*: sistema que utiliza la probabilidad del UBM como factor de normalización.

(2) *GMM(MAP)/UBM+SSM*: sistema que utiliza la combinación lineal del UBM y el SMM como factor de normalización (ecuación 4.6). Al igual que con modelos GMM(EM), se ha analizado el rendimiento del sistema con diferentes coeficientes en la combinación, de donde se ha obtenido que, de nuevo, el mejor coeficiente es α=0.97.

(3) *GMM(MAP)/UBM+Uniform*: sistema que añade una constante ε>0 al modelo de referencia (ecuación 4.7). Como con los modelos GMM(EM) se han realizado experimentos preliminares para establecer el valor óptimo de la constante, obteniendo ε=0.005 como valor óptimo.

En la **Figura 4.11** se muestra el rendimiento de los tres sistemas sobre la partición "AlbID 34/102/68" (Figura 4.1(a)). Las tasas EER y el coste mínimo de estos sistemas se muestran en la **Tabla 4.3**. Atendiendo al EER y al coste C_{DET}^{min}, el mejor sistema es UBM+SSM. Con respecto al sistema GMM(MAP) (denominado, UBM en adelante), obtiene una mejora relativa del 20%. Su curva es, para casi todos los puntos de operación, superior a los otros sistemas. De todas formas, cabe destacar que los tres sistemas presentan un buen comportamiento, ningún sistema sube de un EER del 3.5%. Los sistemas UBM y UBM+Uniform presentan prácticamente la misma curva DET.

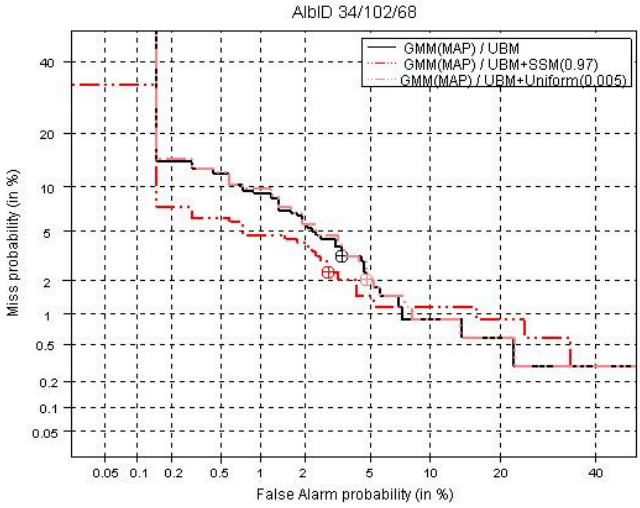

Figura 4.11. *Curvas DET sobre la partición "AlbID 34/102/68" de Albayzín, definida para tareas de identificación en conjuntos abiertos. Se evalúan los tres sistemas siguientes: (1) GMM(MAP), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(MAP)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (3) GMM(MAP)/UBM+Uniform, sistema que añade una constante de pertenencia al modelo de referencia para compensar el ratio y disminuir las falsas alarmas. Los puntos circulares indican el C_{DET}^{min} de cada sistema.*

Tabla 4.3. *EER y C_{DET}^{min} sobre la partición "AlbID 34/102/68", definida para tareas de identificación en conjuntos abiertos, de los tres sistemas siguientes: (1) GMM(MAP), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(MAP)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (3) GMM(MAP)/UBM+Uniform, sistema que añade una constante al modelo de referencia para compensar el ratio y disminuir las falsas alarmas.*

AlbID 34/102/68		
Sistema	**EER**	C_{DET}^{min}
GMM(MAP)	3.5	0.03309
GMM(MAP)/UBM+SSM	2.8	0.02573
GMM(MAP)/UBM+Uniform	3.55	0.03382

Por último, en la **Figura 4.12** se presenta el rendimiento de los mismos tres sistemas sobre la partición "AlbID+Mismatched 34/102/68", que incluye un conjunto de señales no modeladas (Figura 4.1(a+b)). Las tasas EER y los costes C_{DET}^{min} se muestran en la **Tabla 4.4.** Al igual que al emplear modelos de locutor GMM(EM), en este caso vuelve a destacar el alto rendimiento del sistema UBM+SSM. Obtiene una mejora relativa del 47.14% en EER, en promedio, frente al resto de sistemas. Asimismo, la curva DET presenta una mejoría notable para la mayoría de puntos de operación.

Por otro lado, cabe destacar que el rendimiento de los sistemas UBM, o el UBM+Uniform, se deteriora en mucho menor escala que al emplear modelos de locutor GMM(EM), probablemente debido a que los modelos GMM(MAP) son más robustos frente a señales no modeladas, dado que no son modelos independientes, sino adaptados. Generalmente, las fuentes no modeladas en el UBM, tampoco lo estarán en el modelo de locutor, y en consecuencia, el ratio estimado será bajo, y por tanto la señal será rechazada. En todo caso, algunas señales no modeladas tienen mejor cobertura mediante

la combinación del UBM y el SSM, probablemente debido a que no se puede garantizar que una fuente no modelada en el UBM tampoco lo estará en el modelo de locutor adaptado, y viceversa. En este sentido, sobre la partición "AlbID 34/102/68", que contiene únicamente señales modeladas, al emplear modelos GMM(MAP) la tasa P_{fa} del sistema combinado se sitúa en torno al 3.5% en el punto de operación P_{miss}=2%, y en torno al 5% con los otros sistemas (Figura 4.11). Sobre la partición que incluye señales no modeladas (Figura 4.12), en el mismo punto de operación P_{miss}=2%, la tasa P_{fa} decrece al 1.5% con el sistema combinado y se mantiene en torno al 5% con el sistema UBM (o UBM+Uniform). Esto es debido a que el SSM permite rechazar casi todas las señales no modeladas.

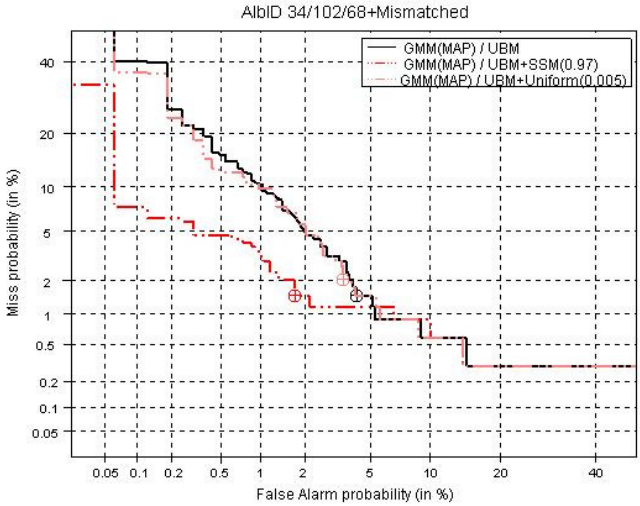

Figura 4.12. *Curvas DET sobre la partición "AlbID+Mismatched 34/102/68" que contiene un conjunto de test con señales no modeladas y se ha definido para tareas de identificación en conjuntos abiertos, de los tres sistemas siguientes: (1) GMM(MAP), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(MAP)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (3) GMM(MAP)/UBM+Uniform, sistema que añade una constante de pertenencia al modelo de referencia para compensar el ratio y disminuir las falsas alarmas. Los puntos circulares indican el C_{DET}^{min} de cada sistema.*

Tabla 4.4. *EER y C_{DET}^{min} sobre la partición "AlbID+Mismatched 34/102/68", que contiene un conjunto de señales con fuentes no modeladas y se ha definido para tareas de identificación sobre conjuntos abiertos, de los tres sistemas siguientes: (1) GMM(MAP), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(MAP)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (3) GMM(MAP)/UBM+Uniform, sistema que añade una constante al modelo de referencia para compensar el ratio y disminuir las falsas alarmas.*

AlbID+Mismatched 34/102/68		
Sistema	**EER**	C_{DET}^{min}
GMM(MAP)	3.5	0.02808
GMM(MAP)/UBM+SSM	1.85	0.01589
GMM(MAP)/UBM+Uniform	3.45	0.02767

4.1.3.5 *Combinación del UBM y el modelo superficial de fuente en la verificación de locutores*

En este apartado, se repite el análisis de diferentes aproximaciones de normalización del modelo de locutor para tareas de verificación. Al igual que en los sistemas de identificación, la normalización se analiza para modelos GMM(EM), en la primera parte, y modelos GMM(MAP), a continuación.

Modelos de locutor GMM(EM)

Para los modelos locutor GMM(EM) se analizan las mismas aproximaciones que en la identificación:

(1) GMM(EM)

(2) GMM(EM)/SSM

(3) GMM(EM)/UBM+SSM

(4) GMM(EM)/UBM+Uniform

Véase en el apartado 4.2.3.4, la sección "Modelos de locutor GMM(EM)", la descripción de cada sistema.

En la **Figura 4.13** se muestran las curvas DET de estos cuatro sistemas sobre la partición "AlbVER 34/102/68*"* (Figura 4.4(a)). Todos los sistemas proporcionan un alto rendimiento, con EER por debajo del 1.5%. En la **Tabla 4.5** se muestra la tasa EER y el coste mínimo C_{DET}^{min} de cada sistema.

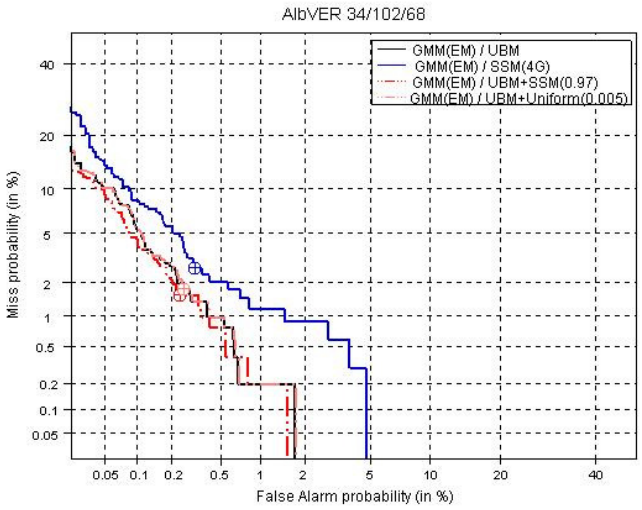

Figura 4.13. *Curvas DET sobre la partición "AlbVER 34/102/68" de Albayzín, definida para tareas de verificación, de los 4 sistemas siguientes: (1) GMM(EM), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(EM)/SSM, sistema que emplea un SSM para normalizar la probabilidad del locutor; (3) GMM(EM)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (4) GMM(EM)/UBM+Uniform, sistema que añade una constante de pertenencia al modelo de referencia para compensar el ratio y disminuir las falsas alarmas. Los puntos circulares indican el C_{DET}^{min} de cada sistema.*

Es especialmente notable el buen comportamiento que presentan los sistemas GMM(EM) (UBM, en adelante), el UBM suavizado (UBM+Uniform) y la combinación lineal de UBM y SSM (UBM+SSM). Las curvas de estos tres sistemas se solapan, por lo que no se puede considerar que el rendimiento de

un sistema sea superior a otro. Quizás destaca el UBM+SSM, el cual presenta una curva ligeramente inferior en la mayoría de puntos del barrido del umbral (véase en la tabla 4.5 cómo su EER y $C_{\mathrm{DET}}^{\mathrm{min}}$ son también ligeramente inferiores). El sistema que sólo emplea el SSM como factor de normalización (denominado sistema SSM), presenta un rendimiento inferior con un EER del 1.1%.

Tabla 4.5. *EER y C_{DET}^{min} sobre la partición "AlbVER 34/102/68" de Albayzín, definida para tareas de verificación, de los cuatro sistemas siguientes: (1) GMM(EM), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(EM)/SSM, sistema que emplea un SSM para normalizar la probabilidad del locutor; (3) GMM(EM)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (4) GMM(EM)/UBM+Uniform, sistema que añade una constante al modelo de referencia para compensar el ratio y disminuir las falsas alarmas.*

AlbVER 34/102/68		
Sistema	**EER**	C_{DET}^{min}
GMM(EM)	0.7	0.004148
GMM(EM)/SSM	1.1	0.005732
GMM(EM)/UBM+SSM	0.55	0.003856
GMM(EM)/UBM+Uniform	0.75	0.004225

En la **Figura 4.14** se muestra el rendimiento de los mismos cuatro sistemas sobre la partición "AlbVER+Mismatched 34/102/68", que incluye señales no modeladas (Figura 4.4(a+b)).

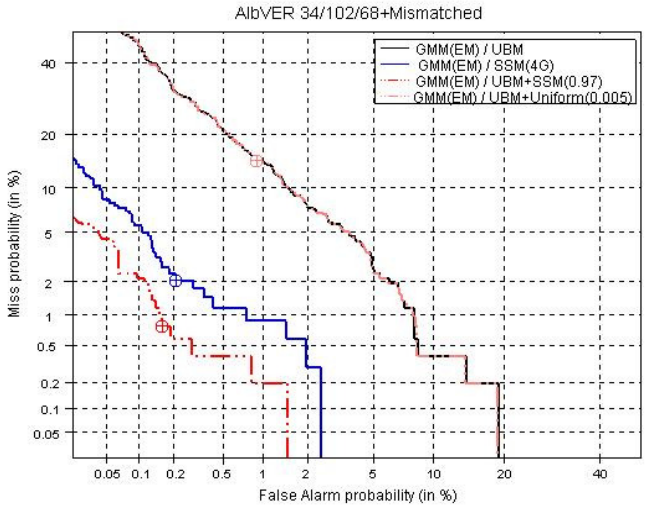

Figura 4.14. *Curvas DET sobre la partición "AlbVER+Mismatched 34/102/68", que contiene un conjunto señales no modeladas y se ha definido para tareas de verificación, de los cuatro sistemas siguientes: (1) GMM(EM), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(EM)/SSM, sistema que emplea el SSM como factor de normalización; (3) GMM(EM)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (4) GMM(EM)/UBM+Uniform, sistema que añade una constante de pertenencia al modelo de referencia para compensar el ratio y disminuir las falsas alarmas. Los puntos circulares indican el C_{DET}^{min} de cada sistema.*

Destaca el alto rendimiento obtenido por el sistema UBM+SSM, cuyo EER es del 0.4% (con una mejoría del 91.49% con respecto a UBM).El sistema SSM presenta la segunda mejor curva, con un EER de 0.95% (y una mejoría del 79.79% con respecto a UBM). Los sistemas UBM y UBM+Uniform presentan un deterioro notable: el EER del sistema UBM se incrementa del 0.7% (Figura 4.13) al 4.7% (Figura 4.14) al añadir señales no modeladas en el conjunto de test. El UBM+Uniform presenta un comportamiento idéntico, con un EER que sube del 0.75% (Figura 4.13) al 4.7% (Figura 4.14). Estos resultados indican que el SSM representa de forma adecuada las señales que provienen de fuentes acústicas no modeladas, y aumenta la robustez del sistema, de forma considerable, al combinarlo linealmente con el UBM. Las tasas EER y los costes C_{DET}^{min} de los sistemas se presentan en la **Tabla 4.6**.

Tabla 4.6. *EER y C_{DET}^{min} sobre la partición "AlbVER+Mismatched 34/102/68", que incluye un conjunto de señales no modeladas y se ha definido para tareas de verificación. Los sistemas analizados son: (1) GMM(EM), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(EM)/SSM, sistema que emplea el SSM como factor de normalización; (3) GMM(EM)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (4) GMM(EM)/UBM+Uniform, sistema que añade una constante al modelo de referencia para compensar el ratio y disminuir las falsas alarmas.*

AlbVER+Mismatched 34/102/68		
Sistema	**EER**	C_{DET}^{min}
GMM(EM)	4.7	0.02334
GMM(EM)/SSM	0.95	0.004113
GMM(EM)/UBM+SSM	0.4	0.002344
GMM(EM)/UBM+Uniform	4.7	0.02331

Modelos de locutor GMM(MAP)

En lo que respecta a modelos de locutor GMM(MAP), se analizan los tres sistemas siguientes:

(1) GMM(MAP)

(2) GMM(MAP)/UBM+SSM

(3) GMM(MAP)/UBM+Uniform

Véase en el apartado 4.2.3.4, la sección "Modelos de locutor GMM(MAP)", la descripción de los sistemas.

En la **Figura 4.15** se presenta el rendimiento de estos tres sistemas sobre la partición "AlbVER 34/102/68" (Figura 4.4(a)). Los tres sistemas obtienen tasas de error P_{fa} y P_{miss} muy bajas en toda la curva. El EER está por debajo del 1% en todos los sistemas (véase la **Tabla 4.7**). La curva DET del UBM+SSM es algo mejor al resto, especialmente para las tasas P_{fa} más pequeñas. Tal como se ha mencionado varias veces a lo largo de este apartado, este comportamiento podría ser debido a que el SSM da cobertura frente a señales que provienen de fuentes no modeladas debidamente en entrenamiento.

Tabla 4.7. *EER y C_{DET}^{min} sobre la partición "AlbVER 34/102/68" de Albayzín, definida para tareas de verificación, de los siguientes cuatro sistemas: (1) GMM(MAP) sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(MAP)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (3) GMM(MAP)/UBM+Uniform, sistema que añade una constante al modelo de referencia para compensar el ratio y disminuir las falsas alarmas.*

AlbVER 34/102/68		
Sistema	**EER**	C_{DET}^{min}
GMM(MAP)	0.85	0.005462
GMM(MAP)/UBM+SSM	0.65	0.0042211
GMM(MAP)/UBM+Uniform	0.85	0.005497

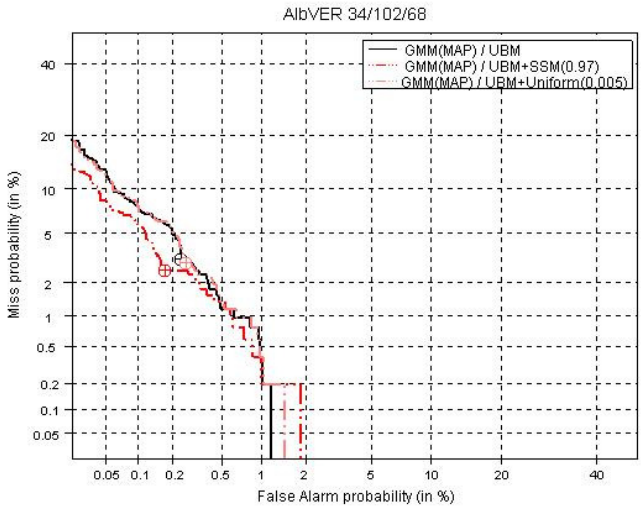

Figura 4.15. *Curvas DET sobre la partición "AlbVER 34/102/68" de Albayzín, definida para tareas de verificación, de los tres sistemas siguientes: (1) GMM(MAP), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(MAP)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (3) GMM(MAP)/UBM+Uniform, sistema que añade una constante de pertenencia al modelo de referencia para compensar el ratio y disminuir las falsas alarmas. Los puntos circulares indican el C_{DET}^{min} de cada sistema.*

Por último, en la **Figura 4.16** se muestra el rendimiento de mismos sistemas sobre la partición "AlbVER+Mismatched 34/102/68", que incluye un conjunto de señales no modeladas (Figura 4.4(a+b)). Al igual que sobre la partición que sólo contiene señales modeladas (véase la Figura 4.15), los tres sistemas presentan un rendimiento muy alto, y su EER está por debajo del 1% en todos los casos (véase la **Tabla 4.8**). El mejor rendimiento, y la mejor curva para todos los puntos de operación, se obtiene con el sistema UBM+SSM. De hecho, obtiene la mejor tasa EER (del 0.4%) y $C_{\mathrm{DET}}^{\mathrm{min}}$.

Al igual que en los experimentos de identificación, se observa que los modelos de locutor GMM(MAP) dan mejor cobertura a señales que provienen de fuentes no modeladas, y que el SSM amplia la cobertura (o la robustez) del UBM ante señales no modeladas. El sistema UBM+SSM presenta una mejoría significativa respecto a la clásica aproximación UBM: su EER es de 0.4%, mientras que el del UBM es de 0.7% (Figura 4.16). Para la partición que sólo contiene señales modeladas (Figura 4.15), está diferencia es más pequeña: el EER del sistema combinado es de 0.65%,

y el del UBM de 0.85%. Por tanto, se puede concluir que la contribución principal del SSM está relacionada con el rechazo de señales no modeladas. Para el sistema UBM+SSM la tasa P_{fa} decrece, en el punto de operación P_{miss}=0.5%, del 0.9% (Figura 4.15) al 0.4% (Figura 4.16), al evaluar señales no modeladas (una mejora relativa del 55.56%). Para el UBM, en el punto de operación P_{miss}=0.5%, la tasa P_{fa} decrece en menor escala, concretamente, del 1% (Figura 4.15) al 0.65% (Figura 4.16), una mejora relativa del 35%. El sistema UBM rechaza debidamente una gran proporción de señales no modeladas, pero no tantas como el sistema UBM+SSM.

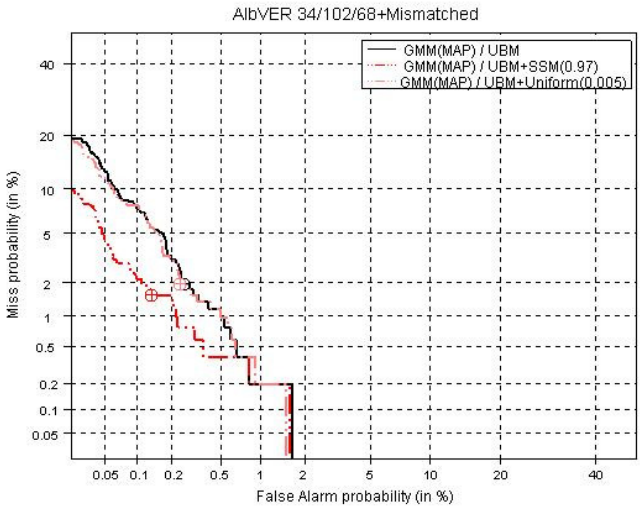

Figura 4.16. *Curvas DET sobre la partición "AlbVER+Mismatched 34/102/68", que contiene un conjunto de señales no modeladas y se ha definido para tareas de verificación, de los tres sistemas siguientes: (1) GMM(MAP), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(MAP)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (3) GMM(MAP)/UBM+Uniform, sistema que añade una constante de pertenencia al modelo de referencia para compensar el ratio y disminuir las falsas alarmas. Los puntos circulares indican el C_{DET}^{min} de cada sistema.*

Tabla 4.8. *EER y C_{DET}^{min} sobre la partición "AlbVER+Mismatched 34/102/68", que contiene un conjunto de señales no modeladas y se ha ha definido para tareas de verificación. Los sistemas analizados son: (1) GMM(MAP), sistema que emplea el UBM como factor de normalización (aproximación clásica); (2) GMM(MAP)/UBM+SSM, sistema que emplea la combinación lineal de UBM y SSM como factor de normalización; y (3) GMM(MAP)/UBM+Uniform, sistema que añade una constante al modelo de referencia para compensar el ratio y disminuir las falsas alarmas.*

AlbVER+Mismatched 34/102/68		
Sistema	**EER**	C_{DET}^{min}
GMM(EM)	0.7	0.004436
GMM(EM)/UBM+SSM	0.4	0.002874
GMM(EM)/UBM+Uniform	0.7	0.004266

4.1.4 Conclusiones del apartado 4.2

En esta primera parte del capítulo se ha descrito la fase de experimentación llevada a cabo con sistemas identificación y verificación de locutores que siguen metodologías clásicas de modelado.

Concretamente, modelos GMM(EM) y modelos GMM(MAP) (estos últimos adaptados a partir de un UBM). La evaluación se ha llevado a cabo sobre particiones definidas sobre la base de datos Albayzín, grabada en condiciones de laboratorio, con señales en castellano, que provienen de 204 locutores. Con objeto de analizar el rendimiento de los sistemas frente a señales generadas a partir de fuentes no modeladas, se ha definido un conjunto de test adicional, denominado *Mismatched*, formado por señales externas a Albayzín: algunas de música, otras grabadas por línea telefónica y algunas otras descargadas aleatoriamente desde Internet.

En esta línea, se ha propuesto una nueva metodología de modelado para representar la fuente de la señal de entrada, y mejorar la robustez de los sistemas ante señales que contienen fuentes no modeladas. La aproximación propuesta se denomina modelo superficial de fuente (SSM, de forma abreviada). Es un GMM de baja resolución, es decir compuesto por una pequeña cantidad de componentes, y estimado a partir de la propia señal a evaluar. Suponiendo que el SSM se ha estimado de forma perfecta, se podría emplear como único factor de normalización, dado que indicaría el grado de aproximación del modelo de locutor a la fuente de la señal a evaluar. Sin embargo, en los experimentos realizados se observa que el SSM presenta una gran sensibilidad a las variaciones específicas de la señal, y en consecuencia puede no representar de forma robusta la fuente de señal de entrada. Como solución, se ha propuesto una combinación lineal del UBM y el SSM para por un lado mejorar la cobertura del UBM y, por otro, ampliar la robustez del SSM.

En las **Tablas 4.9 y 4.10** se presenta el rendimiento de los sistemas de identificación (tabla 4.9) y verificación (tabla 4.10) sobre el corpus de test que contiene únicamente señales modeladas (*"No Mismatch"* en la tabla) y el que incluye un conjunto adicional de señales no modeladas (*"Mismatch"* en la tabla). La configuración de estas particiones de describe en el apartado 4.2.1 para las tareas de identificación en conjuntos abiertos (véase la Figura 4.1), y en el apartado 4.2.2 para tareas de verificación (véase la Figura 4.4). Se presenta el rendimiento de dos sistemas, empleando modelos de locutor GMM(EM) y modelos de locutor GMM(MAP): (1) la aproximación clásica que utiliza la probabilidad del UBM como factor de normalización de la probabilidad del locutor (denominado sistema UBM, en adelante); y (2) la aproximación que utiliza una combinación lineal de las probabilidades del UBM y el SSM como factor de normalización (sistema UBM+SSM, en adelante).

Para el sistema UBM, los modelos GMM(MAP) presentan un rendimiento mejor que los modelos GMM(EM). Su rendimiento se mantiene en el mismo margen al incluir señales no modeladas en el conjunto de test. Al emplear modelos GMM(EM), en cambio, el rendimiento se deteriora considerablemente al incluir señales no modeladas en el corpus de test. Así, la tasa EER del sistema de identificación UBM se incrementa de 1.4% al 8.1% (tabla 4.9), y de 0.7% al 4.7% al utilizar el mismo sistema para la verificación (tabla 4.10). Se cree que este comportamiento puede ser debido a que en los sistemas UBM que emplean modelos de locutor GMM(EM) el ratio se calcula a partir de dos modelos estimados de forma completamente independiente. En consecuencia, dada una señal no modelada (ni en el modelo de locutor ni en el UBM) las probabilidades condicionales de estos

modelos para la señal son impredecibles. De ahí, la probabilidad del locutor puede ser arbitrariamente superior o inferior a la probabilidad del UBM, por lo que el ratio se convierte prácticamente en un valor aleatorio. En los sistemas UBM con modelos GMM(MAP), en cambio, los dos modelos del ratio están relacionados, debido a que el modelo del locutor es un modelo adaptado, estimado a partir del UBM. De hecho, representa el desplazamiento relativo al locutor de las fuentes representadas en el UBM, por lo que generalmente, una fuente no representada en el UBM, tampoco lo estará en el modelo de locutor. Dada una señal no modelada, las probabilidades condicionales de los modelos GMM(MAP) y UBM obtendrán un valor similar para la señal, un valor pequeño por lo general, y en consecuencia, su ratio apuntará hacia el rechazo de este tipo de señales, como debe ser. Por otro lado, debe aclararse que el conjunto "*Mismatched*" consta de señales cuyo contenido acústico es muy diferente al contenido de las señales de entrenamiento, suponiendo además que el locutor objetivo está ausente (por ejemplo, cuando la señal es un fragmento de una canción, el ladrido de un perro, etc.).

Tabla 4.9. *EER de los sistemas de identificación de locutores en conjuntos abiertos. Se analiza el rendimiento de modelos de locutor GMM(EM) y GMM(MAP). Como factor de normalización, se analizan el UBM y la combinación lineal del UBM y el SSM. La experimentación se lleva a cabo sobre las particiones "AlbID 34/102/68"("No mismatch" en la tabla), y "AlbID+Mismatched 34/102/68" que incluye señales no modeladas en en conjunto de test ("Mismatch" en la tabla).*

	EER de los experimentos de identificación en un conjunto de locutores abierto		
	Modelo de Locutor / **Modelo de Referencia**	**GMM(EM)**	**GMM(MAP)**
No Mismatch	**UBM**	1.4	3.5
	UBM+SSM	1.54	2.8
Mismatch	**UBM**	8.1	3.5
	UBM+SSM	1.1	1.85

Por otro lado, una de las conclusiones más significativas que se puede extraer de esta experimentación es el alto rendimiento que presenta el sistema UBM+SSM propuesto en esta tesis. Es, en todos los casos, el sistema con el menor EER, a excepción de la tarea de identificación sobre señales modeladas ("*No Mismatch*" en la tabla 4.9), en la que el sistema UBM (con EER 1.4%) obtiene un rendimiento ligeramente superior al del sistema combinado (con EER 1.54%). Los resultados obtenidos demuestran que el SSM, combinado linealmente con el UBM, mejora la cobertura del UBM. Esta mejoría es muy significativa con modelos de locutor GMM(EM) sobre el conjunto de test con señales no modeladas, dado que el SSM tiene un gran margen de mejora debido a que los sistemas UBM que emplean modelos GMM(EM) presentan un comportamiento muy frágil ante señales no modeladas. Como sistema de contraste a la combinación del UBM y el SSM, se ha analizado el sistema propuesto por Tran [Tran01], que suaviza la probabilidad del UBM mediante una

constante de pertenencia, con objeto de compensar el ratio cuando se manejan señales no modeladas. Este sistema (cuyos resultados se ofrecen en los apartados 4.2.3.4 y 4.2.3.5) presenta un rendimiento muy similar al del sistema UBM. En la mayoría de los experimentos su rendimiento es ligeramente mejor que el del UBM, pero la mejora observada es muy pequeña con respecto a la mejoría que se obtiene con la combinación de UBM y SSM.

Tabla 4.10. *EER de los sistemas de verificación de locutores. Se analiza el rendimiento de modelos de locutor GMM(EM) y GMM(MAP). Como factor de normalización, se analizan el UBM y la combinación lineal del UBM y el SSM. La experimentación se lleva a cabo sobre las particiones "AlbVER 34/102/68" ("No mismatch" en la tabla), y "AlbVER+Mismatched 34/102/68" que incluye señales no modeladas en en conjunto de test ("Mismatch" en la tabla).*

	EER de los experimentos de verificación de locutores		
	Modelo de Locutor / **Modelo de Referencia**	**GMM(EM)**	**GMM(MAP)**
No Mismat	**UBM**	0.7	0.85
	UBM+SSM	0.55	0.65
Mismat	**UBM**	4.7	0.7
	UBM+SSM	0.4	0.4

Por último, cabe destacar que en tareas de identificación (tabla 4.9), los modelos de locutor GMM(EM) presentan (inesperadamente) mejor rendimiento que los modelos de locutor GMM(MAP), cuando el conjunto de test contiene únicamente señales modeladas. Esto puede ser debido a que los modelos de locutor se estiman con 30 o 45 segundos de señal en promedio, lo que probablemente es insuficiente para estimar modelos GMM(MAP) robustos.

4.2 Experimentos de verificación de locutores en la NIST SRE06

Los experimentos que se reportan en esta segunda parte del capítulo se han llevado sobre bases de datos de las evaluaciones NIST SRE 2006 y anteriores. El desarrollo de los sistemas se ha realizado con las bases de datos definidas para las NIST SRE de 2004 y 2005. Para la evaluación de los sistemas se ha utilizado el corpus de la tarea principal de la NIST SRE de 2006 (véase el plan de evaluación en [NIST SRE]). Se trata de una tarea de verificación de locutores (a veces, también denominada "detección de locutores"): dado un locutor objetivo (del cual tan sólo se dispone de un fragmento de señal de unos 150 segundos) y una señal de test (también de unos 150 segundos), se ha de determinar si el locutor objetivo habla en la señal de test.

Esta tarea consta de 51068 experimentos, de los cuales 3616 corresponden a señales del locutor objetivo y 47452 corresponden a señales de un impostor. Cada experimento establece qué señal de test

y qué modelo de locutor objetivo se han de comparar. Además del listado de experimentos, se definen dos listados necesarios para estimar los modelos, uno para hombres, y otro para mujeres, que determinan qué segmento de entrenamiento corresponde a cada locutor objetivo. En esta evaluación se definen 354 modelos de hombres, y 462 de mujeres.

Asimismo, se define un conjunto de reglas (o restricciones) para la evaluación. Entre ellas, destacan las siguientes:

- Cada experimento debe ser independiente al resto, es decir, las decisiones se han de tomar sólo a partir de la información extraída del segmento de test y el modelo del locutor objetivo. No se puede emplear información extraída de otros segmentos u otros modelos. Por ejemplo, no se puede aplicar ninguna técnica de normalización que considere más de un segmento o más de un modelo, ni emplear los datos de la propia evaluación para modelar el espacio de impostores.
- El género de los locutores objetivo es un dato conocido, e indicado de forma explícita en el listado de modelos. Generalmente no se definen experimentos cruzados para este tipo de tareas, es decir, al definir el conjunto de test no se plantean experimentos donde el género del locutor objetivo y el género del locutor real sean diferentes.
- Todas las señales del conjunto de evaluación (entrenamiento y test) siguen el formato SPHERE (*SPeech HEader REsources*), un formato de archivo definido por el NIST y empleado en ficheros de audio. Los ficheros SPHERE, después de una cabecera (generalmente de 1024 bytes), almacenan en formato binario las muestras de la señal de voz. La cabecera consta de diversa información sobre el contenido del fichero: el número de muestras de la señal, la frecuencia de muestreo, la cantidad de canales, la codificación de las muestras, si las muestras están comprimidas o no, así como ciertos datos de interés insertados por el creador del fichero. En la cabecera de los ficheros de la evaluación se incluye el idioma de la conversación y se indica si la conversación se ha grabado o no por canal telefónico. Toda la información proporcionada en la cabecera se puede emplear tanto para generar los modelos como para realizar la clasificación.

En cuanto al idioma de las conversaciones, cabe destacar que la mayoría son en inglés. El conjunto de evaluación contiene algunas señales de entrenamiento y test en árabe, mandarín, ruso y español (adquiridas principalmente de locutores bilingües), con objeto de analizar si al emplear un idioma diferente en entrenamiento y test el rendimiento de los sistemas empeora. De hecho, en las NIST SRE, como resultado oficial adicional, se analiza el rendimiento de los sistemas para el subconjunto constituido por experimentos de la tarea principal en los que el idioma de los segmentos de entrenamiento y test coincide y es inglés. En la NIST SRE de 2006, este subconjunto consta de 23687 experimentos, de los cuales 1846 corresponden a señales del locutor objetivo y 21841 corresponden a señales de un impostor.

Todos los segmentos de entrenamiento y test de la tarea principal se extraen de conversaciones telefónicas, en las que se solicita a los participantes que indiquen cuál es el canal de transmisión (móvil, fijo inalámbrico, fijo normal), así como el terminal (*head-mounted* (de cabeza), terminal

estándar, *ear-bud* (auricular), teléfono personal) empleado. Cabe destacar que, entre las tareas adicionales definidas para la NIST SRE de 2006, hay una tarea de verificación sobre segmentos de test grabados de forma local mediante micrófonos.

4.2.1 Sistemas definidos para la NIST SRE06

Para la evaluación, se han desarrollado dos sistemas:

(1) GMM(MAP) estimados a partir de un UBM

(2) GMM-SVM, con vectores soporte GMM(MAP)

Debido a que en las NIST SRE se conoce el género del locutor objetivo, se estiman modelos UBM dependientes del género (*gender-dependant*): un modelo UBM a partir de segmentos de hombres (denominado UBM-m, de forma abreviada), y otro a partir de segmentos de mujeres (denominado UBM-f, de forma abreviada). En experimentos preliminares (confirmando resultados de trabajos de la bibliografía), se ha observado un rendimiento mejor al emplear modelos UBM dependientes del género que al emplear modelos UBM mixtos. En este trabajo, los modelos UBM constan, cada uno, de 512 componentes gaussianas, estimadas mediante el algoritmo EM en 20 iteraciones. Los segmentos empleados para estimar estos modelos se han extraído del conjunto de entrenamiento del corpus de la NIST SRE de 2004. Este corpus contiene señales con características muy similares al corpus de evaluación: se compone de conversaciones telefónicas grabadas por los mismos canales de transmisión y empleando el mismo tipo de terminales. Con objeto de proporcionar una cobertura adecuada, el material de entrenamiento se ha seleccionado detalladamente, equilibrando los segmentos, en la medida de lo posible, por canal de transmisión y tipo de terminal empleado. Así, se incluyen, como máximo, 50 segmentos de cada combinación canal-terminal, escogidos de forma aleatoria. En consecuencia, el UBM-f se estima a partir de 325 segmentos, y el UBM-m a partir de 350 segmentos.

Los modelos de locutor GMM(MAP) se adaptan a partir del UBM correspondiente al género del locutor objetivo. El coeficiente de adaptación se ha fijado heurísticamente a $\tau=16$ y se adaptan únicamente las medias de las componentes (como en la mayoría de aproximaciones de la bibliografía). La probabilidad de una señal para un modelo de este tipo se estima mediante la estrategia *top-N*, es decir, para cada vector se evalúan las N componentes de mayor probabilidad. En este trabajo, $N=8$.

Por otro lado, los modelos GMM-SVM se entrenan con una muestra positiva correspondiente al locutor objetivo y varias muestras negativas correspondientes a locutores impostores. El sistema debe definir un hiperplano de separación para cada locutor objetivo que separe su región del espacio vectorial de la región del resto de locutores. El modelo de locutor se construye con las muestras que determinan dicho hiperplano. Las muestras del conjunto de impostores, que es el mismo para todos los locutores objetivo, sean hombres o mujeres, se obtienen a partir de 1174 segmentos del conjunto de test del corpus de la NIST SRE de 2004 (evidentemente, este conjunto de impostores debe ser diferente al conjunto empleado para estimar los modelos UBM). La muestra positiva, en cambio, se deriva del segmento de entrenamiento correspondiente al locutor objetivo. En el conjunto de

impostores se incluyen tanto mujeres como hombres, y todo tipo de canales de transmisión y terminales de teléfono que supuestamente se emplean en el conjunto de evaluación. Cada muestra a clasificar es un vector generado a partir del GMM(MAP) estimado con los datos de cada muestra. Debido a que el sistema define modelos UBM dependientes del género, la adaptación MAP del segmento, sea del locutor objetivo o de un impostor, se realiza a partir del UBM que coincide con el género del locutor objetivo. Dado el GMM(MAP) de un segmento y el UBM a partir del cual se ha estimado, el vector se genera de la siguiente manera:

- En primer lugar, se han de concatenar las componentes gaussianas del modelo adaptado.
- Para representar el desplazamiento frente al UBM, y no el valor en el modelo adaptado, a la media del modelo adaptado se le resta la media de la gaussiana correspondiente en el UBM.
- Asimismo, la media de cada gaussiana se normaliza multiplicándola por el ratio entre la raíz cuadrada de su peso y su varianza (véase la ecuación 2.39). Debido a que sólo se adaptan las medias de las componentes, el peso y la varianza de cada componente adaptada coincide con el peso y la varianza de la misma componente en el UBM.
- Por último, en experimentos preliminares se ha observado que el rendimiento de las máquinas de vectores soporte aumenta cuando los vectores expandidos se normalizan, es decir, las componentes se dividen por la norma del vector.

Los experimentos con modelos GMM-SVM se han realizado mediante la herramienta de software libre LibSVM [Chang01], diseñada para el entrenamiento y la evaluación de sistemas SVM. Por otro lado, la dimensión de los vectores es de 10240, debido a que los modelos UBM se definen con 512 componentes y los vectores parametrizados constan de 20 parámetros.

El proceso de parametrización de los conjuntos de datos de las NIST SRE, se inicia descartando los tramos cuya energía es inferior a 30 dB de la energía máxima de la señal, y aplicando un filtro pasa-banda (300Hz-3300Hz) para eliminar una parte del ruido del canal. Las señales, adquiridas a 8 kHz, se analizan en tramos de 20 milisegundos solapados a intervalos de 10 milisegundos. Después, se aplica una ventana de *Hamming* y se calcula una FFT de 256 puntos. Las amplitudes de la FFT se promedian en 20 filtros triangulares solapados, con frecuencias centrales y anchuras según la escala de Mel de percepción auditiva. A continuación, se aplica una DCT sobre el logaritmo de las amplitudes del filtro y se obtienen 10 MFCC. Para aumentar la robustez frente al ruido del canal, se aplican una CMS y FW. Por último, se estiman las primeras derivadas de los parámetros MFCC, obteniendo vectores acústicos de 20 dimensiones.

4.2.2 Resultados de los sistemas definidos para la NIST SRE06

En la **Figura 4.17** se muestra el rendimiento de ambos sistemas para la tarea principal de la NIST SRE06 (experimentos *all*, en adelante), y para un subconjunto de éste, constituido por experimentos cuyos segmentos de entrenamiento y test contienen únicamente habla inglesa (experimentos *eng* en adelante). La tasa EER, y las funciones de coste C_{DET}, C_{DET}^{min}, C_{llr} y C_{llr}^{min} de cada sistema se muestran en

la **Tabla 4.11**. Al igual que en las evaluaciones del NIST (véase [NIST SRE]), los parámetros C_{fa} , C_{miss} y P_{obj} se fijan, respectivamente, a 1.0, 10.0 y 0.01. Se observa un rendimiento notablemente superior del sistema GMM-SVM, tanto en los experimentos *all*, como en los experimentos *eng*. El EER de este sistema es de 8.33% para los experimentos *all*, lo que supone una mejora del 19.28% con respecto a GMM(MAP), cuyo EER es del 10.32%. Para los experimentos *eng*, el EER de GMM-SVM es del 6.39%, superando el rendimiento de GMM(MAP) en un 27.22% (el EER de éste es del 8.78%). Estos resultados ponen de manifiesto el deterioro del sistema al considerar segmentos de test en un idioma diferente al de entrenamiento. El EER de GMM(MAP) mejora en un 14.92%. El sistema GMM-SVM presenta un comportamiento similar, con una mejora del 23.29%.

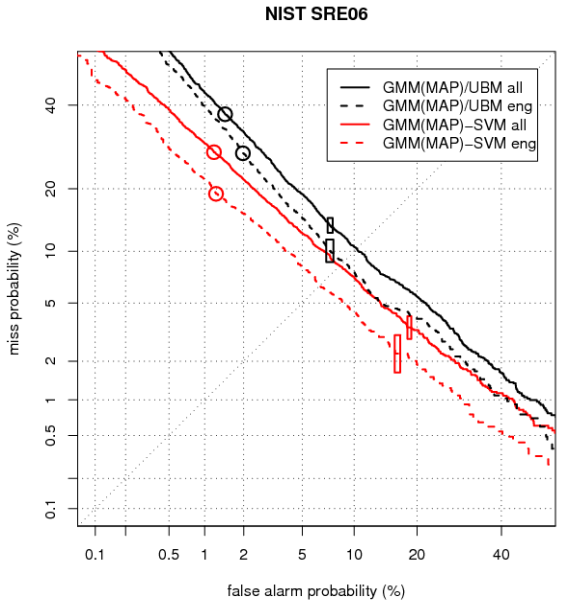

Figura 4.17. *Curvas DET de los sistemas GMM(MAP) y GMM-SVM para la tarea principal de la NIST SRE06 ("all") y para un subconjunto de éste compuesto únicamente por segmentos en ingles ("eng"). El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo, C_{DET}^{min}.*

Tabla 4.11. *Tasa EER y costes C_{llr}, C_{llr}^{min},C_{DET} y C_{DET}^{min} de los sistemas GMM(MAP) y GMM-SVM sobre la tarea principal de la NIST SRE06 ("all") y sobre un subconjunto de éste compuesto únicamente por segmentos en ingles ("eng").*

NIST SRE06		C_{llr}	C_{llr}^{min}	EER	C_{DET}	C_{DET}^{min}
all	GMM(MAP)	0.9537	0.3619	10.32	0.0864	0.05186
	GMM-SVM	0.9965	0.2946	8.33	0.1867	0.03968
eng	GMM(MAP)	0.9496	0.3151	8.78	0.0826	0.0472
	GMM-SVM	0.9962	0.2335	6.39	0.1643	0.0312

En cuanto a los valores de coste mínimo C_{DET}^{min} y C_{llr}^{min} que indican, respectivamente, el coste C_{DET} mínimo en el barrido del umbral y el coste C_{llr} con la calibración óptima de probabilidades, se observa la misma tendencia. GMM-SVM presenta mejor rendimiento que GMM(MAP), y el rendimiento de

ambos sistemas mejora considerablemente sobre los experimentos *eng*. Los valores de C_{DET} y C_{llr} obtenidos en la práctica son bastante peores, debido a que la evaluación se ha realizado sin calibrar ni normalizar las probabilidades (la calibración y normalización de probabilidades mediante datos externos se analiza en los siguientes apartados). En este caso, el umbral de cada sistema se ha fijado de forma empírica a 0.00 para GMM(MAP), y a 0.002 para GMM-SVM.

4.2.3 Resultados de la normalización de probabilidades en los sistemas definidos para la NIST SRE06

En este apartado se describe la aplicación de las normalizaciones Z-norm, T-norm y ZT-norm sobre los dos sistemas, GMM(MAP) y GMM-SVM, desarrollados para su evaluación en la NIST SRE06. Los conjuntos de impostores utilizados para la Z-norm y la T-norm son disjuntos. Se han extraído del conjunto de test de la base de datos de la NIST SRE08. Debido a que la T-norm es más costosa, su conjunto de impostores es más reducido.

El conjunto de pronunciaciones utilizado para estimar la Z-norm es mixto, dado que debe proporcionar cobertura a cualquier señal de entrada. Consta de 550 segmentos extraídos de conversaciones telefónicas, equilibradas en la medida de lo posible en cuanto al género del locutor, canal y terminal empleados en la conversación (con 35 señales de cada tipo, como máximo).

Puesto que se utilizan UBM dependientes del género, para la T-norm se definen dos conjuntos de impostores: un conjunto para locutores objetivo que son hombres (T-norm-m), y otro para mujeres (T-norm-f). El conjunto T-norm-m se compone de 117 segmentos, correspondientes a 117 hombres diferentes. El conjunto T-norm-f reúne 120 segmentos correspondientes a 120 mujeres diferentes. Estos dos conjuntos se definen de la forma más equilibrada posible en cuanto a canales de transmisión y terminales de teléfono (con 15 señales de cada tipo, como máximo).

Resultados de las normalizaciones Z-norm y T-norm

En las **Figuras 4.18 y 4.19** se muestran el rendimiento obtenido al aplicar las aproximaciones Z-norm y T-norm sobre las probabilidades del sistema GMM(MAP), para la tarea principal de la NIST SRE06 (Figura 4.18, experimentos *all,* en adelante), y para un subconjunto de éste, constituido por experimentos que contienen únicamente habla inglesa (Figura 4.19, experimentos *eng,* en adelante).

En ambas figuras se observa la misma tendencia:

- La Z-norm mejora el rendimiento del sistema original (sin normalización) prácticamente para todos los puntos de operación de la curva. Se observa una mejora superior sobre los experimentos *eng* (Figura 4.19) que los experimentos *all* (Figura 4.18).
- La T-norm mejora el rendimiento del sistema original (sin normalización) para los puntos con tasas de falsa alarma, P_{fa}, más pequeñas, concretamente para $P_{fa}<EER$. Sin embargo, a partir de este punto, el rendimiento del sistema empeora considerablemente. En comparación con la Z-norm, sólo cabe destacar la mejora obtenida por la T-norm para los puntos con tasas de falsa alarma pequeñas. Estos resultados sugieren la conveniencia de usar la T-norm en aplicaciones de

alta seguridad, con costes C_{fa} o probabilidades P_{fa} elevados (es decir, cuando la prioridad de las aplicaciones es evitar la intrusión de impostores).

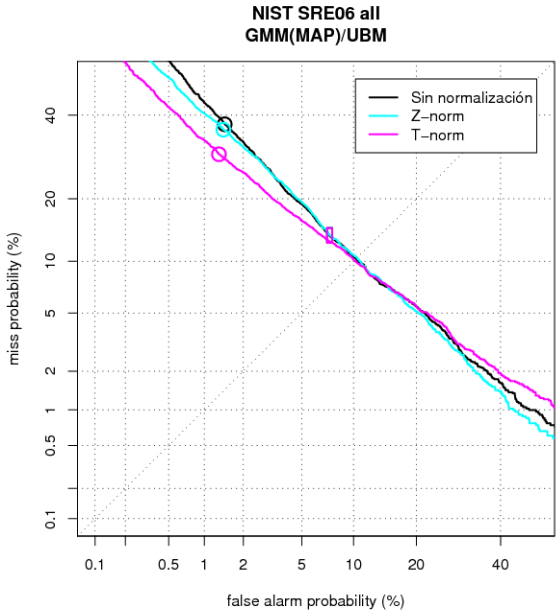

Figura 4.18. *Curvas DET de los experimentos con Z-norm y T-norm sobre el sistema GMM(MAP) para la tarea principal de la NIST SRE 2006. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

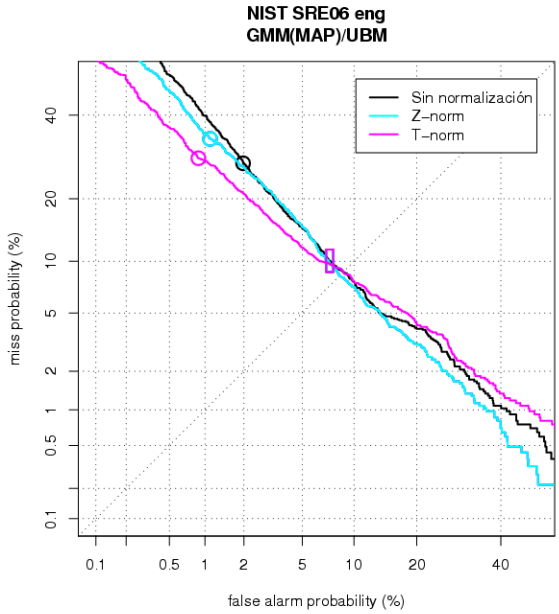

Figura 4.19. *Curvas DET de los experimentos con Z-norm y T-norm sobre el sistema GMM(MAP) para el subconjunto de experimentos en ingles de la tarea principal de la NIST SRE06. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

En las Figuras **4.20 y 4.21** se muestra el rendimiento de Z-norm y T-norm sobre el sistema GMM-SVM para los experimentos *all* (Figura 4.20), y los experimentos *eng* (Figura 4.21).

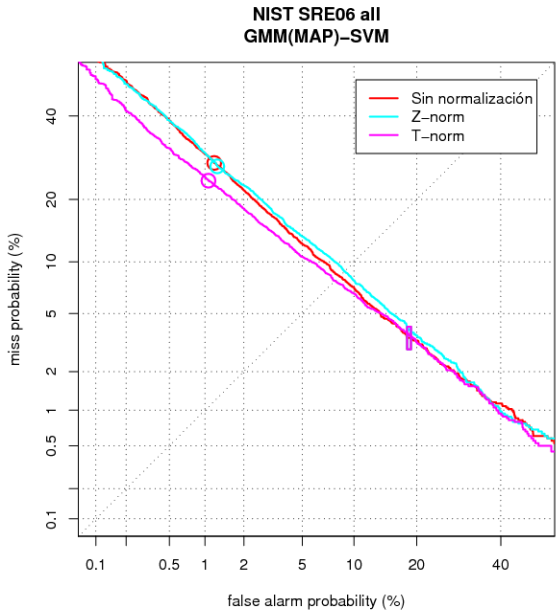

Figura 4.20. *Curvas DET de los experimentos con Z-norm y T-norm sobre el sistema GMM-SVM para la tarea principal de la NIST SRE06. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

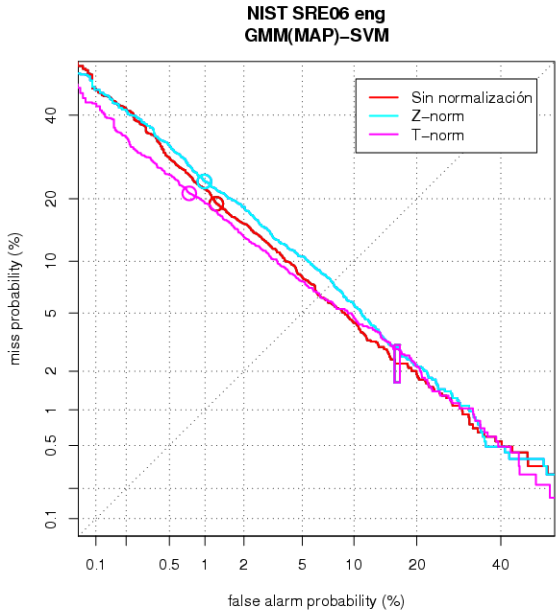

Figura 4.21. *Curvas DET de los experimentos con Z-norm y T-norm sobre el sistema GMM-SVM para el subconjunto de experimentos en ingles de la tarea principal de la NIST SRE06. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

Se observa que:

- La Z-norm deteriora (o en el mejor de los casos, mantiene) el rendimiento del sistema original (sin normalización) para todos los puntos de operación del barrido de umbral. Este comportamiento se da tanto sobre los experimentos *all* como sobre los experimentos *eng*.
- La T-norm, en cambio, presenta un mejor comportamiento que sobre el sistema GMM(MAP). Sobre los experimentos *all*, su rendimiento es superior al del sistema original para todos los puntos de operación de la curva. Sobre los experimentos *eng* se observa un comportamiento similar, aunque para algunos puntos de operación con tasas altas de falsa alarma, $P_{fa}>EER$, el rendimiento de la T-norm es inferior (aunque sólo ligeramente) al del sistema original.

Resultados de las normalizaciones ZT-norm y TZ-norm

Tal como se ha mencionado en el estado de arte, en varios trabajos de la bibliografía, así como en los sistemas presentados a las últimas evaluaciones del NIST, se observa una clara tendencia a aplicar las aproximaciones Z-norm y T-norm de forma secuencial, para así eliminar de forma conjunta la variabilidad de la señal (T-norm) y la variabilidad del locutor (Z-norm). Si la T-norm se aplica sobre probabilidades ya normalizadas con Z-norm, la aproximación se denomina ZT-norm. Si la aplicación se realiza en orden inverso, aplicando primero la T-norm, y a continuación la Z-norm, la aproximación se conoce como TZ-norm. La ZT-norm minimiza en primer lugar la variabilidad relacionada con los locutores, tanto la del locutor objetivo, como la de los impostores requeridos para la T-norm, de forma que todos los modelos presenten una distribución de probabilidad similar. A continuación, para eliminar la variabilidad específica de la señal a evaluar, se aplica una T-norm con los modelos ya normalizados, por lo que se asume que todas las señales del conjunto de evaluación presentarán una distribución de probabilidades similar. La TZ-norm, en cambio, garantiza primero una distribución de probabilidades similar para todas las señales del conjunto de evaluación y para las señales del conjunto de impostores requeridos por la Z-norm. Posteriormente, se aplica una Z-norm, para eliminar la variabilidad del locutor sobre señales ya normalizadas con respecto a su propia variabilidad, de forma que los modelos de todos los locutores objetivo proporcionen distribuciones de probabilidad similares.

En la **Figuras 4.22 y 4.23** se muestra el rendimiento de ZT-norm y TZ-norm sobre el sistema GMM(MAP), para los experimentos *all* (Figura 4.22), y los experimentos *eng* (Figura 4.23).

En las dos figuras se observa una tendencia parecida:

- La ZT-norm mejora considerablemente el rendimiento del sistema original (sin normalización) para las tasas de falsa alarma más pequeñas, concretamente para $P_{fa}<EER$. A partir de este punto, el rendimiento es similar al del sistema original, con algunos puntos que presentan un rendimiento inferior al del sistema original (de forma más notable para los experimentos *eng*).
- La TZ-norm también presenta un rendimiento considerablemente superior al del sistema original para los puntos $P_{fa}<EER$. Asimismo, para estos mismos puntos, muestra una curva mejor que la TZ-norm, tanto sobre los experimentos *all*, como sobre los experimentos *eng*. Al igual que con la ZT-norm, cuando $P_{fa}>EER$, la TZ-norm presenta un rendimiento similar al del sistema original,

con algunos puntos en que se muestra tasas de error claramente superiores a las del sistema original (de nuevo, de forma más notable sobre los experimentos *eng*).

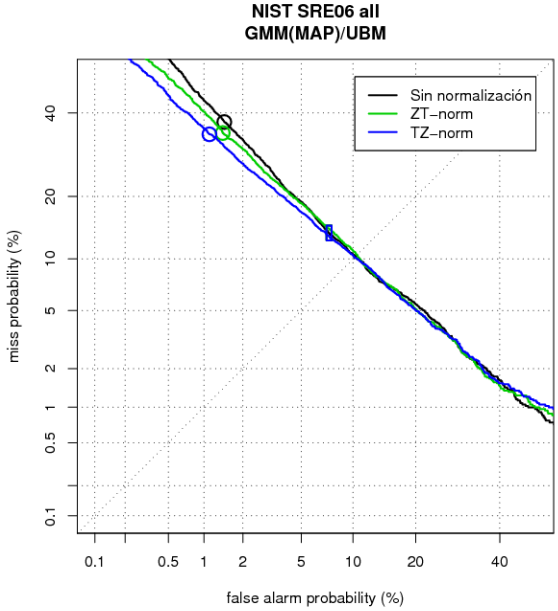

Figura 4.22. *Curvas DET de los experimentos con ZT-norm y TZ-norm sobre el sistema GMM(MAP) para la tarea principal de la NIST SRE de 2006. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El circulo, en cambio, indica el punto de coste mínimo de C_{DET}^{min}.*

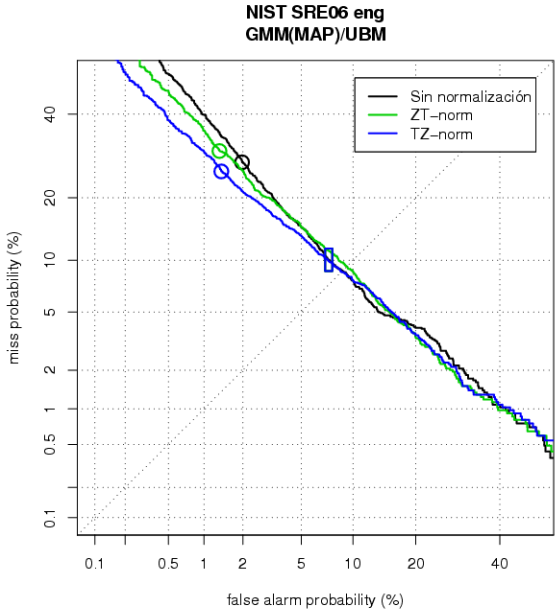

Figura 4.23. *Curvas DET de los experimentos con ZT-norm y TZ-norm sobre el sistema GMM(MAP) para el subconjunto de experimentos en ingles de la tarea principal de la NIST SRE06. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

Por último, en las **Figuras 4.24 y 4.25** se muestran las curvas DET que resultan al aplicar la ZT-norm y la TZ-norm sobre el sistema GMM-SVM, de nuevo sobre el conjunto de experimentos *all* (Figura 4.24), y el conjunto de experimentos *eng* (Figura 4.25).

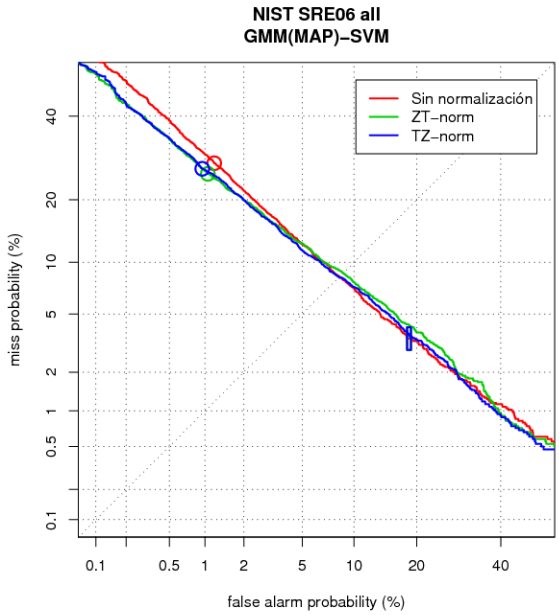

Figura 4.24. *Curvas DET de los experimentos con ZT-norm y TZ-norm sobre el sistema GMM-SVM para la tarea principal de la NIST SRE06. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

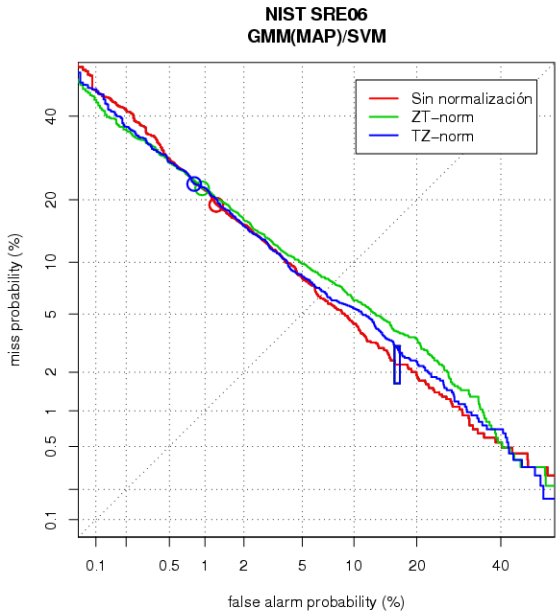

Figura 4.25. *Curvas DET de los experimentos con ZT-norm y TZ-norm sobre el sistema GMM-SVM para el subconjunto de experimentos en inglés de la tarea principal de la NIST SRE06. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

A diferencia del resto de experimentos realizados en este apartado, en este caso se observa un comportamiento muy diferente en los dos conjuntos experimentales. En las curvas obtenidas sobre el conjunto *all*, se observan curvas prácticamente iguales de la TZ-norm y la ZT-norm para $P_{fa}<EER$. Ambas aproximaciones presentan un rendimiento considerablemente mejor al del sistema original, sobre todo para los puntos de P_{fa} más pequeños. No obstante, para $P_{fa}>EER$, la TZ-norm obtiene un rendimiento mejor que la ZT-norm, y similar al del sistema original. Véase como las curvas de ambos sistemas, la del sistema original y la del sistema con TZ-norm, están solapadas, con algunos puntos donde el sistema con TZ-norm es superior al sistema original, y viceversa. Por otro lado, en los resultados obtenidos sobre el subconjunto *eng*, los sistemas con ZT-norm y TZ-norm presentan un rendimiento inferior al del sistema original prácticamente para todos los puntos de operación de la curva (a excepción de aquéllos con tasas P_{fa} más pequeñas, concretamente $P_{fa}<0.5\%$).

Por lo tanto, los resultados obtenidos en la experimentación con sistemas que incorporan ZT-norm y TZ-norm no conducen a conclusiones definitivas. En los experimentos *all* se observa que ambas aproximaciones mejoran el rendimiento del sistema original. Esta mejora es más notable para los puntos de operación con tasas de falsa alarma más pequeñas, hecho que sugiere el empleo de estas metodologías para las aplicaciones de alta seguridad que esperan una gran cantidad de intrusos y quieren minimizar sus efectos, incluso a costa de cometer errores al verificar los locutores objetivo. Sin embargo, al evaluar estas aproximaciones sobre los experimentos *eng* con el sistema GMM-SVM, el rendimiento obtenido es inferior al del sistema original. Queda pendiente analizar con detalle las causas de este comportamiento contradictorio en experimentos futuros, ya que a priori se esperaba que TZ-norm y ZT-norm mejoraran de forma más significativa el rendimiento del sistema original, ya que minimizan de forma conjunta la variablidad de la señal y del locutor. De todas formas, tal como se detalla en las conclusiones del apartado, se cree que este comportamiento podría ser debido a la composición de los conjuntos de impostores T-norm y Z-norm.

Contraste de resultados

A modo de resumen, en la **Tabla 4.12** se presentan las tasas EER y los costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} del sistema GMM(MAP) sin ninguna normalización, y con las metodologías de normalización Z-norm, T-norm, ZT-norm y TZ-norm, analizadas en este apartado. Se observa que al normalizar las probabilidades, los costes C_{llr}, C_{llr}^{min} y C_{DET}^{min} del sistema decrecen, tanto sobre los experimentos *all*, como sobre los experimentos *eng*. La tasa EER, sin embargo, presenta un comportamiento algo irregular: tomando como referencia el sistema original, en algunos casos aumenta y en otros decrece.

Sobre el conjunto de experimentos *all*, el mejor EER se ha obtenido con la TZ-norm (EER 10.07%) con una mejora del 2.42% con respecto al EER del sistema original. El menor coste C_{llr}^{min}, el coste del sistema asumiendo una calibración óptima, se obtiene también con la TZ-norm. Tal como se observaba en las curvas DET (véase la Figura 4.22), la TZ-norm presenta un rendimiento mejor que la ZT-norm. Asimismo, su rendimiento es claramente mejor que el del sistema original para los puntos de la curva con tasas de falsa alarma pequeñas. Este mismo comportamiento se observa para la

T-norm (Figura 4.18), que, en cambio, presenta un rendimiento claramente peor para los puntos correspondientes a tasas de falsa alarma superiores al EER. Debido al buen rendimiento de la T-norm para los puntos con valores de falsa alarma más pequeños (en las NIST SRE la probabilidad P_{fa} de las aplicaciones es muy elevada), es precisamente la T-norm la que proporciona el menor coste C_{DET}^{min}. La Z-norm, en cambio, obtiene un EER ligeramente superior al del sistema original, aunque presenta una curva mejor en la mayoría de puntos de operación (Figura 4.18). El buen comportamiento general de esta curva se refleja en el hecho de que la Z-norm proporciona el mínimo valor de C_{llr}.

Tabla 4.12. *Tasa EER, y costes C_{llr}, C_{llr}^{min},C_{DET} y C_{DET}^{min} del sistema GMM(MAP), sin aplicar ninguna normalización de probabilidades, y con Z-norm, T-norm, ZT-norm y TZ-norm. La evaluación se realiza sobre la tarea principal de la NIST SRE06 ("all") y sobre un subconjunto de experimentos de esta tarea compuesto únicamente por segmentos en ingles ("eng").*

NIST SRE06 GMM(MAP)/UBM		Cllr	Cllr.min	EER	Cdet	Cdet.min
all	Sin normalización	0.9537	0.3619	10.32	0.0864	0.05186
	Z-norm	**0.6714**	0.357	10.45	0.0864	0.05009
	T-norm	0.6918	0.3496	10.09	0.0864	**0.04277**
	ZT-norm	0.7126	0.359	10.38	0.0864	0.0486
	TZ-norm	0.7289	**0.3469**	**10.07**	0.0864	0.04524
eng	Sin normalización	0.9496	0.3151	8.78	0.0826	0.0472
	Z-norm	**0.6559**	**0.3003**	**8.34**	0.0826	0.0446
	T-norm	0.6822	0.304	8.77	0.0826	**0.0376**
	ZT-norm	0.6972	0.3133	9.26	0.0826	0.04353
	TZ-norm	0.7174	0.3017	8.67	0.0826	0.03917

Sobre los experimentos *eng*, TZ-norm presenta mejor rendimiento que ZT-norm (Figura 4.23). La T-norm presenta, de nuevo, el menor coste C_{DET}^{min}, debido a su buen rendimiento para los puntos con tasas de falsa alarma más pequeñas. En cuanto a Z-norm, mejora el rendimiento del sistema original para todos los puntos de operación (Figura 4.19). En este caso, incluso proporciona la mejor tasa EER (con una mejora del 5.01% con respecto al sistema original) y los menores costes C_{llr} y C_{llr}^{min}.

En resumen, los resultados obtenidos con el sistema GMM(MAP) revelan un comportamiento mejor de la Z-norm y la TZ-norm, frente a T-norm y la ZT-norm.

En la **Tabla 4.13** se muestran los resultados experimentales del sistema GMM-SVM. En este caso, la Z-norm presenta un comportamiento considerablemente peor, tanto sobre los experimentos *all* como sobre los experimentos *eng*. Véase cómo, con respecto al sistema original, la Z-norm no mejora sino que empeora la tasa EER y proporciona peores costes C_{llr}^{min} y C_{DET}^{min} (un comportamiento ya observado en las Figuras 4.20 y 4.21). El resto de aproximaciones (T-norm, ZT-norm y TZ-norm) reducen los costes con respecto al sistema original. De nuevo, la TZ-norm es superior a ZT-norm, tanto en los costes C_{llr}^{min} y C_{DET}^{min}, el EER, como sobre las curvas DET obtenidas con ambas aproximaciones (Figura

4.24). En estos experimentos, la mejor aproximación de normalización es la T-norm. Obtiene la menor tasa EER (con una mejora del 5.76% con respecto al sistema original) y los menores costes C_{llr}, C_{llr}^{min} y C_{DET}^{min} de todas las aproximaciones analizadas. Como puede verse en la Figura 4.20, la curva DET del sistema con T-norm es mejor que la del sistema original para todos los puntos de operación.

Tabla 4.13. *Tasa EER, y costes C_{llr}, C_{llr}^{min},C_{DET} y C_{DET}^{min} del sistema GMM-SVM, sin aplicar ninguna normalización de probabilidades, y con Z-norm, T-norm, ZT-norm y TZ-norm. La evaluación se realiza sobre la tarea principal de la NIST SRE06 ("all") y sobre un subconjunto de experimentos de esta tarea compuesto únicamente por segmentos en ingles ("eng").*

NIST SRE06 GMM(MAP)-SVM		Cllr	Cllr.min	EER	Cdet	Cdet.min
all	Sin normalización	0.9965	0.2946	8.33	0.1867	0.03938
	Z-norm	0.6999	0.3061	8.88	0.1867	0.03954
	T-norm	**0.6744**	**0.2767**	**7.85**	0.1867	**0.03445**
	ZT-norm	0.7588	0.2946	8.74	0.1867	0.0369
	TZ-norm	0.8059	0.2877	8.3	0.1867	0.03604
eng	Sin normalización	0.9962	0.2335	6.39	0.1643	0.0312
	Z-norm	0.6683	0.2541	7.64	0.1643	0.0333
	T-norm	**0.6437**	**0.2279**	**6.39**	0.1643	**0.02841**
	ZT-norm	0.7279	0.2587	7.74	0.1643	0.0317
	TZ-norm	0.775	0.2414	6.83	0.1643	0.0313

En los experimentos *eng* se observa una tendencia parecida. La TZ-norm presenta un rendimiento mejor que la ZT-norm, tanto en cuanto al EER como a los costes C_{llr}^{min} y C_{DET}^{min}. Sin embargo, el rendimiento de ambas es peor que el del sistema original, lo que queda reflejado en sus curvas DET (véase la Figura 4.25). Por último, al igual que sobre los experimentos *all*, destaca el rendimiento del sistema con T-norm, que, de nuevo, presenta la mejor tasa EER y los menores costes C_{llr}, C_{llr}^{min} y C_{DET}^{min} de todas las aproximaciones de normalización. En cuanto a su curva DET (Figura 4.21), es mejor que la del sistema original para los puntos correspondientes a las tasas de falsa alarma más pequeñas y ligeramente peor en el resto de puntos.

En definitiva, en lo que al análisis del sistema GMM-SVM respecta, el buen comportamiento que presenta la T-norm es la conclusión más relevante.

4.2.4 Resultados de la fusión de los sistemas definidos para la NIST SRE06

En este apartado se describe la fusión de los sistemas GMM(MAP) y GMM-SVM en experimentos sobre la NIST SRE06. Se analiza también el rendimiento obtenido al fusionar sistemas que aplican normalización de probabilidades, concretamente, con aquellas metodologías que han mostrado un rendimiento mejor en los experimentos realizados en el apartado 4.3.3.

La fusión de sistemas se ha optimizado sobre el conjunto experimental definido para la NIST SRE05 (véase [NIST SRE]). Este corpus se ha reservado exclusivamente para estimar los parámetros de la fusión, debido a que contiene el mismo tipo de señales que el corpus de la tarea principal de la NIST SRE06. El proceso de optimización requiere la estimación de los modelos de los locutores objetivo de la NIST SRE05. Para ello, se emplean los mismos conjuntos de impostores y modelos de referencia que para estimar los modelos de los locutores objetivo de la NIST SRE06. Puesto que no se emplea información auxiliar, la optimización se realiza considerando un único bloque experimental, el conjunto de evaluación definido para la tarea principal de la NIST SRE05.

La herramienta empleada para estimar y aplicar los parámetros de la fusión es *FoCal* [FoCal], que permite realizar la fusión mediante regresión logística. La fusión, además de optimizar el rendimiento, trata de calibrar las puntuaciones del sistema fusionado, por lo que se espera que haya pocas diferencias entre C_{llr}^{min} y C_{llr}, y entre C_{DET}^{min} y C_{DET}. Puesto que las puntuaciones del sistema fusionado se pueden interpretar como probabilidades a posteriori, como umbral de aceptación se establece el umbral de Bayes, $\widehat{\theta_b}$, (el umbral "teórico") definido en la ecuación 2.50. Dados los costes C_{fa} y C_{miss}, y la probabilidad P_{obj}, fijados, respectivamente, a 1.0, 10.0 y 0.01 en las NIST SRE, $\widehat{\theta_b}=2.2925$.

En la **Figura 4.26** se muestra el rendimiento de la fusión de los sistemas GMM(MAP) y GMM-SVM, sin normalización de probabilidades, sobre el propio conjunto de desarrollo, el corpus de la NIST SRE05. La fusión mejora, o en el peor de los casos mantiene, el rendimiento del mejor sistema para todos los puntos de operación de la curva. Asimismo, en la curva del sistema fusionado los puntos C_{DET} (marcado mediante un rectángulo) y C_{DET}^{min} (marcado mediante un círculo) prácticamente coinciden, demostrando que el sistema está debidamente calibrado. Nótese que en las curvas correspondientes a los sistemas originales, los puntos C_{DET} y C_{DET}^{min} están considerablemente alejados, lo que pone de manifiesto la importancia de la fusión también con respecto a la calibración de las puntuaciones.

Las tasas EER y los costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de estos tres sistemas se presentan en la **Tabla 4.14**. Se observa una mejora del 2.85% del sistema fusionado con respecto al EER del mejor subsistema. Asimismo, se observa una gran proximidad entre los costes C_{llr} y C_{llr}^{min}, y C_{DET} y C_{DET}^{min}, respectivamente, sobre todo en comparación con la distancia que presentan los mismos costes en cada subsistema. De aquí, se puede concluir que el sistema fusionado está bien calibrado. Nótese, no obstante, que se trata de una evaluación en modo cerrado, ya que el conjunto de evaluación es el mismo sobre el que se han optimizado los parámetros de la fusión. La verdadera evaluación consiste en aplicar los parámetros estimados para la fusión sobre otro conjunto experimental, tal como se realiza a continuación.

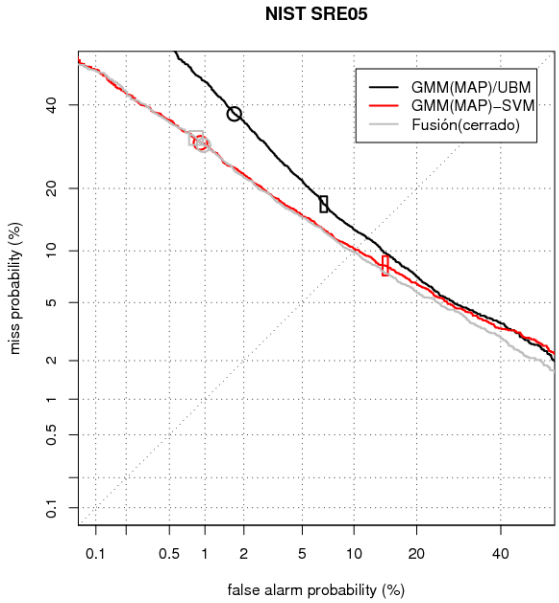

Figura 4.26. *Curvas DET de los sistemas GMM(MAP), GMM-SVM y el sistema que resulta de la fusión – optimizada en cerrado – de ambos sistemas, para la tarea principal de la NIST SRE05, conjunto con el que se entrenan los parámetros de la fusión. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

Tabla 4.14. *Tasa EER y costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de los sistemas GMM(MAP), GMM-SVM y el sistema que resulta de la fusión – optimizada en cerrado – de ambos sistemas. La evaluación se ha llevado a cabo sobre la tarea principal de la NIST SRE05, conjunto reservado para entrenar la fusión.*

NIST SRE05	Cllr	Cllr.min	EER	Cdet	Cdet.min
GMM(MAP)	0.9533	0.4153	11.64	0.08399	0.05434
GMM-SVM	0.9962	0.3571	10.16	0.15079	0.03923
Fusión (cerrado)	0.3615	0.3475	9.87	0.03697	0.03918

En las **Figuras 4.27 y 4.28** se muestra el rendimiento real de la fusión de los sistemas GMM(MAP) y GMM-SVM, para la tarea principal de la NIST SRE06 (Figura 4.27, experimentos *all* en adelante) y para un subconjunto de éste, constituido únicamente por experimentos de habla inglesa (Figura 4.28, experimentos *eng*). Se observa que la curva correspondiente a la fusión de sistemas mejora, o mantiene, el rendimiento del mejor subsistema, prácticamente para todos los puntos de operación de la curva. Asimismo, se observa una gran proximidad entre los puntos C_{DET} y C_{DET}^{min}, lo que demuestra que la fusión produce puntuaciones debidamente calibradas. Además, puesto que los parámetros de la fusión se han estimado sobre el corpus de desarrollo y presentan también buen rendimiento sobre el corpus de evaluación, se puede concluir que el procedimiento de fusión se ha generalizado de forma adecuada.

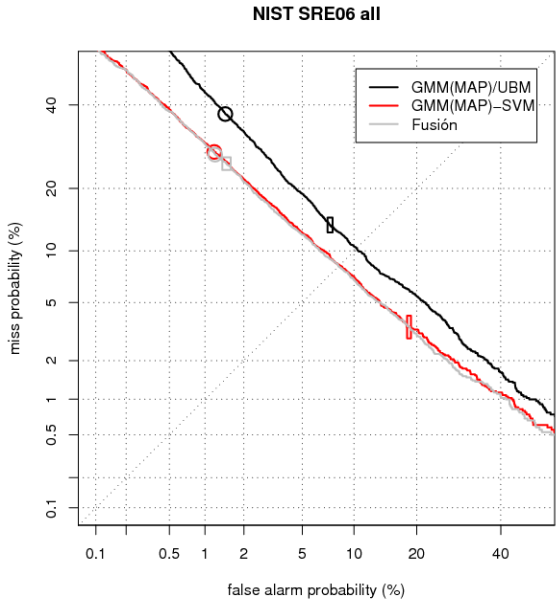

Figura 4.27. *Curvas DET de los sistemas GMM(MAP), GMM-SVM y el sistema que resulta de la fusión de ambos sistemas. La evaluación se ha llevado a cabo sobre la tarea principal de la NIST SRE06. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

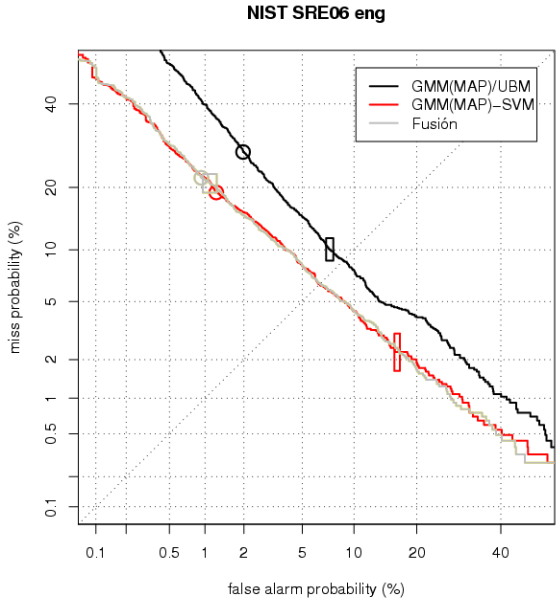

Figura 4.28. *Curvas DET de los sistemas GMM(MAP), GMM-SVM y el sistema que resulta de la fusión de ambos sistemas (fusión optimizada sobre la NIST SRE de 2005). La evaluación se ha llevado a cabo sobre el subconjunto de experimentos de la tarea principal de la NIST SRE06 constituido por segmentos en inglés. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

Las tasas EER y los costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de estos sistemas se presentan en la **Tabla 4.15**. Los costes C_{llr} y C_{llr}^{min}, y C_{DET} y C_{DET}^{min} obtienen, respectivamente, valores muy similares, sobre todo en comparación con la distancia que presentan en los sistemas originales. En los experimentos *all*, el sistema fusionado presenta una mejora del 1.32% en EER, con respecto al mejor subsistema (GMM-SVM). En los experimentos *eng*, el EER se mantiene en el mismo punto.

Tabla 4.15. *Tasa EER y costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de los sistemas GMM(MAP), GMM-SVM y el sistema que resulta de la fusión de ambos sistemas (entrenada sobre el corpus de la NIST SRE de 2005). La evaluación se ha llevado a cabo sobre la tarea principal de la NIST SRE06 ("all") y sobre un subconjunto de experimentos en ingles de esta tarea ("eng").*

NIST SRE06		Cllr	Cllr.min	EER	Cdet	Cdet.min
all	GMM(MAP)	0.9537	0.3619	10.32	0.0864	0.05186
	GMM-SVM	0.9965	0.2946	8.33	0.1867	0.03968
	Fusión	0.321	0.2893	8.22	0.0399	0.0392
eng	GMM(MAP)	0.9496	0.3151	8.78	0.0826	0.0472
	GMM-SVM	0.9962	0.2335	6.39	0.1643	0.0312
	Fusión	0.275	0.2319	6.39	0.0316	0.0312

Para analizar cómo se complementan los sistemas con probabilidades normalizadas, se ha realizado la fusión de los mismos. En los experimentos realizados, el mejor rendimiento se obtiene con la normalización Z-norm para el sistema GMM(MAP) (tabla 4.12), y con la T-norm para el sistema GMM-SVM (tabla 4.13). Debido a que primero se han de estimar los parámetros de la fusión, se han estimado los parámetros de normalización Z-norm y T-norm sobre la NIST SRE05. Para ello, se emplean los mismos conjuntos de impostores que en los experimentos de normalización de la NIST SRE06. En la **Tabla 4.16** se muestran los resultados obtenidos con estas metodologías en los experimentos *all* de la NIST SRE05. Se presentan los costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min}, y la tasa EER de los sistemas GMM(MAP) y GMM-SVM, para los sistemas originales (sin normalización) y para los sistemas con Z-norm y T-norm. El comportamiento de estos sistemas es similar al observado para el conjunto experimental de la NIST SRE06 (véanse las tablas 4.12 y 4.13):

- La Z-norm presenta mejor rendimiento que la T-norm con el sistema GMM(MAP). Aunque la tasa EER obtenida al aplicar la Z-norm es ligeramente superior a la del sistema original. Los costes C_{llr}, C_{llr}^{min}, y C_{DET}^{min} son inferiores al aplicar la Z-norm con respecto a los costes del sistema original.
- Con el sistema GMM-SVM, la T-norm obtiene mejor rendimiento que la Z-norm y que el sistema original. Su tasa EER presenta una mejora del 7.78% con respecto al sistema original.

Estos resultados vuelven a sugerir el empleo de la Z-norm con GMM(MAP) y la T-norm sobre GMM-SVM. El resultado obtenido con su fusión sobre el mismo corpus de la NIST SRE05, se muestra también en la tabla 4.16. Al tratarse de una evaluación realizada en cerrado, cabe esperar un buen comportamiento del sistema fusionado (y así sucede, en efecto). El EER decrece en un 2.35% con respecto al mejor subsistema (GMM-SVM con T-norm) y presenta el mejor rendimiento de todos los sistemas analizados. Asimismo, las puntuaciones del sistema fusionado demuestran estar debidamente calibradas. Por último, comparando estos resultados con los de la fusión de los sistemas originales, sin normalización, (véase la tabla 4.14), se observa una mejora del 7.29% en el EER (9.15% frente a 9.87%), y todos los costes son inferiores (compárense las tablas 4.14 y 4.16).

Tabla 4.16. *Tasa EER y funciones de coste C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de los sistemas GMM(MAP), GMM-SVM. Se analiza el rendimiento de los sistemas sin aplicar ninguna normalización de probabilidades, y al aplicar Z-norm y T-norm. La evaluación se ha realizado sobre la tarea principal de la NIST SRE05. Asimismo, se incluye el rendimiento obtenido al fusionar, en cerrado los sistemas que han mostrado el mejor rendimiento: GMM(MAP)+Z-norm y GMM-SVM+Tnorm.*

NIST SRE05		Cllr	Cllr.min	EER	Cdet	Cdet.min
GMM(MAP) /UBM	Sin normalización	0.9533	0.4153	11.64	0.08399	0.05434
	Z-norm	0.6975	0.4051	12.04	0.08399	0.05262
	T-norm	0.6822	0.4213	12.54	0.08399	0.04485
GMM(MAP) -- SVM	Sin normalización	0.9962	0.3571	10.16	0.15079	0.03923
	Z-norm	0.6707	0.3669	10.92	0.15079	0.0396
	T-norm	0.6666	0.3311	9.37	0.15079	0.0358
Fusión (cerrado)	GMM(MAP)/UBM+Z-norm GMM(MAP)-SVM+T-norm	0.3329	0.326	9.15	0.0356	0.0351

En la **Tabla 4.17** se presenta el rendimiento de la fusión de los sistemas GMM(MAP) con Z-norm y GMM-SVM con T-norm para los experimentos *all* y los experimentos *eng* de la NIST SRE06.

Tabla 4.17. *Tasa EER y costes C_{llr}, C_{llr}^{min}, C_{DET} y C_{DET}^{min} de los sistemas GMM(MAP)+Znorm, GMM-SVM+T-norm y la fusión de ambos sistemas (optimizada sobre la NIST SRE de 2005). La evaluación se ha realizado sobre la tarea principal de la NIST SRE06 ("all") y el subconjunto de experimentos de esta tarea compuesto únicamente por segmentos en inglés ("eng").*

NIST SRE06		Cllr	Cllr.min	EER	Cdet	Cdet.min
alll	GMM(MAP)/UBM+Z-norm	0.6714	0.357	10.45	0.0864	0.05009
	GMM(MAP)-SVM+T-norm	0.6744	0.2767	7.85	0.1867	0.03445
	Fusión	0.2918	0.2725	7.77	0.03459	0.0341
eng	GMM(MAP)/UBM+Z-norm	0.6559	0.3003	8.34	0.0826	0.0446
	GMM(MAP)-SVM+T-norm	0.6437	0.2279	6.39	0.1643	0.02841
	Fusión	0.2506	0.2222	6.23	0.02899	0.02871

De los resultados obtenidos se pueden extraer las mismas conclusiones que en experimentos de fusión previos: la fusión mejora el rendimiento de los subsistemas fusionados (tanto en cuanto al EER como a los costes) y calibra debidamente las puntuaciones.

Por último, en las **Figuras 4.29 y 4.30** se compara el rendimiento de la fusión de los sistemas originales y la fusión de los sistemas con probabilidades normalizadas. La comparación se realiza para los experimentos *all* (Figura 4.29) y los experimentos *eng* (Figura 4.30) de la NIST SRE06.

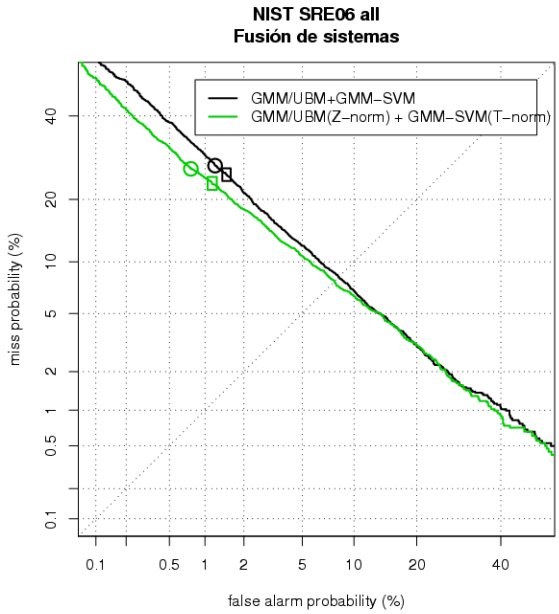

Figura 4.29. *Curvas DET de la fusión de: (1) los sistemas GMM(MAP) y GMM-SVM; y (2) los sistemas GMM(MAP) con Z-norm y GMM-SVM con T-norm. La evaluación se ha llevado a cabo sobre la tarea principal de la NIST SRE06. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

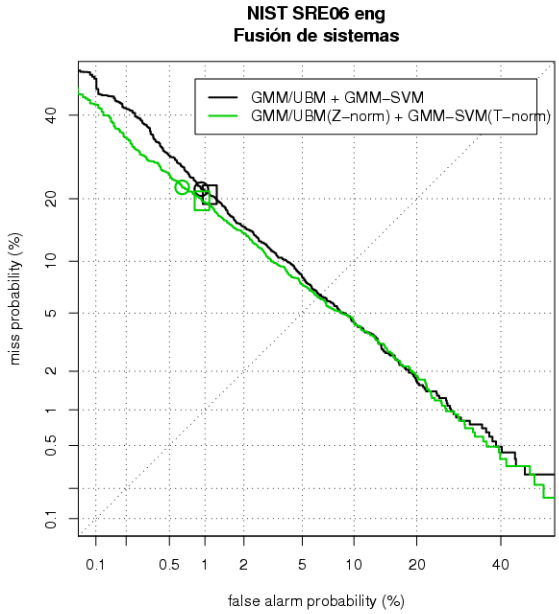

Figura 4.30. *Curvas DET de la fusión de: (1) los sistemas GMM(MAP) y GMM-SVM; y (2) los sistemas GMM(MAP) con Z-norm y GMM-SVM con T-norm. La evaluación se ha llevado a cabo el subconjunto de experimentos en inglés de la tarea principal de la NIST SRE06. El rectángulo marcado en las curvas indica el C_{DET} del sistema, el coste real. El círculo indica el punto de coste mínimo C_{DET}^{min}.*

En ambos casos, se observa una clara superioridad de la fusión con probabilidades normalizadas, dado que se obtiene una curva DET mejor prácticamente para todos los puntos de operación. Esta tendencia se ve reflejada en la tasa EER y en todos los costes. En los experimentos *all*, el EER de la fusión con probabilidades normalizadas es del 7.77% (Tabla 4.17), lo que supone una mejora del 5.47% con

respecto al EER de la fusión de sistemas sin normalización, que es del 8.22% (Tabla 4.15). Los costes presentan el mismo comportamiento, con mejores resultados, en todos los casos, para la fusión con puntuaciones normalizadas. En los experimentos *eng*, se observa un comportamiento idéntico: el EER de la fusión de sistemas con probabilidades normalizadas (un 6.23%) presenta una mejora del 2.5% con respecto al EER de la fusión de sistemas sin aplicar ninguna metodología de normalización (un 6.39%). De nuevo, los costes son inferiores para la fusión con probabilidades normalizadas.

4.2.5 Conclusiones y trabajo futuro

En esta segunda parte del capítulo se describe la experimentación llevada a cabo sobre las bases de datos del NIST. La evaluación se realiza sobre dos conjuntos experimentales de la NIST SRE06: el conjunto experimental de la tarea principal (denominados experimentos *all* de forma abreviada) y un subconjunto de experimentos de esta tarea, compuesto únicamente por señales en inglés (denominados experimentos *eng*). Se han desarrollado dos sistemas: (1) un sistema basado modelos GMM(MAP); y (2) otro basado en GMM-SVM. En los resultados obtenidos se observa una clara superioridad del sistema GMM-SVM (véanse la Figura 4.17 y la tabla 4.11). Por otro lado, los sistemas muestran una gran dependencia del idioma de las pronunciaciones, ya que se observa una mejora considerable al evaluar los sistemas sobre el conjunto *eng*.

La experimentación incluye también la aplicación de metodologías de normalización sobre las puntuaciones obtenidas con ambos sistemas, tales como Z-norm, T-norm, ZT-norm y TZ-norm. Se observa que la Z-norm y la TZ-norm presentan el mejor rendimiento sobre GMM(MAP) (véase la tabla 4.12). Para el sistema GMM-SVM, en cambio, el mejor rendimiento se obtiene con la T-norm (véase la tabla 4.13). A priori, era de esperar un mejor rendimiento de la ZT-norm o la TZ-norm frente a la Z-norm y la T-norm, dado que (supuestamente) la ZT-norm y la TZ-norm minimizan conjuntamente la variabilidad de la señal y del locutor, mientras que la Z-norm trata de minimizar sólo la variabilidad del locutor, y la T-norm sólo la de la señal de evaluación. Queda pendiente analizar con más detalle estos resultados en experimentos futuros. De todas formas, este comportamiento podría ser debido a la composición de los conjuntos de impostores T-norm y Z-norm. Tal como se ha mencionado anteriormente, el rendimiento de la T-norm y la Z-norm mejora si el conjunto de impostores se selecciona de forma muy cuidadosa (véase [McLaren09]), si se definen conjuntos de impostores extensos (lo que incrementa el coste del proceso de normalización), o si se definen conjuntos de impostores T-norm dependientes del locutor (véanse las aproximaciones AT-norm o KL-Tnorm descritas en el apartado 2.4.3). Asimismo, se ha de tener en cuenta que el conjunto de impostores Z-norm definido para este trabajo es mixto y único (contiene tanto hombres como mujeres). Puesto que en las NIST SRE no se prevé ningún experimento cruzado, donde el género del locutor real sea diferente al género del locutor objetivo, se cree que, probablemente, la Z-norm obtendría un rendimiento mejor si se definieran conjuntos de impostores Z-norm dependientes del género, un conjunto para los locutores objetivo hombres, y otro para las mujeres. Generalmente,

resulta más difícil discriminar la voz de una mujer entre mujeres, que la voz de un hombre entre mujeres (y viceversa), razón por la cual no se planifican experimentos cruzados en las NIST SRE. Pero en una situación real, no se puede suponer que se conocerá el género del locutor objetivo, por lo que se ha definido un único conjunto de impostores Z-norm.

El apartado finaliza analizando la fusión de sistemas, tanto la fusión de los sistemas originales, GMM(MAP) y GMM-SVM, como la fusión de los sistemas con puntuaciones normalizadas, aplicando Z-norm sobre GMM(MAP) y T-norm sobre GMM-SVM. Los resultados revelan una ligera mejoría de la curva DET y de la tasa EER al fusionar los sistemas. Asimismo, como cabe esperar, se observa un rendimiento ligeramente mejor de la fusión con probabilidades normalizadas, frente a la fusión de los sistemas originales (véanse las Figuras 4.29 y 4.30). Pero la conclusión más destacable de estos experimentos de fusión, es, quizás, la buena calibración que presentan las puntuaciones fusionadas (véase en las tablas 4.15 y 4.17 la pequeña diferencia que se observa, respectivamente, entre los costes C_{llr} y C_{llr}^{min}, y C_{DET} y C_{DET}^{min}). Por otro lado, sería interesante analizar la fusión con otro tipo de sistemas que no empleen los MFCC como representación acústica, sino otro tipo de parámetros, especialmente parámetros de nivel alto, tales como fonemas, patrones de pronunciación, etc. Cabe esperar que la correlación de sistemas con distintos tipos de parámetros será menor, y en consecuencia, se complementarán de forma más adecuada en la fusión. Asimismo, queda pendiente probar la fusión con información auxiliar adicional, la cual se podría probar primero sobre los sistemas ya definidos, para analizar cómo varía el rendimiento. Para ello, se podrían definir subconjuntos de experimentos según el idioma de los segmentos de entrenamiento y test, el género del locutor objetivo, el canal de grabación de los segmentos, etc.

Por otro lado, debido a que en el corpus de las NIST SRE se incluyen señales grabadas a través de diferentes canales telefónicos y a través de diversos terminales, y tal como se demuestra en la mayoría de sistemas desarrollados para las últimas evaluaciones, sería beneficioso aplicar alguna metodología de compensación de canales. De las metodologías propuestas con este fin en los últimos años, destacan JFA (*Joint Factor Analysis*) [Kenny08], NAP (*Nuissance Attribute Projection*) [WMCampbell06c] y la adaptación por auto-canales (*Eigenchannel Adaptation*) [Kenny04]. JFA y los auto-canales se aplican sobre sistemas basados en GMM; NAP, en cambio, suele aplicarse sobre SVM. Asimismo, sería interesante analizar el rendimiento que presenta HLDA, descrita y evaluada sobre Albayzín y Dihana en el capítulo 3, sobre los conjuntos experimentales de las NIST SRE. HLDA es una transformación lineal que se aplica sobre el espacio parametrizado de las muestras, para normalizar los efectos del canal, decorrelar el espacio, aumentar su capacidad de discriminación y reducir, en la medida de lo posible, la dimensionalidad de los vectores acústicos.

Por último, llegados a este punto, cabe destacar que sería interesante (y queda pendiente como trabajo futuro) analizar el rendimiento de SSM sobre los conjuntos experimentales de la NIST SRE06. SSM se podría aplicar, combinado linealmente con el UBM, o como sistema adicional en la fusión para mejorar el rendimiento del sistema.

Capítulo 5

Seguimiento de locutores en un entorno AmI

Debido al potencial de aplicación que presenta el reconocimiento de locutores como tecnología de seguimiento de usuarios, sobre todo por la naturalidad y transparencia de la interacción con el sistema, en la primera parte de este capítulo se propone y evalúa un sistema de seguimiento de locutores pensado para su aplicación en hogares inteligentes.

A continuación, el capítulo aborda la implantación del sistema propuesto, con un breve resumen de los proyectos de investigación desarrollados en los últimos años en relación con hogares inteligentes, y la descripción de las principales tecnologías empleadas en las aplicaciones de seguimiento de usuarios. Finaliza con la descripción de una aplicación que implementa el sistema de seguimiento propuesto en este trabajo.

5.1 Sistema continuo de seguimiento de locutores

El sistema de seguimiento de locutores que se propone en esta tesis está orientado a su aplicación en hogares inteligentes. En un escenario de este tipo, la cantidad de locutores que debe monitorizar (seguir) el sistema no suele ser muy elevada. Por tanto, la aproximación implementada parte de la precondición de que los usuarios a seguir son conocidos: habitantes o individuos asiduos. El seguimiento se realiza sobre un conjunto predeterminado de locutores, que se detectan simultáneamente. Por otro lado, el escenario planteado requiere un sistema de seguimiento continuo (*online*), que proporcione decisiones inmediatas (de baja latencia). La mayoría de las aproximaciones de seguimiento de locutores son cíclicas o multi-fase, es decir, los datos de entrada se procesan de

principio a fin de forma iterativa. No se pueden aplicar en escenarios que requieran una respuesta en tiempo real. De hecho, se han encontrado muy pocas aproximaciones de seguimiento y diarización de locutores orientadas a su ejecución online o en tiempo real (véase el estado del arte en el apartado 2.5).

Con objeto de cumplir los requisitos de continuidad y baja latencia, el sistema propuesto sigue un algoritmo muy simple [Zamalloa10a]. La segmentación y la clasificación de la señal de entrada se realizan de forma conjunta: se definen y procesan segmentos de longitud fija (específicamente de 1 segundo), a los cuales se les asigna una tasa de detección para cada locutor a reconocer. A continuación, se aplica un proceso de calibración, que mapea linealmente las tasas de detección a probabilidades "teóricas". Las probabilidades calibradas se comparan con un umbral establecido de forma heurística sobre un corpus de desarrollo para decidir si el segmento de entrada corresponde a alguno de los locutores conocidos o a un locutor desconocido. Por otro lado, debido a que las probabilidades se calculan a partir de segmentos de duración muy reducida, el rendimiento del sistema de detección puede ser muy sensible a la variabilidad local del segmento analizado. En consecuencia, con el objetivo de aumentar la robustez del sistema, se propone una metodología de suavizado que utiliza información extraída de segmentos anteriores [Zamalloa10b].

5.1.1 Descripción de los módulos del sistema

A continuación, se describe la funcionalidad de los módulos que componen el sistema

Parametrización acústica

El módulo de parametrización acústica se encarga de extraer las secuencias de vectores de parámetros acústicos a partir de muestras adquiridas a 16 kHz empleando un micrófono del entorno. Las muestras se analizan en tramos de 20 milisegundos, solapados a intervalos de 10 milisegundos. Después, se aplica una ventana de *Hamming* y se calcula una FFT de 512 puntos. Las amplitudes de la FFT se promedian en 24 filtros triangulares solapados, con frecuencias centrales y anchuras según la escala de Mel de percepción auditiva. A continuación, se aplica una DCT sobre el logaritmo de las amplitudes del filtro y se obtienen 12 MFCC. Debido a que la señal de audio se procesa de forma continua, se aplica una aproximación dinámica de la CMS [Huang01], donde la media cepstral se adapta de forma dinámica para cada tramo *i*, tal como se muestra a continuación:

$$\mu_i = \alpha C(i) + (1-\alpha)\mu_{i-1} \tag{5.1}$$

donde α es una constante de tiempo (cuyo valor está en torno a 0.001), $C(i)$ el vector de coeficientes cepstrales del tramo *i*, y μ_{i-1} la media cepstral dinámica del tramo *i-1*. Por último, se estiman las primeras derivadas de los MFCC, obteniendo vectores acústicos de 24 dimensiones.

Cabe destacar que el proceso de parametrización se ha llevado a cabo mediante el framework *Sautrela* desarrollado por Mikel Peñagarikano [Penagarikano08]. Sautrela es un entorno de desarrollo de software versátil para tecnologías del habla, escrito en Java y ofrecido bajo una licencia de distribución libre [Sautrela].

Detección de locutores

El módulo de detección de locutores da como salida, a intervalos de un segundo, una tasa de detección para cada locutor objetivo. Esta longitud se ha determinado de forma empírica, ya que proporciona una resolución temporal y una riqueza espectral relativamente adecuadas, y presenta una latencia aceptable para la mayoría de aplicaciones.

La tasa de detección se calcula mediante GMM(MAP) estimados a partir de un UBM. Estos modelos, además de presentar muy buen rendimiento para el reconocimiento de locutores, permiten la aplicación de la estrategia denominada *top-N*, que aligera el coste del proceso de cómputo de probabilidades y obtiene estimaciones más robustas: la probabilidad de cada observación se evalúa únicamente a partir de las *N* componentes del modelo de locutor que obtienen mayor probabilidad en el UBM.Para la adaptación de los modelos se emplean segmentos que de acuerdo a las anotaciones manuales proporcionadas por los creadores de la base de datos, sólo contienen habla del locutor a entrenar. De esta forma, las componentes gaussianas relacionadas con las fuentes acústicas más observadas en estos segmentos se desplazarán de forma notable, mientras que las componentes relacionadas con fuentes no presentes (música, silencio, otros locutores) se mantienen prácticamente intactas, sin adaptarse. En consecuencia, se supone que el sistema de seguimiento planteado tendrá la capacidad de detectar los segmentos correspondientes a los locutores objetivo y rechaza al resto (segmentos correspondientes a impostores, silencio, música, etc.).

Dado el modelo de locutor λ_l del locutor *l*, y el modelo UBM λ_{UBM}, la tasa de detección $\Delta_t(X)$ del segmento de entrada *X* se estima de la siguiente manera:

$$\Delta_l(X)=L(X|\lambda_l)-L(X|\lambda_{UBM}) \tag{5.2}$$

donde $L(X|\lambda)$ es la log-probabilidad condicional del segmento X dado λ. Una vez el módulo de detección ha estimado las tasas de detección del segmento de entrada *X*, se procede a su calibración.

Calibración del sistema

El módulo de calibración aplica una transformación lineal para convertir las tasas de detección Δ_l, con $l \in [1, L]$, en probabilidades teóricas $C(\Delta_l)$, con $l \in [1, L]$, ajustadas a un amplio rango de posibles escenarios de aplicación. Se aplica con el objetivo de optimizar la determinación de un umbral de aceptación. Los parámetros de la transformación se estiman sobre un corpus de desarrollo, tratando de maximizar la información mutua de las tasas obtenidas en experimentos definidos sobre este corpus, o lo que viene a ser lo mismo, tratando de minimizar la métrica C_{llr}. Para tomar la decisión final sobre un segmento *X*, la tasa calibrada se compara con el umbral de decisión de Bayes (de coste mínimo):

$$\text{decisión}=\begin{cases}\text{locutor objetivo} & \text{si } C(\Delta(X))>\ln\left(\frac{C_{fa}(1-P_{obj})}{C_{miss}P_{obj}}\right)\\ \text{locutor desconocido} & \text{en caso contrario}\end{cases} \tag{5.3}$$

La transformación lineal aplicada corresponde a la función denominada “Z_cal” en el paquete de software FoCal ofrecido bajo licencia libre (véase [FoCal]), una transformación lineal en la que se

acotan los valores mínimos y máximos de las tasas transformadas en función de los valores mínimos y máximos que adquieren respectivamente las tasas de detección de los locutores objetivo e impostores.

Seguimiento de locutores

El módulo de seguimiento de locutores se encarga de decidir si el segmento de entrada corresponde a alguno de los locutores objetivo – y en caso de que así sea, a cuál de ellos corresponde –, o si corresponde a algún locutor desconocido. Esta decisión se toma a partir de las tasas calibradas $C(\Delta_l)$, con $l \in [1, L]$, aplicando bien un criterio de identificación (IDL), o bien un criterio de verificación (VERL).

Al emplear la aproximación IDL, la comparación con el umbral se realiza únicamente para el locutor objetivo de máxima probabilidad $C(\Delta_{l^*}(X))$, donde:

$$l^*(X) = \arg\max_{l \in [1.L]} \{C(\Delta_l(X))\} \tag{5.4}$$

Así, se toma una única decisión para cada segmento de entrada *X*: *X* ha sido pronunciado por el locutor objetivo l^* si $C(\Delta_{l^*}(X)) > \theta_I$; en caso contrario, *X* ha sido pronunciado por un locutor desconocido. Obviamente, esta aproximación no puede detectar segmentos solapados, en los que dos o más locutores intervienen a la vez, ya que la decisión se toma únicamente en base a la probabilidad del locutor más probable.

Al emplear la aproximación VERL, en cambio, la comparación con el umbral se realiza para las tasas de todos los locutores objetivo. El sistema decide que el locutor l habla en el segmento *X* si $C(\Delta_l(X)) > \theta_V$, donde θ_V es el umbral de aceptación. En caso contrario, decide que *l* no habla en *X*. Por lo tanto, esta aproximación permite la detección de segmentos en los que hablan dos o más locutores, dado que el sistema considera que todos los locutores cuya probabilidad cumple la condición $C(\Delta_l(X)) > \theta_V$ aparecen en *X*. De todas formas, si el sistema, por diseño, no permite detectar segmentos solapados, ambas aproximaciones proceden del mismo modo.

Suavizado para mejorar la robustez del sistema

Las tasas de detección se estiman a partir de segmentos de duración corta (1 segundo), por lo que el rendimiento del sistema puede resultar perjudicado debido a la variabilidad local del segmento. Con objeto de mejorar la robustez del sistema frente a este tipo de variabilidades, el sistema incluye un módulo de suavizado, que emplea información obtenida de segmentos anteriores. Así, las tasas de detección se estiman a partir de segmentos de mayor longitud. Suponiendo que en los segmentos previos no se ha dado un cambio de locutor (o un cambio de la fuente acústica), emplear las muestras correspondientes a dichos segmentos para estimar las tasas de detección produce tasas más precisas y robustas, y en consecuencia, el rendimiento del sistema aumenta. Este método de suavizado no impide cumplir con los requisitos de procesamiento continuo y latencia baja, dado que toda la información adicional se extrae de segmentos ya procesados.

En la práctica, este módulo de suavizado consiste en combinar linealmente las tasas de detección obtenidas para los últimos *w* segmentos de un segundo de duración. Estos *w* segmentos (entre los que

se encuentra el último segmento analizado) se ponderan mediante una función rectangular (pesos uniformes) o triangular (pesos que disminuyen de forma lineal hacia atrás en el tiempo). La función triangular da mayor importancia a los segmentos más recientes de la ventana de análisis, mientras que la función rectangular da la misma importancia a todos los segmentos.

5.2 Experimentos del sistema de seguimiento continuo

En este apartado se describe la fase de evaluación del sistema propuesto para el seguimiento continuo (y de baja latencia): la configuración experimental y los resultados obtenidos.

5.2.1 Configuración experimental

5.2.1.1 La base de datos AMI

Los experimentos se han llevado a cabo sobre la base de datos de reuniones AMI, de acceso público bajo una licencia libre. Se trata de un conjunto de datos multimodal con interacciones humanas en tiempo real, en el contexto de reuniones grabadas en un entorno inteligente. Las grabaciones se han realizado en tres salas equipadas con diversos micrófonos y cámaras, que recogen señales sincronizadas de audio y video. Se ha elegido esta base de datos por su estrecha relación con el escenario AmI (hogar inteligente) para el que se ha diseñado el sistema de seguimiento de locutores continuo. Cabe destacar que no se conoce ninguna otra aproximación de seguimiento que emplee este conjunto de datos para la experimentación.

El desarrollo y la evaluación del sistema se han realizado a partir de un subconjunto de datos del corpus. Concretamente, a partir de las grabaciones realizadas en la sala de reuniones de Edimburgo, que incluye 15 sesiones en total; específicamente, las sesiones ES2002-ES2016, con 4 reuniones de media hora de duración, en promedio, por cada sesión. Cada sesión consta de 4 reuniones, en las que toman siempre parte los mismos 4 participantes. Tres de estos cuatro participantes se escogen como locutores objetivo del sistema, mientras que al otro participante se le asigna el papel de impostor. La selección de locutores objetivo/impostor se ha realizado cuidadosamente (y no de forma aleatoria), para evitar que el género del participante seleccionado como impostor facilite su discriminación. Es decir, para las sesiones en las que tres de los cuatro participantes son hombres, se establece que el impostor ha de ser uno de los hombres y no la mujer (y viceversa). Las señales de audio se obtienen sumando las señales grabadas por los micrófonos de boca de cada participante.

Con objeto de hacer una evaluación del sistema en condiciones de independencia estadística se han definido dos subconjuntos disjuntos que se componen de sesiones diferentes (y de hecho, de locutores diferentes): uno para el desarrollo del sistema; y otro, para su evaluación. El conjunto de desarrollo consta de 8 sesiones (32 reuniones) y se emplea para ajustar los parámetros de configuración del sistema. El conjunto de evaluación contiene las 7 sesiones restantes (28 reuniones), y se emplea para evaluar el rendimiento del sistema con los parámetros ajustados sobre el corpus de desarrollo.

Tanto el conjunto de desarrollo como el de evaluación se dividen, a su vez, en dos subconjuntos: el de entrenamiento, con dos reuniones de cada sesión seleccionadas aleatoriamente; y el conjunto de test, compuesto por las dos reuniones restantes. En la **Figura 5.1** se muestra una representación esquemática de las particiones creadas.

Grabaciones sala Edinburgo (*4 reuniones por sesión*)			
Desarrollo *8 sesiones*		**Evaluación** *7 sesiones*	
Entrenamiento *2 reuniones por sesión* (selección aleatoria)	**Test** *2 reuniones restantes* (de cada sesión)	**Entrenamiento** *2 reuniones por sesión* (selección aleatoria)	**Test** *2 reuniones restantes* (de cada sesión)
➢ Estimación UBM ➢ Estimación modelos de locutor	➢ Configuración del sistema	➢ Estimación UBM ➢ Estimación modelos de locutor	➢ Evaluación del sistema

Figura 5.1 *Representación esquemática de las particiones creadas en la base de datos AMI para la evaluación del sistema de seguimiento continuo*

5.2.1.2 Estimación del UBM

Se han definido dos sistemas de detección de locutores que difieren en los datos empleados para la estimación del UBM:

- El sistema *UBM-g* emplea 15 reuniones AMI grabadas en dos salas distintas a la de Edimburgo. Esto significa que el UBM se estima a partir de un conjunto de datos independiente de los conjuntos de desarrollo y evaluación, por lo que cabe esperar un cierto desajuste acústico entre el UBM y los datos de entrenamiento y test. Las reuniones se han seleccionado cuidadosamente con objeto de, en la medida de lo posible, equilibrar en género el conjunto de datos empleado.
- El sistema *UBM-t*, por el contrario, emplea únicamente las reuniones del conjunto de entrenamiento. Por tanto, en este caso, el UBM es dependiente del conjunto de locutores objetivo, ya que se estima a partir sus propios datos de entrenamiento.

El modelo UBM-g presenta la ventaja de que una vez calculado, se aplica a cualquier conjunto de evaluación, mientras que el modelo UBM-t es dependiente del conjunto de evaluación, y en consecuencia se ha de reestimar de forma específica para cada conjunto de locutores objetivo. Ambos modelos, UBM-g y UBM-t, son mixtos, independientes del género del locutor y constan de 256 gaussianas, estimadas en 20 iteraciones del algoritmo EM.

5.2.1.3 Comparación del sistema de seguimiento continuo con un sistema de referencia

El rendimiento del sistema continuo se ha comparado con el de un sistema de referencia que opera de forma *offline* (es decir, de forma no continua), el cual, al igual que muchas de las aproximaciones de seguimiento de la bibliografía, consta de dos fases secuenciales desacopladas: (1) la segmentación acústica del recurso de audio a procesar; y (2) la detección de los segmentos definidos en la fase anterior. Los puntos de cambio se detectan mediante una metodología clásica de segmentación

acústica basada en métricas estadísticas (la metodología se describe más abajo), similar a la conocida aproximación BIC [Chen98]. A continuación, se determina el locutor más probable de cada segmento, a partir de los modelos de locutor GMM(MAP) estimados en la fase de entrenamiento.

La comparación trata de analizar el grado de degradación del sistema continuo, que segmenta y detecta los locutores de forma conjunta, con respecto a un sistema que aplica ambas fases de forma desacoplada. Cabe destacar que ambas aproximaciones (el sistema continuo y el de referencia) emplean los mismos modelos de locutor.

Metodología de segmentación acústica del sistema de referencia

El sistema de referencia sigue un algoritmo no-supervisado muy simple para realizar la segmentación acústica. Los puntos de cambio se determinan desde un punto de vista puramente acústico. El algoritmo considera una ventana deslizante *W* compuesta por *N* vectores acústicos, y estima la probabilidad de que en la mitad de la ventana haya un punto de cambio. A continuación, desplaza la ventana en *K* vectores hacia adelante, y repite este proceso hasta alcanzar el final de la secuencia de vectores, es decir, el final del recurso de audio a procesar. Para estimar la probabilidad de un punto de cambio, cada ventana se divide en dos mitades W_{izq} y W_{dcha}, se estima una distribución Gaussiana de covarianza diagonal con cada mitad, y se calcula su CLR. El CLR se registra como probabilidad de cambio en el punto situado en la mitad de *W*. De esta forma, se obtiene una secuencia de CLRs a post-procesar con objeto de determinar las fronteras hipotéticas de los segmentos. El procedimiento depende de dos parámetros de configuración: (1) el umbral de aceptación τ; y (2) la longitud mínima de segmento δ. En la práctica, se decide que un punto *t* es realmente un punto de cambio acústico si su CLR es superior a τ y además no se detecta ningún punto con mayor CLR en el intervalo $[t-\delta, t+\delta]$ (para una descripción más detallada de la metodología de segmentación, véase [Rodriguez07]).

5.2.2 Resultados de los experimentos de seguimiento

5.2.2.1 Sistema de seguimiento continuo vs sistema de referencia

En las **Figuras 5.2 y 5.3** se comparan las curvas DET del sistema continuo, de latencia baja, (denominado sistema RT, en adelante) con un sistema de referencia (sitema REF, en adelante), sobre los conjuntos de desarrollo (Figura 5.2) y evaluación (Figura 5.3). Se contrastan las dos aproximaciones de modelado UBM-g y UBM-t. Para tomar las decisiones se sigue la aproximación IDL, por lo que del conjunto de test se excluyen los segmentos en los que interviene más de un locutor.

Como cabe esperar, los resultados obtenidos sobre ambos conjuntos, muestran un rendimiento superior del sistema REF con respecto al sistema RT. De todas formas, se observa un buen rendimiento del sistema RT, el cual toma las decisiones al instante y únicamente a partir del segmento correspondiente al último segundo. Estableciendo como base el rendimiento del sistema REF, sobre el conjunto de desarrollo, el EER pasa de 20.11 a 21.47 al emplear UBM-g (un deterioro relativo del 6.76%), y de 16.93 a 19.94 al emplear UBM-t (un deterioro relativo del 17.77%). Sobre el conjunto de

evaluación, se observa la misma tendencia, donde tomando de nuevo el sistema REF como base, el EER pasa de 25.81 a 26.20 con UBM-g (un deterioro relativo del 1.55%), y de 19.07 a 21.14 con UBM-t (un deterioro relativo del 10.85%).

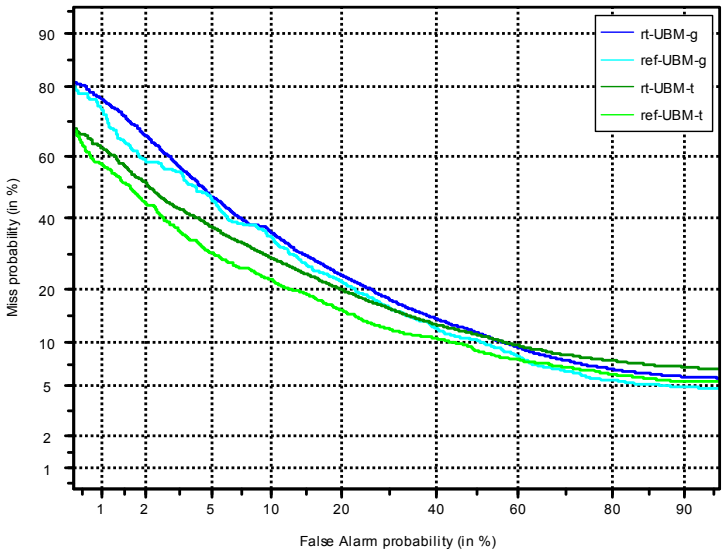

Figura 5.2. *Curvas DET del sistema continuo(rt) y el sistema de referencia (ref) para el seguimiento de locutores con las aproximaciones UBM-g y UBM-t sobre el conjunto de desarrollo del corpus AMI. El módulo de detección sigue la metodología empleada en tareas de identificación de locutores.*

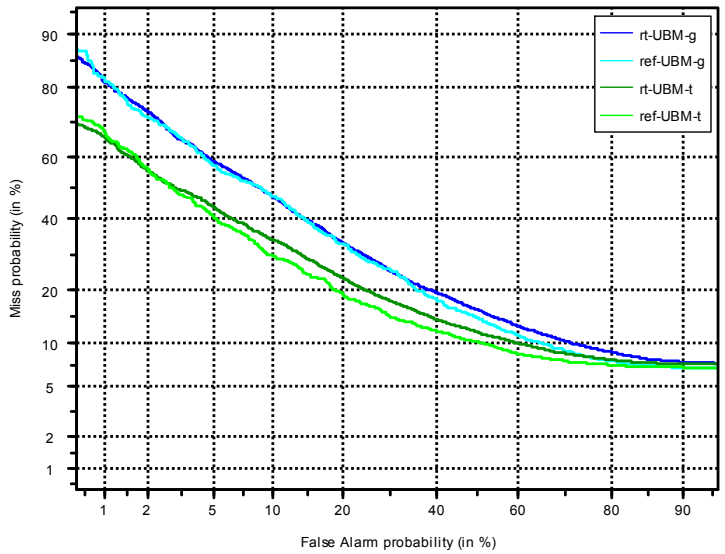

Figura 5.3. *Curvas DET del sistema continuo (rt) y el sistema de referencia (ref) para el seguimiento de locutores al emplear las aproximaciones UBM-g y UBM-t sobre el conjunto de evaluación de AMI. El módulo de detección sigue la metodología empleada en tareas de identificación de locutores.*

La aproximación UBM-t presenta un rendimiento similar sobre ambos conjuntos (el EER pasa del 19.94 sobre el conjunto de desarrollo al 21.14 sobre el conjunto de evaluación), lo cual demuestra que la configuración del sistema (determinada sobre el conjunto de desarrollo) es también adecuada para el conjunto de evaluación. Al emplear la aproximación UBM-g, por el contrario, el sistema presenta una

mayor degradación: el EER pasa del 21.47 en desarrollo al 26.20 en evaluación). Todo parece indicar que estimar el UBM a partir de señales grabadas en condiciones acústicas dispares y con otros locutores hace que la cobertura acústica del UBM empeore y que además la configuración del sistema presente una menor robustez que la del sistema que emplea un UBM estimado en la misma sala y a partir de los propios locutores objetivo.

En la **Tabla 5.1** se muestran la precisión, la cobertura y la medida-F de los cuatro sistemas desarrollados (UBM-g sistema continuo, UBM-g sistema de referencia, UBM-t sistema continuo, UBM-t sistema de referencia) con probabilidades calibradas y no-calibradas. Estos resultados corresponden al punto de operación de la curva DET considerado óptimo, es decir al umbral de aceptación óptimo. Para las probabilidades calibradas el umbral se establece sobre el corpus de desarrollo, en función de las probabilidades a priori y los costes de la aplicación final. Para las tasas de detección no-calibradas, el umbral se fija a 0.

Los resultados presentados en la tabla 5.1 evidencian la validez del proceso de calibración, que mejora de forma considerable el rendimiento (medida-F) de todos los sistemas. Los resultados ponen de manifiesto que el sistema de referencia extrae información relevante de la segmentación acústica, pero en función del escenario y la latencia requerida, dicha segmentación puede no ser aplicable. En esas situaciones, el sistema de seguimiento continuo propuesto en este trabajo presenta características más adecuadas, ya que toma las decisiones de forma continua y con una latencia baja, a costa de una ligera degradación del rendimiento del sistema.

Tabla 5.1. *Precisión (PRC), cobertura (RCL) y medida-F del sistema continuo (rt) y el sistema de referencia (ref) para el seguimiento de locutores al emplear las aproximaciones UBM-g y UBM-t sobre las conjuntos de desarrollo (dev) y evaluación (eval) del corpus AMI. El módulo de detección sigue la metodología empleada en tareas de identificación de locutores.*

		No-calibrado			Calibrado		
		PRC	**RCL**	**medida-F**	**PRC**	**RCL**	**medida-F**
Dev	**rt-UBM-g**	0.66	0.92	0.77	0.81	0.8	0.81
	ref-UBM-g	0.67	0.93	0.78	0.82	0.82	0.82
	rt-UBM-t	0.67	0.91	0.77	0.82	0.83	0.82
	ref-UBM-t	0.69	0.92	0.79	0.84	0.86	0.85
Eval	**rt-UBM-g**	0.69	0.92	0.78	0.78	0.82	0.8
	ref-UBM-g	0.69	0.93	0.79	0.78	0.84	0.81
	rt-UBM-t	0.71	0.91	0.8	0.81	0.85	0.83
	ref-UBM-t	0.72	0.92	0.81	0.81	0.87	0.84

5.2.2.2 Sistema de seguimiento continuo con suavizado

En las **Figuras 5.4 y 5.5** se muestran las curvas DET del sistema UBM-t con las funciones de suavizado, sobre el conjunto de desarrollo y evaluación, respectivamente. La combinación se realiza

con pesos uniformes (función rectangular, función *fs* en adelante), y pesos linealmente decrecientes (función triangular, función *ft* en adelante).

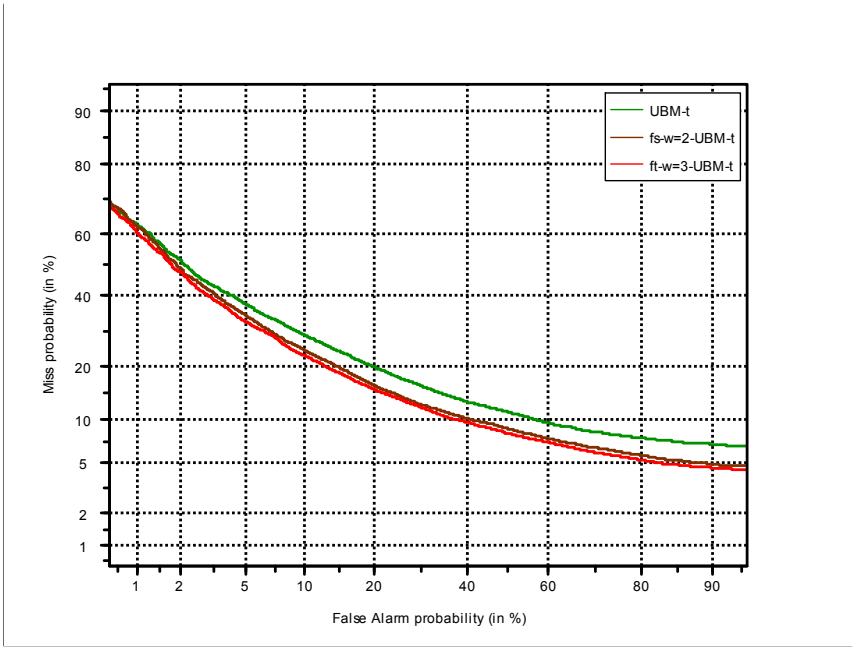

Figura 5.4. *Curvas DET del sistema de seguimiento continuo (UBM-t) sobre el conjunto de desarrollo del corpus AMI al aplicar dos tipos de suavizado: (1) una ventana rectangular sobre los dos últimos segmentos (fs-w=2-UBM-t); y (2) una ventana triangular sobre los tres últimos segmentos (ft-w=3-UBM-t). Como referencia, se incluye la curva del sistema original sin suavizado (UBM-t).*

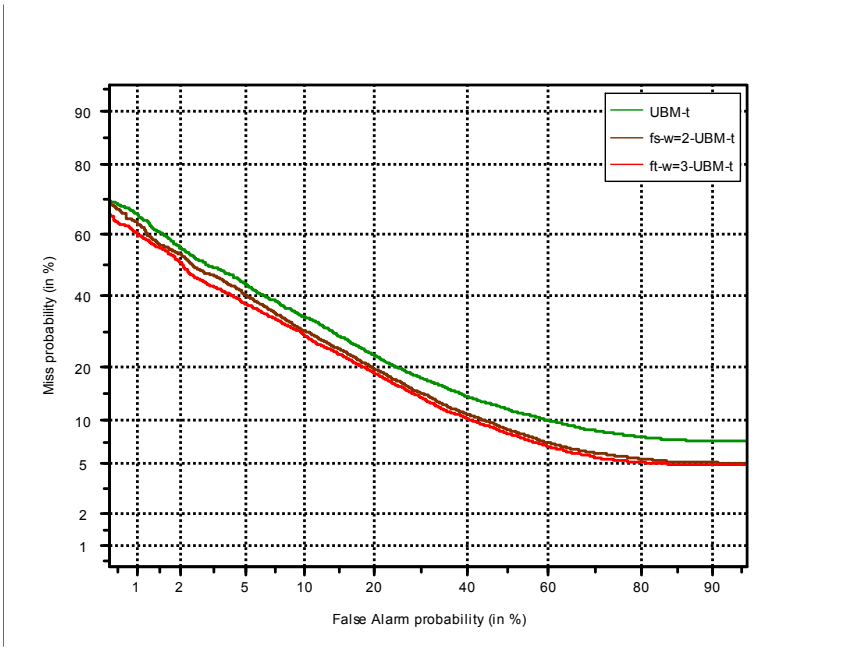

Figura 5.5. *Curvas DET del sistema de seguimiento continuo (UBM-t) sobre el conjunto de evaluación del corpus AMI al aplicar dos tipos de suavizado: (1) una ventana rectangular sobre los dos últimos segmentos (fs-w=2-UBM-t); y (2) una ventana triangular sobre los tres últimos segmentos (ft-w=3-UBM-t). Como referencia, se incluye la curva del sistema original sin suavizado (UBM-t).*

El tamaño óptimo de la ventana se ha determinado de forma heurística sobre el conjunto de desarrollo. Su valor depende de la longitud media de los turnos de locutor. Para *fs*, el valor óptimo es $w = 2$. Para *ft*, $w = 3$. Tal como se aprecia en ambas figuras, emplear información adicional extraída

de segmentos anteriores mejora el rendimiento del sistema consistentemente. Sobre el conjunto de desarrollo (Figura 5.4), el EER del sistema sin suavizado pasa de 18.94 a 16.91 con *fs* ($w = 2$), y a 16.26 con *ft* ($w = 3$), lo que se traduce a una mejora del 10.72% y del 14.15%, respectivamente. Sobre el conjunto de evaluación (Figura 5.5) se observa la misma tendencia: el EER del sistema original ($w = 1$) pasa de 21.14 a 19.37 con *fs* ($w = 2$) y a 18.64 con *ft* ($w = 3$), cifras que demuestran, respectivamente, una mejora relativa del 8.37% y del 11.82%.

La precisión, la cobertura y la medida-F del sistema original sin suavizado y del sistema *ft* se presentan en la **Tabla 5.2**. Se incluyen las tasas obtenidas antes y después de aplicar la calibración. Los resultados vuelven a poner de manifiesto la eficiencia del proceso de calibración. La medida-F del sistema *ft* es superior a la del sistema original, con una mejora relativa del 3.53% y del 3.49% en los conjuntos de desarrollo y evaluación, respectivamente.

Tabla 5.2. *Precisión (PRC), cobertura (RCL) y medida-F del sistema de seguimiento continuo (con UBM-t) sobre las conjuntos de desarrollo (dev) y evaluación (eval) del corpus AMI. Se compara el rendimiento del sistema sin suavizado y el sistema que aplica el suavizado triangular sobre los tres últimos segmentos (ft-w=3-UBM-t). El módulo de detección sigue la metodología empleada en tareas de identificación de locutores.*

		No-calibrado			**Calibrado**		
		PRC	**RCL**	**medida-F**	**PRC**	**RCL**	**medida-F**
Dev	**UBM-t**	0.67	0.91	0.772	0.82	0.83	0.822
	ft-w=3-UBM-t	0.69	0.93	0.79	0.84	0.86	0.851
Eval	**UBM-t**	0.71	0.92	0.8	0.81	0.85	0.831
	ft-w=3-UBM-t	0.73	0.94	0.822	0.83	0.89	0.86

A continuación, se muestran los resultados obtenidos utilizando el sistema UBM-g con suavizado, mediante las funciones *fs* y *ft* anteriormente descritas. Una vez más, el tamaño óptimo de las ventanas se determina de forma heurística sobre el corpus de desarrollo: 2 segundos ($w = 2$) para *fs*, y 3 segundos para *ft* ($w = 3$). En las **Figuras 5.6 y 5.7** se muestran las curvas DET de los tres sistemas (sistema UBM-g, sistema UBM-g con *fs* y sistema UBM-g con *ft*) sobre los conjuntos de desarrollo y evaluación, respectivamente. De nuevo, se observa que el suavizado mejora de forma consistente el rendimiento del sistema. Sobre el conjunto de desarrollo, el EER del sistema pasa de 21.47 a 20.53 al aplicar *fs* ($w = 2$), y a 20.47 al aplicar *ft* ($w = 3$), lo que supone mejoras relativas del 4.38% y del 4.66%, respectivamente. Sobre el conjunto de evaluación el EER pasa de 26.21 a 25.69 con *fs* ($w = 2$) y a 25.3 con *ft* ($w = 3$), con mejoras relativas del 1.98% y 3.47%, respectivamente. Como con la aproximación UBM-t, la función triangular proporciona mejor rendimiento que la función rectangular.

En la **Tabla 5.3** se presentan la precisión, la cobertura y la medida-F del sistema original UBM-g y del sistema UBM-g ft ($w = 3$) sobre los conjuntos de desarrollo y evaluación, antes y después de la calibración. Atendiendo a la medida-F, el sistema UBM-g *ft*, presenta mejoras relativas del 2.11% y el 2.88% sobre los conjuntos de desarrollo y evaluación, respecto al sistema original UBM-g. Estos resultados vuelven a confirmar que el empleo de información adicional proveniente de segmentos

anteriores mejora el rendimiento del sistema, manteniendo los requisitos de continuidad y latencia baja que imponen los escenarios para los que se ha definido el sistema, y que el calibrado de las tasa mejora consistentemente el rendimiento del sistema.

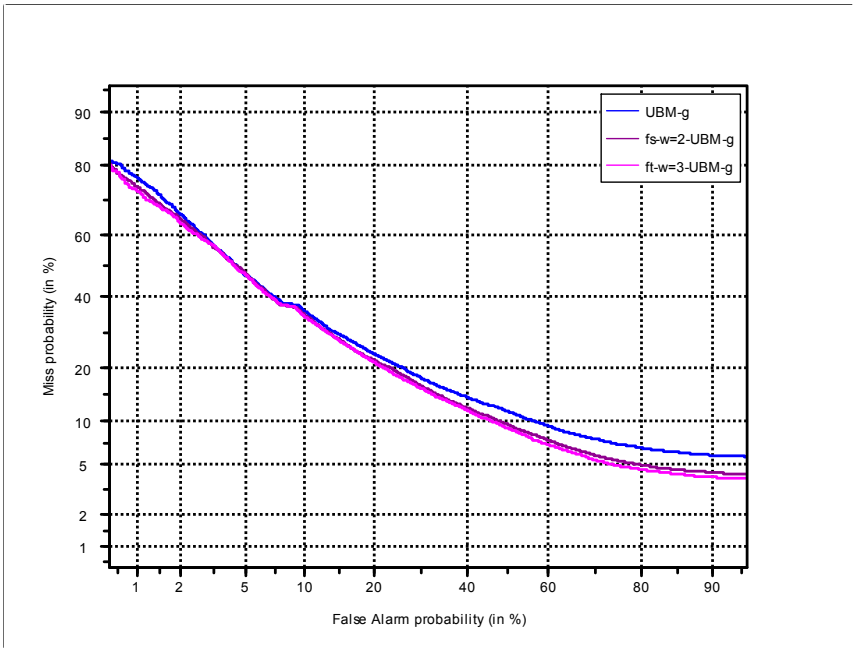

Figura 5.6. *Curvas DET del sistema de seguimiento continuo (UBM-g) sobre el conjunto de desarrollo del corpus AMI, al aplicar dos tipos de suavizado: (1) una ventana rectangular sobre los dos últimos segmentos (fs-w=2-UBM-g); y (2) una triangular sobre los tres últimos segmentos (ft-w=3-UBM-g). Como referencia, se incluye la curva del sistema original sin suavizado (UBM-g).*

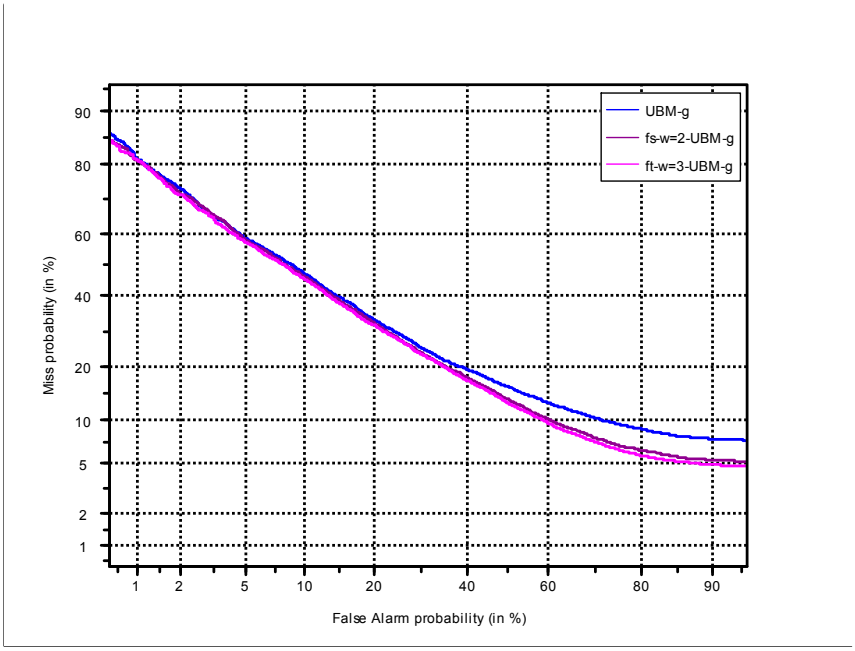

Figura 5.7. *Curvas DET del sistema de seguimiento continuo (UBM-g) sobre el conjunto de evaluación del corpus AMI, al aplicar dos tipos de suavizado: (1) una ventana rectangular sobre los dos últimos segmentos (fs-w=2-UBM-g); y (2) una triangular sobre los tres últimos segmentos (ft-w=3-UBM-g). Como referencia, se incluye la curva del sistema original sin suavizado (UBM-g).*

Tabla 5.3. *Precisión (PRC), cobertura (RCL) y medida-F del sistema de seguimiento continuo (con UBM-g) sobre los conjuntos de desarrollo (dev) y evaluación (eval) del corpus AMI. Se compara el rendimiento del sistema original (sin suavizado) con el sistema que aplica suavizado triangular sobre los tres últimos segmentos (ft-w=3-UBM-g). El módulo de detección sigue la metodología empleada en tareas de identificación de locutores.*

		No-calibrado			Calibrado		
		PRC	**RCL**	**medida-F**	**PRC**	**RCL**	**medida-F**
Dev	**UBM-g**	0.66	0.92	0.769	0.81	0.8	0.806
	ft-w=3-UBM-g	0.67	0.94	0.784	0.81	0.83	0.823
Eval	**UBM-g**	0.69	0.92	0.785	0.78	0.82	0.8
	ft-w=3-UBM-g	0.7	0.95	0.806	0.79	0.86	0.823

5.2.2.3 Detección de segmentos solapados

El seguimiento de locutores puede también realizarse según la aproximación VERL que permite la detección de segmentos en los que dos o más locutores intervienen a la vez. El sistema asume que todos los locutores cuya probabilidad supera el umbral de aceptación hablan en el segmento analizado.

En las **Figuras 5.8 y 5.9** se muestran las curvas DET de los sistemas UBM-g y UBM-t con la aproximación VERL, sobre los conjuntos de desarrollo y evaluación, respectivamente. Dado que VERL permite la detección de segmentos solapados, los sistemas se evalúan tanto excluyendo como incluyendo los segmentos solapados en el conjunto de test. Sobre el conjunto de desarrollo (Figura 5.8), el EER de UBM-g pasa de 14.72 a 18.71 al incluir los segmentos solapados (un deterioro relativo del 27.1%). De forma parecida, el EER del sistema UBM-t pasa de 11.68 a 15.77 (una degradación del 35.02%). Los resultados obtenidos sobre el conjunto de evaluación (Figura 5.9) presentan una tendencia muy similar: el EER pasa de 18.14 a 22.99 con UBM-g (una degradación del 26.72%), y de 12.03 a 17.00 con UBM-t (una degradación del 41.32%). Así, aunque la aproximación VERL permite realizar seguimiento de locutores con detección de segmentos solapados, el rendimiento de los sistemas mejora notablemente al restringir la evaluación a los segmentos no solapados.

Por otro lado, al igual que con la aproximación IDL, el sistema UBM-t presenta un rendimiento muy similar sobre los conjuntos de desarrollo (EER=11.68) y evaluación (EER=14.72). El sistema UMB-g ofrece un rendimiento peor y se degrada considerablemente al pasar del conjunto de desarrollo (EER = 14.72) al de evaluación (EER=18.14), de nuevo, supuestamente, debido al empeoramiento de la cobertura del UBM.

En la **Tabla 5.4** se muestran la precisión, la cobertura y la medida-F del sistema UBM-g con VERL sobre los conjuntos de desarrollo y evaluación. Se compara el rendimiento del sistema al excluir e incluir los segmentos solapados en el conjunto de test. Como referencia, se muestran también las tasas obtenidas por el sistema UBM-g con la aproximación IDL. En este sentido, atendiendo a la medida-F, la aproximación IDL muestra un mejor rendimiento. No obstante, si analizamos los EER obtenidos con ambas aproximaciones, el sistema UBM-g con IDL se sitúa en 21.47 y 26.20 sobre los conjuntos de desarrollo y evaluación, respectivamente (véanse las Figuras 5.2 y 5.3). Al emplear UBM-g con

VERL, el EER es de 14.72 en desarrollo y de 18.14 en evaluación (véanse las Figuras 5.8 y 5.9). Parecen resultados contradictorios, pero, como veremos a continuación, tienen una explicación si reparamos en la naturaleza de las curvas DET.

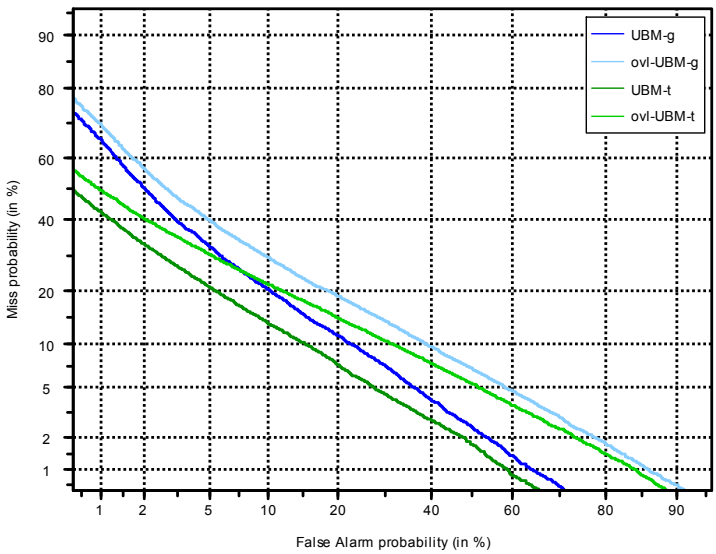

Figura 5.8. *Curvas DET del sistema de seguimiento continuo, con UBM-g y UBM-t, sobre el conjunto de desarrollo del corpus AMI. El módulo de detección sigue la metodología empleada en tareas de verificación, lo que permite la evaluación del sistema incluyendo segmentos solapados (ovl).*

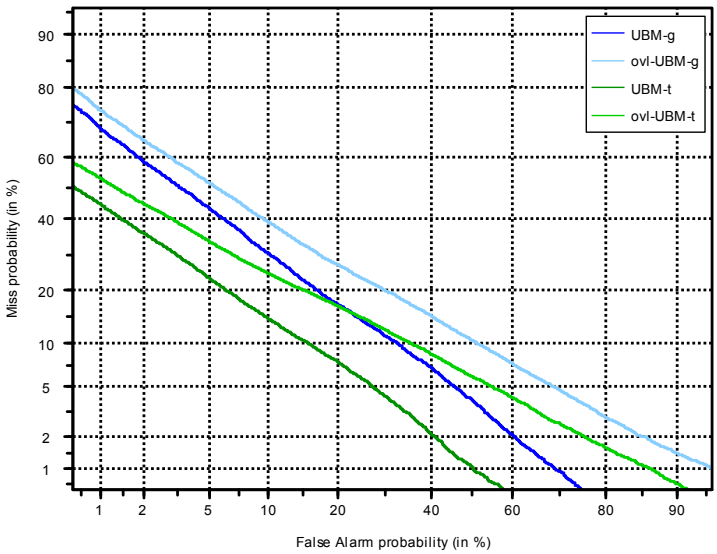

Figura 5.9. *Curvas DET del sistema de seguimiento continuo con UBM-g y UBM-t, sobre el conjunto de evaluación del corpus AMI. El módulo de detección sigue la metodología empleada en tareas de verificación, lo que permite la evaluación del sistema incluyendo los segmentos solapados (ovl).*

Tabla 5.4. *Precisión (PRC), cobertura (RCL) y medida-F del sistema de seguimiento continuo (UBM-g) sobre los conjuntos de desarrollo (dev) y evaluación (eval) de AMI. El módulo de detección sigue la metodología empleada en tareas de verificación, lo que permite la evaluación del sistema incluyendo segmentos solapados (ovl-verl-UBM-g). Como referencia, se muestran los resultados para el sistema que sigue la metodología empleada en tareas de identificación (idl-UBM-g).*

		No-Calibrado			Calibrado		
		PRC	**RCL**	**medida-F**	**PRC**	**RCL**	**medida-F**
Dev	**verl-UBM-g**	0.39	0.95	0.557	0.72	0.78	0.736
	ovl-verl-UBM-g	0.36	0.95	0.527	0.71	0.71	0.711
	idl-UBM-g	0.66	0.92	0.769	0.81	0.8	0.806
Eval	**verl-UBM-g**	0.35	0.97	0.513	0.61	0.78	0.687
	ovl-verl-UBM-g	0.32	0.97	0.478	0.60	0.74	0.664
	idl-UBM-g	0.69	0.92	0.785	0.78	0.82	0.8

La curva DET permite comparar la capacidad global de discriminación de distintos sistemas de detección evaluados sobre el mismo conjunto experimental. En este sentido, se ha de tener en cuenta que el conjunto de experimentos sobre el que se calculan las curvas DET es distinto para las aproximaciones IDL y VERL. La aproximación VERL estima la curva con las tasas de detección de todos los locutores objetivo dado cada segmento del conjunto de test (cada segmento se evalúa más de una vez), mientras que la aproximación IDL estima la curva sólo con la tasa del locutor objetivo más probable dado cada segmento del conjunto de test. Por tanto, las curvas DET del sistema con VERL se estiman a partir de una mayor cantidad de experimentos que los sistemas con IDL, y de hecho, no son curvas comparables. La aproximación VERL, se enfrenta a una cantidad mucho más elevada de segmentos impostor que la aproximación IDL. Debido a que la aproximación VERL rechaza correctamente la mayoría de estos segmentos, las tasas de falsa alarma son inferiores a las tasas que obtiene el sistema al aplicar IDL, y en consecuencia, se obtienen curvas DET mejores.

La medida-F, sin embargo, no depende del volumen del conjunto experimental y no se define en función del porcentaje de segmentos correctamente clasificados, sino en función del tiempo clasificado (detectado) correctamente. En consecuencia, las medidas-F de los sistemas con IDL y VER son directamente comparables. De ahí, al emplear tasas no calibradas, los resultados obtenidos muestran un rendimiento superior de la aproximación IDL, en porcentajes relativos del 38.06 % y 53.02% sobre los conjuntos de desarrollo y evaluación, respectivamente. Con las tasas calibradas, debido al propio efecto que produce la calibración, estas diferencias se atenúan a 9.51% en desarrollo, y 16.45% en evaluación. Así pues, el proceso de calibración mejora considerablemente el rendimiento del sistema con VERL, debido a que la calibración corrige una gran parte de los errores de falsa alarma cometidos por el sistema, a costa de provocar algunos errores de detección. Esto queda reflejado en que el sistema con VERL calibrado muestra una precisión mucho más alta con respecto al sistema no calibrado. En este sentido, al comparar la mejora que introduce la calibración en los sistemas basados en las aproximaciones IDL y VERL, se observa que el sistema VERL presenta una

mejora del 32.14% en la medida-F, mientras que para el sistema IDL esta mejora es tan sólo del 4.81%.

Por otro lado, la evaluación del sistema IDL exige la exclusión de los segmentos solapados del conjunto de test, ya que sólo puede detectar un único locutor para cada segmento de entrada. El sistema VERL, en cambio, permite la detección simultánea de varios locutores. Por ello, en la tabla 5.4 se incluyen las tasas obtenidas sobre todos los segmentos del conjunto de test, incluyendo los segmentos solapados. Los resultados indican que al considerar los segmentos solapados empeora el rendimiento del sistema. Aunque el empeoramiento observado en la medida-F es ligeramente inferior al empeoramiento observado en el EER, en ambos casos se pone de manifiesto que como cabía esperar el sistema tiene una mayor tasa de error con los segmentos solapados.

Por último, en la **Tabla 5.5** se presentan la precisión, la cobertura y la medida-F de los sistemas UBM-t con VERL e IDL, incluyendo las tasas obtenidas al considerar todos los segmentos (solapados o no) del conjunto de test. En general, se confirman los resultados obtenidos con el sistema UBM-g. La aproximación IDL da mejores resultados que la aproximación VERL, y la calibración se revela como un paso imprescindible. De hecho, la calibración mejora la medida-F del sistema IDL en un 6.48% y un 3.87% sobre los conjuntos de desarrollo y evaluación, respectivamente. Para el sistema VERL, estas mejoras son del 20.61% en desarrollo y del 22.68% en evaluación. Al comparar los resultados obtenidos con el sistema VERL sobre el conjunto de test estándar (que excluye los segmentos solapados) y sobre el conjunto de test extendido que incluye los segmentos solapados, la medida-F sufre una degradación del 5.45% con tasas de detección no calibradas, y del 3.49% con tasas calibradas.

Tabla 5.5. *Precisión (PRC), cobertura (RCL) y medida-F del sistema de seguimiento continuo (con UBM-t) sobre las conjuntos de desarrollo (dev) y evaluación (eval) de AMI. El módulo de detección sigue la metodología empleada en tareas de verificación de locutores, lo que permite la evaluación del sistema incluyendo los segmentos solapados (ovl-verl-UBM-t). Como referencia, se añaden los resultados para el sistema que sigue la metodología de tareas de identificación (idl-UBM-g).*

		No-calibrado			**Calibrado**		
		PRC	**RCL**	**medida-F**	**PRC**	**RCL**	**medida-F**
Dev	**verl-UBM-t**	0.51	0.94	0.66	0.79	0.8	0.796
	ovl-verl-UBM-t	0.47	0.94	0.624	0.76	0.78	0.768
	idl-UBM-t	0.67	0.91	0.772	0.82	0.83	0.822
Eval	**verl-UBM-t**	0.49	0.96	0.648	0.76	0.83	0.795
	ovl-verl-UBM-t	0.44	0.96	0.604	0.72	0.81	0.761
	idl-UBM-t	0.71	0.92	0.8	0.81	0.85	0.831

5.3 AmISpeaker: Implantación del sistema de seguimiento de locutores continuo

En este apartado se detalla el proceso de implantación del sistema de seguimiento continuo propuesto y evaluado en el apartado anterior. Se describe la arquitectura y la implementación de la aplicación de seguimiento, denominada *AmISpeaker*, que se ha desarrollado.

5.3.1 Descripción de la aplicación AmISpeaker

Debido a la naturalidad que presenta el habla como medio de interacción entre personas y máquinas, y el potencial de aplicación que presenta el seguimiento de locutores como tecnología de seguimiento de usuarios de un entorno AmI, se ha desarrollado una aplicación de seguimiento de locutores, denominada *AmISpeaker*, la cual implementa el sistema de seguimiento continuo propuesto y evaluado en la primera parte de este capítulo. Las características más destacables de dicho sistema son la continuidad (es decir, la capacidad e operar de forma *online*), la baja latencia en sus decisiones y el buen rendimiento que presenta con respecto a un sistema tradicional, de referencia, que, como la mayoría de aproximaciones de la bibliografía sigue un procedimiento iterativo, no aplicable en aproximaciones continuas. La continuidad y baja latencia son características esenciales para cualquier aplicación en un entorno AmI.

AmiSpeaker parte de un conjunto de locutores objetivo y una señal de entrada continua, sobre la cual aplica técnicas del campo de la identificación de locutores, para determinar los instantes de tiempo en los que habla alguno de los locutores objetivo. La señal a analizar se adquiere mediante un micrófono embebido en el entorno AmI, en este caso, en el hogar inteligente. Si el entorno AmI se delimitara por zonas, con uno o varios micrófonos por cada zona, el mismo procedimiento de seguimiento de locutores se podría aplicar para determinar la posición o localización relativa de los locutores. El desarrollo actual del sistema no incluye ni la delimitación del espacio por zonas ni la localización de los locutores.

Por otro lado, AmISpeaker sigue el paradigma SOC [Papazoglou07], ya que SOC promueve el desarrollo de aplicaciones distribuidas, ligeras e interoperables a partir de servicios. Estos servicios se entienden como entidades autónomas, independientes del lenguaje de programación o sistema operativo, entidades que se pueden descubrir, publicar, describir y componer de forma dinámica. El elemento principal de SOC, es la arquitectura orientada a servicios (SOA) [SOAa] [SOAb] [Arsajani04]. SOA define una estructura lógica, distribuida e interoperable para el diseño de aplicaciones distribuidas a partir de servicios como componentes de software. Los sistemas se han de construir a partir de servicios que cubren las necesidades de las aplicaciones que interactúan con el usuario final o, las necesidades de otros servicios distribuidos en la red. Para ello, emplea interfaces estándares públicos y descubribles. En definitiva, SOA propone la estructuración en servicios de las componentes de software. En los últimos años, se han desarrollado diversas tecnologías para la

definición de este tipo de componentes, que posibilitan la implementación de aplicaciones. Entre ellas, cabe destacar los servicios web [WebServices] [Weerawarana05], UPnP (*Universal Plug and Play*) [UPnP], Jini [Jini], OSGi [OSGi], etc. AmISpeaker es un servicio de seguimiento de locutores que sigue el estándar SOA, concretamente a través de la tecnología UPnP.

A continuación, se describen la arquitectura de software del servicio AmISpeaker y la fase de implementación de la aplicación.

5.3.1.1 Arquitectura de software del servicio de seguimiento de locutores

En la **Figura 5.10** se ofrece una representación esquemática del servicio, el cual, a su vez, se compone de varios servicios auxiliares:

Figura 5.10. *Representación esquemática de un servicio de seguimiento de locutores integrable en las aplicaciones que siguen una arquitectura orientada a servicios (SOA).*

- **Servicio de detección de locutores** (*Speaker Detection Service*, SDS): toma como entrada las muestras de audio adquiridas por un micrófono del entorno, y da como salida, cada segundo, la probabilidad de cada locutor objetivo para las muestras correspondientes al último segundo de

señal. Para ello, el servicio dispone de tres módulos (descritos en el apartado 5.2.1): (1) un módulo de parametrización en tiempo real que extrae la secuencia de vectores acústicos de las muestras adquiridas por el micrófono; (2) un módulo de detección de locutores, que a partir de un segmento parametrizado (de un segundo de duración) y los modelos de los locutores objetivo, estima la tasa de detección de cada locutor; y, por último, (3) un módulo de calibración, el cual mapea las tasas estimadas a probabilidades o ratios teóricos.

- **Servicio de seguimiento de locutores** (*Speaker Tracking Service*, STS): se encarga de realizar el seguimiento de locutores de un espacio *E*, en base a las probabilidades emitidas por el servicio SDS_{XE} asociado al micrófono m_{XE} de *E*; es decir, determina la identidad del locutor (o locutores) que en ese instante hablan en el micrófono m_{XE}. Es el servicio que decide si el segmento corresponde a alguno de los locutores objetivo del sistema: si es así, reporta su identidad; en caso contrario, asume que el segmento proviene de un locutor desconocido o es un segmento de silencio, ruido, etc. La toma de decisiones se realiza mediante el módulo de seguimiento de locutores, el cual puede ejecutarse como una tarea de identificación de locutores, o como una tarea de verificación de locutores (véase el apartado 5.2.1). Por otro lado, el servicio incluye también un módulo de suavizado, que trata de mejorar la robustez del sistema empleando información obtenida de segmento anteriores.
- **Servicio de seguimiento de locutores de una zona** *(Zone Speaker Tracking Service,* ZSTS*):* un servicio ZSTS se basa en la conceptualización de que un entorno AmI se puede dividir en zonas separadas, independientes e interconectadas (por ejemplo, las habitaciones de una casa), y cada zona en espacios contiguos, pegados, los cuales pueden o no ser independientes (supóngase una cocina en la que distinguen el espacio para preparar las comidas y el espacio para su consumo). De esta forma, tal como se representa en la Figura 5.10, si en cada espacio *E* se instala un micrófono (o más de uno), y el ZSTS conoce la ubicación física de dichos micrófonos y además analiza los resultados de seguimiento emitidos por los STS de la zona (recuérdese que cada servicio STS se asocia con un micrófono), el ZSTS puede inferir los movimientos realizados por los locutores, es decir, puede determinar su localización espacial relativa (como por ejemplo, que el usuario se encuentra en el espacio 1A de la zona A, o que se ha movido del espacio 1A al espacio 1C). Para ello, obviamente, es también imprescindible que se almacene el historial de los movimientos y las detecciones realizadas. De forma alternativa, el ZSTS puede también aplicar herramientas de inferencia estadística para indicar el grado de fiabilidad de un resultado obtenido con cada STS de la zona, para aumentar la robustez de un STS, en el caso de que un ZSTS observe como salida el mismo locutor en dos STS asociados a dos micrófonos muy próximos (de un mismo espacio), o disminuirla cuando se detecte que dos STS próximos emiten como salida dos locutores distintos.
- **Servicio de seguimiento de locutores de un entorno AmI** *(AmI Speaker Tracking Service,* ASTS): el funcionamiento de un servicio ASTS es muy similar al de un ZSTS. Se basa en la conceptualización de que un entorno AmI se divide en varias zonas, donde cada zona contiene un

ZSTS que realiza el seguimiento de los locutores presentes, indicando cual es su identidad, así como su localización relativa con respecto a los espacios definidos en la zona. Para ello, analiza los resultados emitidos por el servicio ZSTS de cada zona, de la misma forma que realiza un ZSTS con respecto a un STS.

5.3.1.2 Implementación de la aplicación

La aplicación AmISpeaker sigue la arquitectura descrita más arriba, pero debido a la carencia de un escenario AmI real, el desarrollo sólo se ha realizado hasta el nivel del servicio STS (el servicio de seguimiento local). Este nivel incorpora el sistema de seguimiento de locutores propuesto y evaluado en la primera parte del capítulo. Queda pendiente como trabajo futuro el desarrollo, la evaluación e implementación de los servicios ZSTS y ASTS.

Por tanto, *AmISpeaker*, en su desarrollo actual, es un servicio STS (servicio de seguimiento) realizado mediante la tecnología UPnP. Consta de un SDS (un servicio de detección) asociado a un micrófono embebido en el entorno. Este servicio, a intervalos periódicos, emite la tasa de detección de cada locutor objetivo dado el segmento de audio correspondiente al intervalo. La señal analizada proviene de la señal adquirida por el micrófono al que está asociado el servicio. Por defecto, los intervalos, es decir los segmentos analizados, son de un segundo de duración, coincidiendo con la longitud empleada para la evaluación del sistema de seguimiento sobre la base de datos AMI. La aplicación AmISpeaker, que incluye un servicio STS, es la que realiza el seguimiento de los locutores. Determina, de forma continua, el locutor más probable del último segmento analizado, pudiendo emplear alguna función de suavizado para ello. El funcionamiento es el siguiente:

- AmISpeaker se suscribe a un servicio SDS, mediante el cual se puede activar y desactivar la detección de locutores de la zona correspondiente al micrófono que está asociado el SDS.
- El método que activa la detección permite especificar la longitud (en tiempo) de los segmentos a analizar.
- Mientras la detección sigue activa, SDS estima de forma continua las tasas de detección de los locutores objetivo dado cada segmento analizado. Además, comunica a todos sus suscriptores las tasas estimadas.
- AmISpeaker, anuncia las decisiones tomadas mediante el servicio STS a todos sus suscriptores (los servicios ZSTS, si los hubiera), y además las muestra en pantalla.

5.4 Conclusiones

En este capítulo se ha presentado un sistema de seguimiento de locutores para su aplicación en hogares inteligentes o en cualquier otro escenario AmI que presente una cantidad limitada de usuarios (conocidos de antemano). El escenario planteado requiere un sistema de seguimiento continuo, que proporcione respuestas y decisiones inmediatas, de latencia baja. La mayoría de aproximaciones de seguimiento de locutores de la bibliografía no observan los requisitos de continuidad y baja latencia.

Son aproximaciones iterativas o multi-fase, que procesan los datos de principio a fin de forma cíclica, ya que se diseñan para procesar recursos de audio previamente grabados. Con el objetivo de cumplir estos requisitos de continuidad y latencia baja, el sistema desarrollado sigue un algoritmo muy simple, que clasifica y segmenta la señal de audio de entrada de forma simultánea. El sistema procesa y toma las decisiones sobre segmentos de longitud fija. La detección de locutores se realiza por medio modelos GMM(MAP) que presentan un buen rendimiento y son eficientes computacionalmente.

Asimismo, el sistema pretende abordar el desarrollo de aplicaciones de seguimiento de locutores según la especificación SOA, para alcanzar la interoperabilidad y movilidad que requiere un entorno AmI. Se ha definido una arquitectura SOA con dos tipos de servicios: servicios de seguimiento de locutores (STS); y servicios de detección de locutores (SDS), auxiliares, a los cuales se suscriben los STS (véase la Figura 5.10). Los SDS capturan la señal de audio grabada por un micrófono embebido en el entorno AmI y determinan una tasa de detección para cada locutor objetivo a intervalos periódicos de 1 segundo. Por medio de la composición de servicios, el STS analiza las salidas de SDS y toma las decisiones de seguimiento correspondientes.

El sistema se ha desarrollado y evaluado por medio de la base de datos de reuniones AMI. Se ha definido un conjunto de desarrollo para configurar los parámetros del sistema, y un conjunto de evaluación para evaluar el rendimiento del sistema con la configuración establecida en desarrollo. Asimismo, se definen dos sistemas: UBM-g, en el que el UBM se estima a partir de un conjunto de datos independiente de los conjuntos de desarrollo y evaluación; y UMB-t, en el que el UBM es dependiente de los locutores objetivo, ya que se estima a partir de sus datos de entrenamiento. Por otro lado, el esquema de seguimiento permite también la aplicación de dos aproximaciones diferentes: (1) identificación de locutores (denominada IDL, de forma abreviada), donde la decisión se toma en base a la tasa de detección del locutor más probable; y (2) verificación de locutores (denominada VERL), donde para tomar las decisiones se analizan las tasas de detección de todos los locutores objetivo, de modo que todos aquellos locutores cuyas tasas de detección superen el umbral son aceptadas, y en consecuencia pueden detectarse varios locutores simultáneamente.

Los resultados muestran un mejor rendimiento de los modelos UBM-t que los UBM-g. Esto sugiere la estimación de un UBM específico al entorno (sala) de evaluación y a los propios locutores, con la ventaja de no necesitar ningún conjunto de datos adicional al de entrenamiento para estimar los modelos de locutor. No obstante, es necesario estimar un UBM específico para cada conjunto de locutores objetivo. UBM-g, en cambio, puede aplicarse sobre cualquier conjunto de locutores objetivo.

En cuanto a las aproximaciones IDL y VERL, destaca la superioridad en rendimiento de IDL. La aproximación VERL permite detectar varios locutores simultáneamente, pero su rendimiento sobre el conjunto de test restringido (sin segmentos solapados) es peor que el de la aproximación IDL. Si se considera el conjunto de test completo, en base a la medida-F, el sistema VERL sufre una degradación del 3.7% con UBM-g y del 3.96% con UBM-t, lo que sugiere explorar otro tipo de metodologías para la detección de segmentos solapados.

El rendimiento del sistema propuesto se ha comparado con el de un sistema de referencia que sigue un procedimiento desacoplado, con una fase de segmentación y otra de detección. El sistema de referencia ofrece un rendimiento mejor que el sistema propuesto. Con la aproximación UBM-t la medida-F pasa de 0.84 con el sistema de referencia a 0.83 con el sistema propuesto (lo que supone un deterioro del 1.2%). Pero en función de la latencia requerida por el escenario, la segmentación previa del recurso de audio puede no resultar posible, y en consecuencia, ser preferible la aplicación del sistema propuesto en este trabajo, el cual cubre los requisitos de continuidad y latencia baja, a costa de una muy pequeña degradación del rendimiento, que consideramos aceptable para la mayoría de escenarios de aplicación en entornos AmI.

Por otro lado, cabe destacar la eficiencia del proceso de calibración, introducido en el sistema con el propósito de establecer de forma automática un umbral de aceptación óptimo. Teóricamente, la calibración realiza la transformación de las tasas de detección a ratios de probabilidad reales, tratando de maximizar la información mutua de las tasas obtenidas sobre el conjunto de desarrollo. La calibración ha conllevado mejoras notables de rendimiento en todos los experimentos realizados, sin aumentar el coste computacional ni la latencia del sistema, requisitos fundamentales del sistema planteado.

Por último, con el objetivo de mejorar la robustez del sistema frente a la variabilidad local proveniente de la corta longitud de los segmentos analizados, se ha aplicado un esquema de suavizado que realiza una combinación lineal de las tasas de detección en un cierto número de segmentos anteriores. El suavizado proporciona, en términos de EER, una mejora del 11.82% con UBM-g y del 3.47% con UBM-t. En términos de la medida-F, el rendimiento del sistema mejora en un 3.47% al emplear UBM-t, y en un 2.88% al emplear UBM-g. La longitud de la ventana de suavizado utilizada en estos experimentos es 3. Esta longitud se ha determinado de forma empírica sobre el conjunto de desarrollo. Queda pendiente analizar nuevos métodos de suavizado, o incluso métodos de segmentación computacionalmente eficientes (de latencia baja) que determinen la longitud óptima de la ventana de análisis. Asimismo, se podría estudiar la aplicación de una función de suavizado a nivel de tramo, y no a nivel de segmento como se ha hecho en este trabajo.

De cara al futuro, se plantean las siguientes líneas de trabajo:

- El empleo de un *back-end* de detección de locutores más robusto, basado en SVM, con un kernel GLDS o vectores soporte GMM. Además, se podría introducir un módulo dedicado a fusionar los diferentes sistemas de detección (p.ej. GMM(MAP) y GLDS-SVM, etc.).
- La introducción de un módulo de segmentación acústica, de baja latencia, que operando de forma continua, determine las fronteras de los segmentos a transferir al modulo de detección de locutores.
- La modificación de la configuración del sistema a un sistema como el planteado en [Markov07] (cuya descripción se encuentra en el apartado 2.5.2), un sistema auto-suficiente de

administración-cero, que posea la capacidad de aumentar y disminuir el conjunto de locutores objetivo de forma automática.

- El desarrollo y la evaluación del servicio de seguimiento de locutores de una zona (ZSTS), definida en la arquitectura propuesta. Esta idea requiere la estructuración de un laboratorio o sala inteligente que cumpla los requisitos necesarios para ello.
- La extensión de la arquitectura SOA propuesta para el seguimiento de locutores (Figura 5.10) a un *framework* de reconocimiento del habla. Este *framework* daría cobertura a la mayoría de aplicaciones del campo del procesamiento de voz (aplicaciones de reconocimiento de locutores, de idioma, del habla, del estado emocional del locutor, etc.) que sigan el estándar SOA. El *framework* de voz se podría plantear mediante una arquitectura en tres capas con diferentes servicios:
    - La capa inferior se compondría de un único servicio de captura de audio que publique en la red la señal de audio grabada desde un micrófono. Se plantea la introducción de un servicio por cada micrófono del entorno.
    - La capa intermedia se compondría de dos tipos de servicios (posiblemente interconectados) y asociados a cada micrófono del entorno: (1) servicios de segmentación acústica de la señal de entrada; y (2) servicio de detección de voz/no-voz que filtren la señal.
    - La capa superior, que daría cobertura a las aplicaciones del entorno, consistiría en diferentes servicios del campo de procesamiento de voz: servicio de reconocimiento de locutores, del idioma, del mensaje emitido, de detección del estado emocional del locutor, etc. que operarían en base a las señales publicadas por los servicios de la capa intermedia.

Capítulo 6

Conclusiones y trabajo futuro

6.1 Resumen global de las conclusiones de la tesis

Esta tesis ha pretendido abordar la integración de los sistemas de reconocimiento de locutores en entornos AmI, con objeto de proveer una interfaz natural para la detección, adaptación y monitorización del usuario. Este objetivo implica un amplio rango de tecnologías y líneas de investigación, y parte de la necesidad de madurez que presentan algunas de estas tecnologías y técnicas involucradas en su realización. Para ofrecer una interacción natural, los sistemas, además de ser proactivos y personalizados, deberán estar embebidos en el entorno, lo que implica la necesidad de sistemas que, con un rendimiento aceptable, requieran una mínima cantidad de recursos en lo que a computación y almacenamiento respecta. Además, deberán determinar sus respuestas con una latencia baja, prácticamente de forma instantánea. Serán sistemas continuos y auto-configurables (de administración cero), garantizando que el usuario pueda consumir los servicios prestados en cualquier instante, sin necesidad de llevar a cabo una tarea específica de entrenamiento o configuración. Por otro lado, cabe destacar que el trabajo no trata de abordar la integración definitiva (objetivo que abarca una gran diversidad de disciplinas científicas), sino que pretende realizar una primera aproximación, analizando, por un lado, las posibilidades de reducción del coste computacional de los sistemas de reconocimiento de locutores, y, por otro lado, la definición de sistemas continuos y de baja latencia.

A continuación, se resumen las conclusiones más relevantes de las diferentes líneas de trabajo que se han seguido en la elaboración de la tesis.

6.1.1 Reducción de dimensionalidad de la parametrización acústica

En lo que a la reducción de dimensionalidad de la representación acústica de las señales respecta, se han estudiado y validado diferentes metodologías que, partiendo de un conjunto de parámetros acústicos estándar (MFCC, la energía y sus derivadas) empleado por la mayoría de sistemas tanto para el reconocimiento del habla como para el reconocimiento de locutores, tratan de definir subconjuntos reducidos atendiendo a su aplicación en la clasificación de locutores. Su objetivo es ahorrar esfuerzo computacional con una baja degradación o incluso una mejora del rendimiento del sistema.

Como metodología principal, se ha propuesto y comprobado la validez de los algoritmos genéticos (GA) para la selección de un subconjunto óptimo de parámetros. La evaluación se ha llevado a cabo sobre las bases de datos Albayzín y Dihana que contienen distintos tipos de habla y provienen de canales diferentes (Albayzín contiene habla grabada en condiciones de laboratorio, mientras que Dihana contiene habla adquirida por canal telefónico). De ahí:

- Se ha determinado que los parámetros estáticos (MFCC) son más relevantes que los dinámicos (sus derivadas) para el reconocimiento de locutores.
- Se ha determinado que el uso de la energía, el primer cepstral, solo es valioso en condiciones de laboratorio, ya que desaparece en los subconjuntos de pequeña dimensión en Dihana.
- Se ha validado el empleo de GA para resolver este tipo de búsquedas heurísticas, dado que los resultados obtenidos han sido consistentes a lo largo de todos los experimentos realizados. La prevalencia de los parámetros estáticos y la falta de robustez de la energía es un comportamiento que se observa en todos ellos.

Por lo tanto, si debido a limitaciones de almacenamiento y capacidad computacional se tuviera que emplear un conjunto reducido de parámetros, los parámetros estáticos constituirían la mejor solución (conjunto que se debería aumentar con la energía si el habla se adquiere en laboratorio).

Como metodología de contraste, se ha analizado el empleo de las transformaciones lineales PCA, LDA y HLDA. Las transformaciones lineales son técnicas muy empleadas para la reducción de dimensionalidad de un espacio vectorial, ya que no sólo seleccionan parámetros, sino que además combinan y reordenan. En este caso, se ha observado que:

- En aquellos casos en que las señales puedan presentar correlaciones (como la que implica un canal con ruido) las trasnformaciones lineales presentan un mejor rendimiento que los GA, lo que puede atribuirse al hecho de que la aproximación con GA sólo lleva a la selección de parámetros, no a recombinaciones.
- En casos con habla limpia, los GA han mostrado un rendimiento mejor, como consecuencia de que optimizan una función objetivo que precisamente es la tasa de reconocimiento del sistema, frente a las transformaciones lineales que buscan optimizar criterios estadísticos de máxima variabilidad (PCA), máxima discriminación (LDA), etc.
- De estas dos conclusiones se plantea la posibilidad de generalizar la aplicación de los GA para que permitan transformaciones de los vectores, como se comentará más adelante.

6.1.2 Metodologías de modelado computacionalmente eficientes, robustos y adaptables

La integración de los sistemas de reconocimiento locutores en dispositivos computacionales con recursos reducidos, requiere el empleo de metodologías de modelado y de detección eficientes y adaptables. En este sentido, se han estudiado metodologías de modelado de referencia definidas en los últimos años. La evaluación se ha llevado a cabo sobre base de datos locales (de ámbito estatal), y bases de datos creadas para las evaluaciones del NIST reconocidas a nivel mundial.

La primera fase de evaluación ha consistido en comparar el rendimiento de las metodologías más clásicas, GMM(EM) y GMM(MAP), sobre Albayzín. De forma adicional, con el objetivo de analizar el comportamiento de estas metodologías al detectar señales que contienen fuentes acústicas no modeladas en entrenamiento, el conjunto de test se ha ampliado con señales de música, conversaciones telefónicas y audio descargado al azar desde internet. Se observa que:

- Los modelos GMM(MAP) presentan un mejor rendimiento que los modelos GMM(EM), tanto para tareas de identificación como para tareas de verificación, y de forma especial sobre el conjunto de test ampliado, ya que su rendimiento se mantiene en el mismo margen de eficiencia que sobre el conjunto inicial sin señales no modeladas.
- El rendimiento de los modelos GMM(EM) se deteriora considerablemente al incluir las señales no modeladas. Se cree que puede ser debido a que, en estos sistemas, el ratio se calcula a partir de dos modelos estimados de forma independiente. En consecuencia, ante una señal no modelada, las probabilidades condicionales son prácticamente impredecibles, y en consecuencia, el ratio se convierte en un valor aleatorio. En los modelos GMM(MAP), en cambio, el modelo de locutor representa el desplazamiento relativo al locutor de las fuentes modeladas en el UBM, por lo que generalmente, una fuente no representada en el UBM, tampoco lo estará en el modelo de locutor. Ante una señal no modelada, las probabilidades condicionales del ratio obtendrán un valor similar, pequeño por lo general, apuntando hacia el rechazo de este tipo de señales, como debe ser.

A raíz de este estudio, y con objeto de ampliar la robustez de estas metodologías de modelado en la detección de señales no modelas, surge la definición del modelo SSM, denominado modelo superficial de fuente. Trata de representar la fuente de la señal de entrada, mediante un GMM de baja resolución, es decir un GMM compuesto por una pequeña cantidad de componentes, estimado a partir de la propia señal a evaluar. El SSM es una metodología propuesta en esta tesis, la cual:

- Se podría emplear como único factor de normalización del modelo de locutor, si su estimación fuera perfecta, dado que indicaría el grado de aproximación del modelo de locutor a la fuente de la señal a evaluar. Sin embargo, en los experimentos realizados se ha observado que SSM presenta una gran sensibilidad a las variaciones específicas de la señal, y en consecuencia, puede no representar de forma robusta la fuente.
- Presenta un alto rendimiento combinado linealmente con el UBM. La combinación lineal del UBM y el SSM como factor de normalización del modelo de locutor se plantea para, por un lado, mejorar la cobertura del UBM, y por otro, ampliar la robustez del SSM. Es una solución que ha

mostrado mejoras significativas, sobre todo con los modelos GMM(EM) que presentaban un mayor margen de mejora. De hecho, es el alto rendimiento que presenta la combinación de UBM y SSM una de las conclusiones más significativas que se puede extraer de esta experimentación.

Por último, cabe destacar que la experimentación sobre Albayzín se ha realizado con el objetivo de analizar el comportamiento de las metodologías clásicas sobre señales cuyo contenido acústico es muy diferente al contenido de las señales de entrenamiento, y suponiendo en todo caso que el locutor objetivo está ausente.

En la segunda fase de evaluación, llevada a cabo sobre la tarea principal del corpus de la NIST SRE06, se han analizado dos de las mejores técnicas de modelado que se han definido en los últimos años: GMM(MAP) y GMM-SVM, junto con diversas metodologías para normalización de probabilidades (Z-norm, T-norm, ZT-norm), y la calibración y fusión de sistemas. En los resultados obtenidos:

- Se observa una clara superioridad del sistema GMM-SVM.
- Las metodologías de normalización conducen a una ligera mejoría del rendimiento.
- Se observa que Z-norm y ZT-norm presentan el mejor rendimiento al emplear modelos GMM(MAP), mientras que la T-norm es la mejor opción con GMM-SVM.
- La fusión de los sistemas definidos produce también una ligera mejora en el rendimiento y calibra debidamente los sistemas.

De forma implícita, la evaluación sobre los conjuntos experimentales del NIST ha conllevado el análisis y la aplicación de las métricas de evaluación de referencia. Pero quizás, el logro más destacable de este estudio es la participación en la evaluación del NIST del 2008 (la NIST SRE08), para la cual se definieron los sistemas descritos en [Zamalloa08d], con buenos resultados teniendo en cuenta que era un marco de evaluación desconocido para todos los participantes involucrados en el desarrollo de los sistemas. Posteriormente, se participó en la evaluación de sistemas reconocimiento del idioma que organizó el NIST en 2009 (NIST LRE09). El sistema definido en este último caso obtuvo muy buenos resultados. Incluía, entre otros, dos sistemas GMM-SVM que, al igual que para el reconocimiento de locutores, han obtenido buenos resultados también para la detección del idioma (véase [Penagarikano09]). Por último, cabe destacar que, a nivel de grupo de investigación, se siguen elaborando algunas de las líneas de trabajo que se han abierto en esta tesis. Así pues, con objeto de perfeccionar las técnicas empleadas y mejorar el rendimiento de los sistemas de reconocimiento de locutores, se ha participado en la evaluación NIST SRE de este mismo año (la NIST SRE10).

6.1.3 Diseño y evaluación de un sistema de seguimiento de locutores continuo y de latencia baja

Debido al potencial de aplicación que presenta el reconocimiento de locutores como tecnología de seguimiento de usuarios, sobre todo por la naturalidad y transparencia que permite en la interacción con el sistema, se ha propuesto un sistema de seguimiento de locutores continuo, de baja latencia, para

su aplicación en un hogar inteligente como escenario AmI (cantidad limitada de usuarios conocidos: habitantes o individuos asiduos). Para ello, primeramente, se han analizado las aproximaciones de seguimiento y diarización de locutores de la bibliografía. Se ha observado que la mayoría no cumplen los requisitos de continuidad y latencia baja, ya que siguen un procedimiento desacoplado, con una fase de segmentación y otra de detección, las cuales se aplican secuencialmente. Con el objetivo de cumplir estos los requisitos de continuidad y latencia baja, el sistema definido sigue un algoritmo muy simple: la segmentación y clasificación se realizan de forma conjunta mediante el procesamiento de segmentos de longitud fija; y a cada segmento se le asigna una probabilidad para cada locutor objetivo por medio de modelos GMM(MAP) que presentan un buen rendimiento para la detección de segmentos de duración limitada y son eficientes computacionalmente.

Al igual que para las tareas de identificación y verificación de locutores, la evaluación del sistema de seguimiento ha requerido un estudio previo de las infraestructuras necesarias. Se ha elegido la base de datos de reuniones AMI para la fase experimental, debido a que es una base de datos estándar grabada en un entorno inteligente muy similar al entorno de aplicación para el cual se ha diseñado el sistema. Con objeto de contrastar su rendimiento, el sistema propuesto se ha comparado con un sistema de referencia que sigue un procedimiento desacoplado, con una fase de segmentación y otra de detección. La conclusión es que:

- Es factible la utilización del sistema, ya que la degradación relativa resulta ser sólo del 1.2% (con respecto a la medida-F). En función de la latencia requerida por el escenario de aplicación, la segmentación previa del recurso de audio puede no resultar posible, y en consecuencia, resultaría preferible la aplicación de un sistema de seguimiento continuo, que cumple los requisitos de continuidad y latencia baja, a costa de una muy pequeña degradación del rendimiento, que se puede considerar aceptable para la mayoría de escenarios de aplicación de un entorno AmI.

Para mejorar la robustez del sistema propuesto frente a la variabilidad local consecuencia de la corta longitud de los segmentos, se ha propuesto la aplicación de un esquema de suavizado que combina linealmente las tasas de detección de un cierto número de segmentos anteriores. Con ventanas de 3 segundos (longitud determinada de forma empírica sobre un conjunto de desarrollo, ya que depende de la media de duración de los turnos de locutor):

- El rendimiento del sistema presenta una mejora del 3.18%, en promedio (con respecto a la medida-F). De todas formas, queda pendiente analizar nuevos métodos de suavizado, o incluso métodos de segmentación de latencia baja que determinen la longitud óptima de la ventana de análisis.

La definición del sistema pretende también abordar el desarrollo de aplicaciones de seguimiento de locutores según la especificación SOA, un estándar que ofrece un marco muy apropiado para el desarrollo de aplicaciones de un entorno AmI, dado que proporciona movilidad e interoperabilidad a las aplicaciones. En esta línea, se ha desarrollado una aplicación de seguimiento de locutores, denominada *AmISpeaker*, que consta de dos tipos de servicios:

(1) Servicios de detección de locutores que capturan la señal de audio grabada por un micrófono y determinan una tasas de detección para cada locutor objetivo a intervalos periódicos

(2) Servicios de seguimiento de locutores, los cuales por medio de la composición de servicios, toman las decisiones de seguimiento en base a las salidas de los servicios de detección del entorno.

Como aportación final, con el objetivo de analizar la aplicabilidad de las tecnologías analizadas y del sistema definido, se ha desarrollado un demostrador de *AmISpeaker*.

6.1.4 Publicaciones

En la página web del grupo de trabajo (concretamente, en la dirección http://gtts.ehu.es/gtts/?p=publications&b=zamalloa) se muestra el listado de las contribuciones realizadas en congresos y workshops nacionales e internacionales. Como se puede observar, el listado incluye algunas contribuciones correspondientes a la colaboración que este trabajo ha supuesto para otros temas de estudio. A continuación, se listan únicamente las que están relacionadas con el ámbito de investigación de la tesis:

- Zamalloa M., Rodriguez-Fuentes L.J., Bordel G., Penagarikano M., Uribe J.P. **Low-Latency Online Speaker Tracking on the AMI Corpus of Meeting Conversations**. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'10*, Texas, USA, Marzo 2010.
- Zamalloa M., Penagarikano M., Rodriguez-Fuentes L.J., Bordel G., Uribe J.P. **An Online Speaker Tracking System for Ambient Intelligence Environments**, *Proceedings of the Second International Conference on Agents and Artificial Intelligence*, *ICAART*, Enero 2010.
- Zamalloa M., Rodríguez-Fuentes LJ., Penagarikano M., Bordel G., Uribe JP. **Feature Selection vs Feature Transformation in Reducing Dimensionality for Speaker Recognition**. *Proceedings of VJTH 2008*, Bilbao, Noviembre 2008.
- Zamalloa M., Bordel G., Rodríguez LJ., Penagarikano M., Uribe JP. **Feature Dimensionality Reduction through Genetic Algorithms for Faster Speaker Recognition**. *Proceedings of Eusipco 2008*, Suiza, Agosto 2008.
- Zamalloa M., Rodriguez L.J., Penagarikano M., Bordel G., Uribe JP. **Increasing Robustness to Acoustically Uncovered Signals through Shallow Source Modelling in Speaker Verification**. *Proceedings of Eusipco 2008*. Laussane, Suiza, Agosto 2008.
- Zamalloa M., Rodríguez-Fuentes LJ., Penagarikano M., Bordel G., Uribe JP. **Comparing Genetic Algorithms to Principal Component Analysis and Linear Discriminant Analysis in Reducing Feature Dimensionality for Speaker Recognition**. *Proceedings of ACM GECCO 2008*, Atlanta, USA, Julio 2008.
- Zamalloa M., Rodriguez L.J., Penagarikano M., Bordel G., Uribe JP. **Improving Robustness in Open Set Speaker Identification by Shallow Source Modelling**. *Proceedings of Speaker Odyssey 2008*, Stellenbosch, Sudáfrica, Enero 2008.

- Zamalloa M., Bordel G., Rodriguez LJ., Penagarikano M., Uribe JP. **Selección y Pesado de Parámetros Acústicos mediante Algoritmos Genéticos para el Reconocimiento de Locutor**. *Proceedings of IVJTH 2006*, Noviembre 2006.
- Zamalloa M., Bordel G., Rodriguez LJ., Penagarikano M., Uribe JP. **Using Genetic Algorithms to Weight Acoustic Features for Speaker Recognition**. *Proceedings of ICSLP 2006*. Pittsburgh, Pensilvania, Septiembre 2006.
- Zamalloa M., Bordel G., Rodríguez LJ., Penagarikano M. **Feature Selection based on Genetic Algorithms for Speaker Recognition**. *Proceedings of IEEE Speaker Odyssey 2006*, San Juan, Puerto Rico, Junio 2006.

Como se puede observar, se han realizado aportaciones en todas las líneas de trabajo que se han seguido en la tesis.

Por ultimo, cabe destacar la participación en los siguientes comités internacionales de revision de contribuciones:

- *Review Committee of IEEE International Conference on Acoustics, Speech and Signal Processing 2009*, *ICASSP'10*, Marzo 2010.
- *Review Committee of IEEE International Conference on Acoustics, Speech and Signal Processing 2009*, *ICASSP'09*, Abril 2009.
- *Review Committee of IEEE International Conference on Acoustics, Speech and Signal Processing 2009*, *ICASSP'08*, Abril 2008.
- *Technical Review Comittee of IEEE International Conference of the IEEE Workshop on Automatic Speech Recognition and Understanding*, *ASRU'09*, Diciembre 2009.

6.2 Futuras líneas de trabajo

Para dar continuidad al trabajo realizado en esta tesis, a continuación se sugieren algunas líneas de actuación:

- La comparación de los GA y las transformaciones lineales como metodologías de reducción de dimensionalidad, sugiere su combinación para así aprovechar los puntos fuertes de ambas tecnologías. Se podría diseñar un GA orientado a la búsqueda de la matriz de transformación que maximice la tasa de reconocimiento de locutores o realizar la selección con GA de un espacio transformado. Por otro lado, sería también interesante analizar el rendimiento de estas mismas estrategias sobre una base de datos estándar más amplia, que incluya una mayor diversidad en cuanto a la cantidad de sesiones por locutor y de canales de grabación, como por ejemplo los corpus de las evaluaciones del NIST.
- Sería interesante analizar el rendimiento de metodologías de modelado que no empleen MFCC como representación acústica, sino parámetros de alto nivel, tales como fonemas, palabras, patrones de pronunciación, etc. Entre estas, se encuentran los modelos consistentes en SVM que

analizan las frecuencias relativas de unigramas, bigramas y trigramas de palabras o fonemas extraídos mediante un reconocedor del habla. Debido a que, supuestamente, la correlación de sistemas con distintos tipos de parámetros es menor, y en consecuencia se complementan mejor en la fusión, la fusión con un sistema que emplee parámetros de alto nivel permitiría mejorar el rendimiento de los sistemas basados en MFCC.

- Ha quedado pendiente probar la fusión con información auxiliar adicional, tarea para la cual se podrían definir subconjuntos de experimentos según el idioma de los segmentos de entrenamiento y test, el género del locutor objetivo, el canal de grabación de los segmentos, etc.
- Sería interesante analizar el comportamiento del modelo superficial de fuente (SSM) sobre los conjuntos experimentales del NIST. SSM se podría aplicar, combinado linealmente con el UBM, o como sistema adicional en la fusión para mejorar el rendimiento del sistema.
- Debido a que los corpus definidos para las evaluaciones del NIST incluyen señales grabadas a través de diferentes canales telefónicos y a través de diversos terminales, sería beneficioso estudiar la aplicación de alguna metodología de compensación de canales. De las metodologías propuestas en los últimos años, destacan JFA, NAP y la adaptación por auto-canales. Se podría también analizar de forma exhaustiva el comportamiento de HLDA, aplicado en Albayzín y Dihana, sobre los conjuntos experimentales del NIST.
- En lo que respecta al sistema de seguimiento de locutores se podría utilizar un *back-end* de detección de locutores más robusto, basado en modelos SVM-GLDS o GMM-SVM. Se podría realizar un estudio que, por un lado, permitiría comparar el rendimiento del sistema con diferentes *back-ends* de detección, y por otro, posibilitaría la introducción de un módulo de fusión de los sistemas de detección.
- Entre las posibles mejoras del sistema de seguimiento, se encuentra la incorporación de un módulo de segmentación acústica de baja latencia, y la modificación de la configuración del sistema a un sistema auto-suficiente de administración-cero, que posea la capacidad de aumentar y disminuir el conjunto de locutores objetivo automáticamente. Por otro lado, con objeto de ampliar la funcionalidad del sistema, se podría modificar su diseño de forma que, en base a la localización de las señales de audio detectadas en el entorno, pueda estimar (posiblemente de forma relativa) la localización de los locutores. La presencia y localización son datos contextuales muy importantes en un entorno AmI, ya que normalmente la información requerida por el usuario depende de su posición instantánea.

Capítulo 7

Bibliografía

[Aarts09] Aarts E., Ruyter B. New Research Perspectives on Ambient Intelligence. Journal of Ambient Intelligence and Smart Environment, vol. 1, pp. 5-14, 2009.

[Abowd05] Abowd G.D., Mynatt E.D. Designing for the Human Experience in Smart Environments. In D.J. Cook and S.K. Das, editors, Smart Environments: Technology, Protocols, and Applications, pp. 153-174, Wiley, 2005.

[Ajmera04] Ajmera J., McCowan I., Bourlard H. Robust Speaker Change Detection. *IEEE Signal Processing Letters*, vol 11, no 8, pp 649-651, 2004.

[Alcocer04] Alcocer N., Castro M., Galiano I. Granel R., Grau S., Griol D. Adquisición de un Corpus de Diálogo: DIHANA. *III Jornadas en Tecnologías del Habla*, pp. 131-134, Valencia, Noviembre 2004.

[AMI] AMI Consortium: http://www.amiproject.org

[AMI Corpus] The AMI Corpus: http://corpus.amiproject.org/

[Amigo] Amigo: Ambient Intelligence for the networked home environment. http://www.hitech-projects.com/euprojects/amigo/

[Anguera06] Anguera X. Robust Speaker Diarization for Meetings. *PhD Thesis dissertation*. Octubre 2006.

[Aronowitz05] Aronowitz H., Burshtein D., Amihood A. A Session-GMM Generative Model Using Test-Utterance Gaussian Mixture Modelling for Speaker Verification. *Proceedings of IEEE ICASSP 2005*, vol 1, pp 733-736, 2005.

[Arsajani04] Arsanjani A. Service Oriented Modelling and Architecture: http://www.ibm.com/developerworks/library/ws-soa-design1/

[Atal76] Atal B.S. Automatic Recognition of Speakers from their Voices, *in Proceedings of the IEEE*, vol. 63, nº 4, pp. 460-475, 1976.

[Auckenthaler00] Auckenthaler R., Carey M., Lloyd Thomas H. Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, vol 10, pp 42-54, 2000.

[AwareHome] Aware Home Research Initiative: http://awarehome.imtc.gatech.edu/

[Barras06] Barras B., Zhu X., Meignier S., Gauvain J.L. Multi Stage Speaker Diarization of Broadcast News. IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, nº5, pp. 1505-1512, Septiembre 2006.

[Ben02] Ben M., Blouet R., Bimbot F. A Monte-Carlo Method for Score Normalization in Automatic Speaker Verification using Kullback-Leibler distances. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'02, vol. 1, pp. 689-692, Orlando, USA, Mayo 2002.

[Ben04] Ben M., Betser M., Bimbot F., Gravier G. Speaker Diarization using bottm-up Clustering based on a Parameter-derived Distance between adapted GMMs. In *Proceedings of the International Conference on Signal and Language Processing*, *INTERSPEECH 2004*, pp. 2329-2332, Corea, Octubre 2004.

[Bernardin07] Bernardin K., Stiefelhagen R. Audio-Visual Multi-Person Tracking and Identification for Smart Environments. In Proceedings of the 15th International Conference on Multimedia, MM'07, pp. 661-670, Bavaria, Alemania, Septiembre 2007.

[Bimbot04] Bimbot F., Bonastre JF., Fredouille C., Gravier G., Magrin-Chagnolleau I., Meignier S., Merlin T., Ortega-Garcia J., Petrovska-Delacrétaz D., Reynolds DA. A tutorial on Text-Independent Speaker Verication. *Eurasip Journal on Applied Signal Processing*, vol 4, pp 430-451, 2004.

[Boakye04] Boakye, K., Peskin, B. Text-Constrained Speaker Recognition on a Text-Independent Task. *In Proceedings Odyssey-04 Speaker and Language Recognition Workshop*, Toledo, Spain ,2004.

[Bonastre00] Bonastre J.F., Delacourt P., Fredouille F., Merlin T., Wellekens C. A Speaker Tracking System based on Speaker Turn Detection for Nist Evaluation. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2000.

[Bordel09] Bordel G., Diez M., Landera I., Nieto S., Penagarikano M., Rodriguez-Fuentes L.J., Varona A., Zamalloa M. Hearch: A Multilingual Spoken Document Retrieval, IEEE Automatic Speech Recognition and Understanding Workshop, ASRU09, Italia, Diciembre 2009.

[Brandschain08] Brandschain L., Cieri C., Graff D., Neely A., Walker K. Speaker Recognition: Building the Mixer4 and Mixer5 Corpora. *In Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC'08*, Marruecos, Mayo 2008.

[Brummer06] Brummer N., du Preez J. Application-independent Evaluation of Speaker Detection. *Computer Speech and Language*, vol 20, pp 230-275, 2006.

[Brummer07] Brummer N., Burget L., Cernocky H., Glembek O, Grezl F., Karafiat M., van Leeuwen DA., Matejka P., Schwarz P., Strasheim A. Fusion of Heterogeneus Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. *IEEE Transactions on Audio, Speech and Language Processing*, vol 15, no 7, 2007.

[Burges98] Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, vol 2, pp 121-168, 1998.

[Burget04a] Buget L. Combination of Speech Features using Smoothed Heteroscedastic Linear Discriminant Analysis, *Proceedings of ICSLP 2004*, Jeju Island, Korea, Octubre 2004.

[Burget04b] Buget L. Complementarity of Speech Recognition Systems and System Combination, *PhD dissertation*, Brno University of Technology, September 2004.

[Burget07] Burget L., Matejka P., Schwarz P., Glembek O. Cernocky JH. Analysis of Feature Extraction and Channel Compensation in a GMM Speaker Recognition System. *IEEE Transactions on Audio, Speech, and Language Processing*, vol 17, no 7, pp 1979-1986, 2007.

[Busso05] Busso C., Hernanz S., Chu C.W., Kwon S.I., Lee S., Georgiou G., Cohen I., Narayanan S. Smart Room: Participant and Speaker Localization and Identification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'05, vol. 2, pp. 117-120, Philadelphia, USA, Abril 2005.

[Carey91] Carey M., Parris E., Bridle J. A Speaker Verification System using Alpha-nets, *in Proceedings of International Conference on Acoustics, Speech and Signal Processing* (ICASSP'91), vol. 1, pp. 397-400, Toronto, Canada, Mayo 1991.

[Casacuberta91] Casacuberta F., García R., Llisterri J., Nadeu C., Pardo JM., Rubio A. Development of Spanish Corpora for Speech Research (Albayzín). *in G. Castagneri Ed., Proceedings of the Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assessment Methods*, pp. 26-28, Septiembre 1991.

[Chang01] Chang CC., Lin CJ. LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[Charbuillet06] Charbuillet C., GA B., Chetouani M., Zarader J.L. Filter Bank Design for Speaker Diarization based on Genetic Algorithms. In Proceedings of the IEEE ICASSP'06, Toulouse, France, 2006.

[Charbuillet07] Charbuillet C., Bruno G. Chetouani M., Zarader J.Z. Multi Filter Bank Approach for Speaker Verification based on Genetic Algorithm. *Proceedings of NOLISP 2007*, pp 105-113, 2007.

[Charlet97] Charlet D., Jouvet D. Optimizing Feature Sets for Speaker Verification. *Pattern Recognition Letters*, vol 18, no 9, pp 873-879, 1997.

[Chen98] Chen SS., Gopalakrishnan PS. Speaker Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[CHILL] Computers in the Human Interaction Loop: http://chil.server.de/

[Cieri06] Cieri C., Andrew W., Campbell JP., Doddington G., Godfrey J., Huang S., Libermann M., Martin A., Nakasone H., Przybocki M., Walter K. The Mixer and Transcript Reading Corpora: Resources for Multilingual Crosschannel Speaker Recognition Research. *In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC'06*, Genoa, Italia, 2006.

[CLEAR] CLEAR 2007, Classification of Events, Activities and Relationships Evaluation and Workshop: http://www.clear-evaluation.org/

[Collet05] Collet M., Charlet D., Bimbot F. A Correlation Metric for Speaker Tracking using Anchor Models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 713-716, Philadelphia, USA, Mayo 2005.

[Collet06] Collet M., Charlet D., Bimbot F. Speaker Tracking by Anchor Models using Speaker Segment Cluster Information. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 1009-1012, Toulouse, Francia, Mayo 2006.

[Cook09] Cook D.J., Augusto J.C., Jakkula V.R. Ambient Intelligence: Technologies, Applications, and Opportunities. Pervasive and Mobile Computing, vol. 5, no. 4, pp. 277-298, Agosto 2009.

[Courant53] Courant R., Hilbert D. Methods of Mathematical Physics. *Interscience*, 1953.

[Cristianini00] Christianini N., Shame-Taylor J. Support Vector Machines. *Cambridge, UK: Cambridge University Press*, 2000.

[Davis80] Davis SB., Mermelstein P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol 28, no 4, pp 357-366, 1980.

[Dalmasso09] Dalmasso E., Castaldo F., Laface P., Coibro D., Vair C. Loquendo-Politecnico di Torino's 2008 NIST Speaker Recognition Evaluation System. *In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP'09*, pp. 4213-4216, Taiwan, Abril 2009.

[Delacourt00] Delacourt P., Wellekens CJ. DISTBIC: A Speaker-based Segmentation for Audio Data Indexing. *Speech Communication*, vol 22, pp 111-126, 2000.

[Demirekler99] Demirekler M., Haydar A. Feature Selection using a Genetics-based Algorithm and its Application to Speaker Identification. *Proceedings of the IEEE ICASS 1999*, pp 329-332, Phoenix, Arizona, 1999.

[Dempster77] Dempster AP., Laird NM., Rubin DB. Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society- Series B*, vol. 39, no 1, pp 1-38, 1977.

[Dey01] Dey A. K. Understanding and Using Context. *Personal and Ubiquitous Computing, Springer-Verlag,* vol. 5, no. 1, pp. 4-7, 2001.

[DET] DET Curves: http://www.nist.gov/speech/tools/DETware v2.1.targz.htm

[Doddington85] Doddington G.R. Speaker Recognition – Identifying people by their voices. *In Proceedings of the IEEE*, vol. 73, pp. 1651-1664, 1985.

[Doddington00] Doddington GR., Przybocki MA., Martin AF., Reynolds DA. The NIST Speaker Recognition Evaluation – Overview, Methodology, Systems, Results, Perspective. *Speech Communication*, no. 31, pp. 225-254, 2000.

[Doddington01] Doddington G.R. Speaker Recognition based on Idiolectal differences between Speakers. *In Proceedings of Eurospeech 2001*, pp. 2521-2524, 2001.

[Duda00] Duda RO., Hart PE., Stork DG. Pattern Classification (2nd Edition), *Wiley Interscience*, 2000.

[Dunn] Dunn J.S., Podio F. Biometrics Consortium Website: http://www.biometrics.org.

[Dunn00] Dunn R.B., Reynolds D.A., Quatieri T.F. Approaches to Speaker Detection andTracking in Conversational Speech, Digital Signal Processing 10, pp. 93-112, 2000.

[ECJ] ECJ 16: http://cs.gmu.edu/~eclab/projects/ecj/

[ELRA] ELRA: http://www.elra.info/

[Errity07] Errity A., McKenna J. A Comparative Study of Linear and Nonlinear Dimensionality Reduction for Speaker Identification. Proceedings of IEEE Intenational Conference on Digital Signal Processing, 2007.

[ESTER] ESTER Evaluation Campaign: http://www.afcp-parole.org/ester/index.html

[EVALITA] Evaluation of NLP and Speech Tools for Italian: http://evalita.fbk.eu/

[Fant70] Fant G. Acoustic Theory of Speech Production, Mouton, The Hague, The Netherlands, 1970.

[Faundez-Zanuy05] Faundez-Zanuy M., Monte-Moreno E. State-of-the-Art in Speaker Recognition. *IEEE Aeroespace and Electronics System Magazine*, vol 7, no 5, pp 7-12, 2005.

[Ferrer06] Ferrer L., Shriberg E., Kajarekar S., Stockle A., Sönmez K., Venkataraman A., Bratt H. The Contribution of Cepstral and Stylistic Features to SRI's 2005 NIST Speaker Recognition Evaluation System. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'06*, vol. 1, pp. 101-104, Toulouse, Francia, Mayo 2006.

[Ferrer07] Ferrer L., Shriberg E., Kajarekar S. Sönmez K. Parameterization of Prosodic Feature Distributions for SVM Modeling in Speaker Recognition. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'07*, Hawai, USA, 2007.

[Ferrer08a] Ferrer L., Sönmez K., Shriberg E. An Anticorrelation Kernel for Improved System Combination in Speaker Verification. *In Proceedings of Odyssey 2008: The Speaker and Language Recognition Workshop*, Stellenbosch, Sudáfrica, Enero 2008.

[Ferrer08b] Ferrer L., Graciarena M., Zymnis A., Shriberg E. System Combination using Auxiliary Information for Speaker Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'08, pp. 4853-4856, Las Vegas, USA, Marzo 2008.

[Finkenzeller03] Finkenzeller K. RFID Handbook: Fundamentals and Applications in Contactless Smart Cards and Identification. *Wiley, 2nd Edition*, 2003.

[Fisher36] Fisher R.A. The use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, vol. 7, pp. 179-188, 1936.

[Fletcher87] Fletcher R. Practical Methods of Optimization. *John Wiley and Sons, Inc., 2nd Edition*, 1987.

[FoCal] Brummer N. Focal toolkit: http://niko.brummer.googlepages.com/jfocal

[Fukunaga90] Fukunaga K. Introduction to Statistical Pattern Recognition. *Elsevier, Nueva York, 2º Edición*, 1990.

[Furui81] Furui S. Comparison of Speaker Recognition Methods using Static Features and Dynamic Features. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 342-350, 1981.

[Gales99] Gales M.J.F. Semi-tied covariance matrices for hidden markov models. *IEEE Trans. Speech and Audio Processing*,vol. 7, pp. 272–281, 1999.

[Galliano05] Galliano S., Geoffrois E., Mostefa D., Choukri K., Bonastre JF., Gravier G. The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In Proceedings of the International Conference on Spoken and Language Processing, INTERSPEECH 2005, pp. 1-4, Septiembre 2005.

[Gatica-Perez07] Gatica-Perez D., Lathoud G., Odobez J., McCowan I. Audiovisual Probabilistic Tracking of Multiple Speakers in Meetings. IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 2, Febrero 2007.

[Gauvain94] Gauvain JC., Lee CH. Maximum a posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transcations on Speech and Audio Processing*, vol 2, no 2, 1994.

[Gauvain95] Gauvain, J.L., Lamel, L.F., Prouts, B. Experiments with Speaker Verification Over the Telephone. *In Pardo, J.M., Enriquez, E., Ortega, J., Ferreiros, J., Macias, J., Valverde, F.J. (eds.) Proc. EUROSPEECH*, Madrid, 1995.

[Garcia-Romero04] García-Romero D., Fierrez-Aguilar J., Gonzalez-Rodriguez J., Ortega-Garcia J. On the Use of Quality Measures for Text-Independent Speaker Recognition, In Proceedings of the Speaker and Language Recognition Workshop, Toledo, Mayo 2004.

[Gish94] Gish H., Schmidt M. Text-independent Speaker Identification. *IEEE Signal Processing Magazine*, vol 11, no 4, 1994.

[Glover06] Glover B., Bhatt H. RFID Essentials. *O'Reilly*, 2006.

[Godfrey94] Godfrey J., Graff D., Martin A. Public Databases for Speaker Recognition and Verification. *ESCA Workshop on Automatic Speaker Recogntion, Identification, and Verification*, Martigny, Suiza, Abril 1994.

[Goldberg89] Goldberg GE. Genetic Algorithms in Search, Optimization and Machine Learning, *Addison-Wesley*, 1989.

[Guo09] Guo W., Long Y., Li Y., Pan L., Wang E., Dai L., iFLY System for the NIST 2008 Speaker Recognition Evaluatio. *In Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP'09*, pp. 4209-4212, Taiwan, Abril 2009.

[Hagras04] Hagras H., Callaghan V., Colley M., Clarke G., Pounds-Cornish A., Duman H. Creating an Ambient-Intelligence Environment using Embedded Agents. IEEE Intelligent Systems, vol. 19, no. 6, pp. 12-20, 2004.

[Hansmann01] Hansmann U., Merk L., Nicklous M., Stober T. *Pervasive Computing Handbook*. Springer-Verlag, 2001.

[Hautamäki08] Hautamäki V., Kinnunen T., Kärkkäinen I., Tuononen M., Saastamoinen J., Fränti P. Maximum a Posteriori Estimation of the Centroid Model for Speaker Verification. IEEE Signal Processing Letters, vol 15, pp. 162-165.

[Hébert08] Hébert M. Text-dependent Speaker Recognition. *Springer Handbook of Speech Processing*, Benesty, Sondhi, Huang Edt., Springer 2008, pp. 743-763.

[Helal05] Helal S., Mann W., El-Zabadani H., King J., Kaddoura Y., Jansen E. The Gator Tech Smart House: A Programmable Pervasive Space, IEEE Computer, no. 38, vol. 3, pp. 50-60, 2005.

[Hermansky94] Hermanksy H., Morgan N. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*,vol. 2, no. 4, pp. 578–589, 1994.

[Hermansky90] Hermansky H. Perceptual Linear Prediction (PLP) Analysis for Speech. *JASA*. Pp 1738-1752, 1990.

[Higgins91] Higgins A., Bahler L., Porter J. Speaker Verification using Randomized Phrase Prompting, *Digital Signal Processing*, vol. 10, no 2, pp. 89-106, 1991.

[Higgins93] Higgins A., Bhaler L., Porter J. Voice Identification using Nearest Neighbor Distance Measure. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 375-378, Tokyo, Japon, 1993.

[Holland75] Holland J.H. Adaptation in natural and artificial systems. *University of Michigan Press*, 1975 (reprinted in 1992 by MIT Press, Cambridge, MA).

[Hong05] Hong QY., Kwong S. A Genetic Classification Method for Speaker Recognition. *Engineering Applictions for Artificial Intelligence*, vol 18, no 1, pp 13-19, 2005.

[HomeLab] HomeLab, Our Testing Ground for a Better Tomorrow. http://www.research.philips.com/technologies/projects/homelab/

[Hsu03] Hsu CN., Yu HC., Yang BH. Speaker Verification without Background Speaker Models. *Proceedings of IEEE ICASSP 2003*, vol 2, pp 233-236, 2003.

[Huang01] Huang X., Acero A., Hsiao-Wuen H. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001.

[iButtons] I Button: http://www.maxim-ic.com/products/ibutton/ibuttons/

[ICSI] The ICSI Meeting Corpus: http://www.icsi.berkeley.edu/Speech/mr/

[Ince07] Ince N.F., Min C.H., Tewfik A.H. In-Home Assistive System for Traumatic Brain Injury Patients. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'07, vol. 2, pp. 565-568, Honolulu, Hawai, Abril 2007.

[Intille06] Intille S.S., Larson K., Munguia-Tapia E., Beaudin J.S., Kaushik P., Nawyn J., Rockinson R. Using a Live-in Laboratory for Ubiquitous Computing Research, In Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. eds. PERVASIVE 2006. LNCS, vol. 3968, pp. 349–365. Springer, Heidelberg, 2006.

[ISTAG99] Weyrich C. Orientations for Work Programme 2000 and beyond. *European Commission Report*, 1999

[ISTAG01] Ducatel K., Bogdanowicz M., Scapolo F., Leijten J., Burgelman J.C. Scenarios for Ambient Intelligence in 2010, *European Commission Report*, 2001.

[ISTAG03] ISTAG. Ambient Intelligence: from vision to reality, *European Commission Report*, 2003.

[Istrate05] Istrate D., Scheffer N., Fredouille N., Bonastre J.F. Broadcast News Speaker Tracking for ESTER 2005 Campaign. *Proceedings of Interspeech'05*, 2005.

[Jaakkola98] Jaakkola T.S., Haussler D. Exploiting Generative Models in Discriminative Classifiers, *In Advances in Neural Information Processing 11*, Editorial M.S. Kearns, S.A. Solla, D.A. Cohn (MIT Press, Cambridge), pp. 487-193, 1998.

[Jain00] Jain AK., Duin RP., Mao J. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 22, no 1, pp 4-37, 2000.

[Jang01] Jang G.J., Lee T.W., Oh J.H. Learning Statistically Efficient Features for Speaker Recognition. Proceedings of IEEE International Conference on Acoustics, Signal and Speech Processing (ICASSP'01), 2001.

[Jin97] Jin H., Kubala F., Schwartz R. Automatic Speaker Clustering. In Proceedings of DARPA Speech Recognition Workshop, Chantilly, VA, Febrero 1997.

[Jin00] Jin Q., Waibel A. Application of LDA in Speaker Recognition. Proceedings of International Conference on Speech and Language Processing (ICSLP'00), 2000.

[Jini] Jini.org http://www.jini.org

[Johnson98] Johnson SE., Woodland PC. Speaker Clustering using Direct Maximization of the MLLR-adapted Likelihood. *In Proceedings of the International Conference on Spoken and Language Processing, ICSLP-98*, Australia, Noviembre 1998.

[Jolliffe02] Jolliffe IT. Principal Component Analysis (Second Edition). *Springer*, 2002.

[JPCampbell97] Campbell J.P. Speaker Recognition: A Tutorial. *Proceedings of the IEEE*, vol. 85, no. 9, pp 1437-1462, Septiembre 1997.

[JPCampbell99] Campbell JP., Reynolds DA. Corpora for the Evaluation of Speaker Recognition Systems. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'99*, vol. 2, pp. 829-832, Phoenix, USA, Marzo 1999.

[JPCampbell03] Campbell J.P., Reynolds D.A., Dunn R.B. Fusing High-level and Low-level features for Speaker Recognition. *In Proceedings of Eurospeech 2003*, Geneva, Italia, 2003.

[JPCampbell04] Campbell JP., Nakasone H., Cieri C., Miller D., Walker K., Martin AF., Przybocki MA. The MMSR Bilingual and Crosschannel Corpora for Speaker Recognition. *In Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC'04*, Lisboa, Portugal, 2004.

[Kaila09] Kaila L., Hyvönen J., Ritala M., Mäkinen Ville, Vanhala J. Development of a Location-aware Speech Control and Audio Feedback System. In Proceedings of the IEEE International Conference on Pervasive Computing and Communications, PERCOM'09, pp. 1-4, Texas, USA, Marzo 2009.

[Kajarekar09] Kajarekar S.S., Scheffer N., Graciarena M., Shriberg E., Stolcke A., Ferrer L., Bocklet T. The SRI NIST 2008 Speaker Recognition Evaluation System. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'09, pp.4205-4208, Taiwan, Abril 2009.

[Kemp00] Kemp T., Schmidt M, Westphal M., Waibel A. Strategies for Automatic Segmentation of Audio Data. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 3, pp. 1423-1426, 2000.

[Kenny04] Kenny P., Dumouchel P. Disentangling Speaker and Channel Effects in Speaker Verification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'04, vol. 1, pp. 37-40, Mayo 2004.

[Kenny08] Kenny P. Ouellet P. Dehak N., Gupta V., Dumouchel P. A Study of Interspeaker Variability in Speaker Verification. IEEE Transactions on Audio, Speech and Language Processing, vol. 16, pp. 980-988, Julio 2008.

[Kinnunen01] Kinnunen T., Franti P. Speaker Discriminative Weighting Method for VQ-based Speaker Identification. *Proceedings of Audio and Video Based Biometric Authentication 2001*, pp 150-156, 2001.

[Kinnunen06] Kinnunen T., Karpov E., Franti P. Real-time Speaker Identification and Verification. *IEEE Transactions on Audio, Speech and Language Processing*, vol 14, no 1, 2006.

[Kinnunen09] Kinnunen T., Saastamoinen J., Hautamäki V., Vinni M., Fränti P. Comparative Evaluation of Maximum A Posteriori Vector Quantization and Gaussian Mixture Models in Speaker Verification, Pattern Recognition Letters, vol. 30, pp. 341-347, 2009.

[Kotti08] Kotti M., Moschou V., Kotropoulos C. Speaker Segmentation and Clustering. *Signal Processing*, vol 88, pp 1091-1124, 2008.

[Krum00] Krum J., Harris S., Meyers B., Brummit B., Hale M., Shafer S. Multi-Camera Multi-Person Tracking for EasyLiving. In Proceedings of the third IEEE International Workshop on Visual Surveillance, VS'2000, pp. 3-10, Dublin, Irlanda, 2000.

[Kumar97] Kumar N. Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition. *Ph.D. dissertation, Johns Hopkins University*, Baltimore, 1997.

[Kwon05] Kwon S., Narayanan S. Unsupervised Speaker Indexing using Generic Models. *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, Septiembre 2005.

[LDC] Linguistic Data Consortium: http://www.ldc.upenn.edu/.

[Le07] Le VB., Mella O., Fohr D. Speaker Diarization using Normalized Cross Likelihood Ratio. *In Proceedings of the International Conference on Spoken and Language Processing, INTERSPEECH 2007,* pp. 1869-1872, Bélgica, Agosto 2007.

[Lima02] Lima C.B., Alcaim A., Apolinário J.A. On the use of PCA in a GMM and Ar-Vector models for text-independent speaker verification. Proceedings of IEEE International Conference on Digital Signal Processing, 2002.

[Li88] Li K.P., Porter J.E. Normalizations and Selection of Speech Segments for Speaker Recognition Scoring. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP'88, vol. 1, pp. 595-598, Nueva York, USA, Abril 1998.

[Linde80] Linde Y., Buzo A. Gray RM. An Algorithm for Vector Quantizer Design, *IEEE Transactions on Communications*, vol 28, no 1, pp 84-95, 1980.

[Lippmann93] Lippmann RP., Kukolich L., Singer E. LNKnet: Neural Network, Machine Learning and Statistical Software for Pattern Classification. *Lincoln Laboratory Journal*, vol 6, no 2, pp 249-268, 1993.

[Liu04] Liu D., Kubala F. Online Speaker Clustering. *Proceedings of IEEE ICASSP 2004*, Montreal, Canada, Mayo 2004.

[Liu05] Liu D., Kiecza D., Kubala F. Online Speaker Adaptation and Tracking for Real-time Speech Recognition. *In Proceedings of the International Conference on Spoken and Language Processing, INTERSPEECH 2005*, Portugal, Septiembre 2005.

[Lu05] Lu L., Zhang HJ. Unsupervised Speaker Segmentation and Tracking in Real-time Audio Content Analysis. *Multimedia Systems*, vol 10, pp 332-343, 2005.

[Luque06] Luque J., Morros R., Anguita J., Farrus M., Macho D., Marqués F., Martínez C., Vilaplana V., Hernando J. Multimodal Person Identification in a Smart Room. *Proceedings of IVJTH 2006*, Zaragoza, Noviembre 2006.

[Luque07] Luque J., Morros R., Garde A., Anguita J., Farrus M., Macho D., Marqués F., Martínez C., Vilaplana V., Hernando J. Audio, Video and Multimodal Person Identification in a Smart Room. In Stiefelhagen R. and Garofolo J. eds. CLEAR 2006, LNCS 4122, pp. 258-269, Springer-Verlag Berlin Heidelberg, 2007.

[Magrin-Chagnolleau99] Magrin-Chagnolleau I., Rosenberg AE., Parthasarathy S. Detection of Target Speakers in Audio Databases. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'99*, vol. 2, pp. 821-824, Phoenix, USA, Mayo 1999.

[Malegaonkar07] Malegaonkar S.A., Ariyaeenia A.M., Sivakumaran P. Efficient Speaker Change Detection using Adapted Gaussian Mixture Models. *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, nº6, August 2007.

[Markov07] Markov K., Nakamura S. Never-Ending Learning System for On-line Speaker Diarization. In Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU'07, pp. 699-704, Diciembre 2007.

[Martin97] Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M. The DET Curve in the Assessment of Detection Task Performance. *Proceedings of Eurospeech 1997*, Rhodes, Grecia, pp 1895-1898, 1997.

[Martin01] Martin AF., Przybocki M. Speaker Recognition in a Multi-Speaker Environment. *Proceedings of European Conference on Speech Communication Technologies*, Aalborg, Dinamarca, Septiembre 2001.

[Martin07] Martin AF. Evaluations of Automatic Speaker Classification Systems. *Lecture Notes in Artificial Intelligence, Springer-Verlag*, 2007.

[Martin09] Martin A.F., Greenberg C.S. NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels. *In Proceedings of the International Conference on Spoken and Language Processing, INSTERSPEECH 2009*, pp. 2579-2582, Brighton, UK, Septiembre 2009.

[Mary06] Mary L, Yegnanarayana B. Prosodic Features for Speaker Verification. *In Proceedings of the International Conference on Spoken and Language Processing, INSTERSPEECH 2006*, Pittsburgh, Pensilvania, Septiembre 2006.

[Mayoue09] Mayoue A., Dorizzi B., Allano L., Chollet G., Hennebert J., Petrovska-Delacretaz D., Verdet F. BioSecure Multimodal Evaluation Campaign 2007. Guide to Biometric Reference Systems and Performance Evaluation, Springer London, pp. 327-372, Abril 2009.

[McLaren09] McLaren M., Vogt R., Baker B., Sridharan S. Improved GMM-based Speaker Verification using SVM-driven Impostor Dataset Selection. *In Proceedings of the International Conference on Spoken and Language Processing, INTERSPEEH 2009*, pp. 1267-1270, Brighton, UK, Septiembre 2009.

[MeetingRoom] Meeting Room: http://www.is.cs.cmu.edu/meeting_room/

[Meignier04] Meignier S., Moraru D., Fredouille C., Besacier L., Bonastre JF. Benefits of Prior Acoustic Segmentation for Automatic Speaker Segmentation. *Proceedings of ICASSP 2004*, Montreal, Canada, Mayo 2004.

[Meignier06] Meignier S., Moraru D., Fredouille C., Bonastre JF., Besacier L. Step-by-step and Integrated Approaches in Broadcast News Speaker Diarization. *Computer Speech and Language*, vol 20, pp 303-330, 2006.

[Moraru05] Moraru D., Ben M., Gravier G. Experiments on Speaker Tracking and Segmentation in Radio Broadcast News. *Proceedings of Interspeech'05*, 2005.

[Mori01] Mori K., Nakawaga S. Speaker Change Detection and Speaker Clustering using VQ Distortion for Broadcast News Speech Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'01, vol. 1, pp. 413-416, Utah, 2001.

[Mostefa07] Mostefa D., Moreau N., Choukri K., Potamianos G., Chu SM., Tyagi A., Casas JR., Turmo J., Cristoforetti L., Tobia F., Pnevmatikakis A., Mylonakis V., Talantzis F., Burger S., Stiefelhagen R., Bernardin K., Rochet C. The CHIL audiovisual corpus for lecture and meeting analysis inside smart romos. *Lang Resources & Evaluation*, vol. 41, pp. 389–407, 2007.

[Nakawaga93] Nakawaga S., Suzuki H. A new Speech Recogntion Method based on VQ-distortion measure and HMM. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'93*, pp. 676-679, 1993.

[Newman96] Newman N., Gillick L., Ito Y. McAllaster D., Peskin B. Speaker Verification through Large Vocabulary Continuous Speech Recognition. In Proceedings of ICSLP 1996, vol. 4, pp. 2419-2422, Philadelphia, USA, Octubre 1996,

[NIST SRE] NIST SRE home: http://www.nist.gov/speech/tests/sre/

[NIST Meeting] NIST Pilot Meeting Corpus: http://www.nist.gov/speech/test_beds/

[NIST RT] NIST Rich Transcription Evaluation Project: http://www.itl.nist.gov/iad/mig/tests/rt/

[Papazoglou06] Papazoglou MP., Traverso P., Dustdar S., Leymann F., Kramer BJ. Service-oriented computing: A research roadmap. *In Francisco Cubera, Bernd J. Kramer, and Michael P. Papazoglou,editors, Service Oriented Computing (SOC); Dagstuhl Seminar Proceedings*. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Alemania, 2006.

[Papazoglou07] Papazoglou MP., Traverso P., Dustdar S., Leymann F. Service-oriented Computing: State of the Art and Research Challenges. *IEEE Computer Society*, November 2007.

[Patané01] Patané G., Russo M. The Enhanced LBG Algorithm. *Neural Network Archive*, vol 14, no 9, pp 1219-1237, 2001.

[Oliveira03] Oliveira L.S., Sabourin R., Bortolozzi F., Suen CY. A Methodology for Feature Selection using Mutilobjective Genetic Algorithms for Handwritten Digit String Recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, vol 17, no 6, pp 903-929, 2003.

[Orr00] Orr R.J., Abowd G.D. Smart Floor: A Mechanism for Natural User Identification and Trackin. In Proceedings of the Conference on Human Factors in Computing Systems (CHI 2000), pp. 275-276, The Hague, Holanda, Abril 2000.

[OSGi] OSGi, The Dynamic Module System for Java: http://www.osgi.org

[Papazoglou07] Papazoglou MP., Traverso P., Dustdar S., Leymann F. Service-oriented Computing: State of the Art and Research Challenges. *IEEE Computer Society*, November 2007.

[Pearson1901] Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, vol. 2, no. 6, pp. 559-572, 1901.

[Pelecanos01] Pelecanos J., Sridharan S. Feature Warping for Robust Speaker Verification. *Proceedings of Speaker Odyssey 2001*. Crete, Grecia, Junio 2001.

[Penagarikano08] Penagarikano M., Bordel G., Bilbao S., Zamalloa M., Rodriguez L.J. Sautrela: un Entorno de Desarrollo Versátil para las Tecnologías del Habla. In *Proceedings of VJTH 2008*, pp. 233-235, Bilbao, Noviembre 2008.

[Penagarikano09] Penagarikano M, Varona A, Zamalloa M., Rodriguez-Fuentes L.J., Bordel G., Uribe J.P. University of the Basque Country +Ikerlan System for NIST 2009 Language Recognition Evaluation. NIST 2009 Language Recognition Evaluation Workshop (véase http://gtts.ehu.es/gtts/NT/fulltext/PenagarikanoLRE09.pdf).

[Potamitis04] Potamitis I., Chen H., Tremoulis G. Tracking of Multiple Moving Speakers with Multiple Microphone Arrays. IEEE Transactions on Speech and Audio Processing, vol. 12, no. 5, pp. 520-529, Septiembre 2004.

[Przybocki99] Przybocki MA., Martin AF. The 1999 Nist Speaker Recognition Evaluation, using Summed Two-Channel Telephone Data for Speaker Detection and Speaker Tracking. *Proceedings of Eurospeech 1999*, Budapest, Hundaria, Septiembre 1999.

[Przybocki04] Przybocki M., Martin A. NIST Speaker Recognition Evaluation Chronicles. *In Proceedings of Odyssey 2004: Speaker and Language Recognition Workshop*, pp. 15-22, Toledo, Junio 2004.

[Przybocki07] Przybocki M., Martin AF., Le AN. NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora – 2004, 2005, 2006. IEEE Transactions on Audio Speech and Language Processing, vol. 15, no. 7, pp. 1951-1959, Septiembre 2007.

[Rabiner93] Rabiner L.R. Fundamentals of Speech Recognition. *Englewood Cliffs, New Jersey, Prentice Hall*, 1993.

[Ramachandran02] Ramachandran R.P., Farrel K.R., Ramachandran R., Mammone R.J. Speaker Recognition – General Classifier Approaches and Data Fusion Methods. Pattern Recognition, vol. 35, pp. 2801-2821, 2002.

[Ramos08] Ramos C., Augusto J.C., Shapiro D. Ambient Intelligence – the Next Step for Artificial Intelligence. IEEE Intelligent Systems, vol. 23, no. 2, pp. 15-18, Marzo 2008.

[Ramos-Castro07] Ramos-Castro D., Fierrez-Aguilar J., Gonzalez-Rodriguez J., Ortega-Garcia J. Speaker Verification Using Speaker- and Test-dependent Fast Score Normalization. *Pattern Recognition Letters*, vol. 28, pp. 90-98, 2007.

[Reynolds94] Reynods DA. Experimental Evaluation of Features for Robust Speaker Identification. *IEEE Transactions on Speech and Audio Processing*, vol 2, no 4, pp 639-643, 1994.

[Reynolds95a] Reynolds DA., Rose RC. Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing*, vol 3, no 1, pp 72-83, 1995

[Reynolds95b] Reynolds DA. Speaker Identification and Verification using Gaussian Mixture Speaker Models. *Speech Communication*, vol 17, pp 91-108, 1995.

[Reynolds97] Reynolds DA. Comparison of Background Normalization Methods for Text-Independent Speaker Verification. *Proceedings of Eurospeech 1997*. Rhodes, Grecia, Septiembre 1997.

[Reynolds98] Reynolds DA., Singer E., Carlson BA., O'Leary GC., McLaughlin JJ., Zissman MA. Blind Clustering of Speech Utterances based on Speaker and Language Characteristics. In Proceedings of the International Conference on Spoken and Language Processing, ICSLP'98, Australia, Diciembre 1998.

[Reynolds00] Reynolds DA., Quatieri TF., Dunn RB. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, vol 10, pp 19-41, 2000.

[Reynolds03] Reynolds D.A., Andrews W., Campbell J.C., Navratil J., Peskin B., Adami A., Jin Q., Klusacek D., Abramson J., Mihaescu R., Godfrey J., Jones D., Xiang B. The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition. *In Proceedings of ICASSP 2003*, 2003.

[Reynolds05] Reynolds D.A., Torres-Carrasquillo P. Approaches and applications of audio diarization. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2005, Toulouse, Francia, 2005.

[Reynolds08] Reynolds D.A., Campbell W.M. Text-Independent Speaker Recognition, Springer Handbook of Speech Processing, Benesty, Sondhi, Huang Edt., Springer 2008, pp. 763-785.

[Rodriguez-Fuentes04] Rodriguez-Fuentes LJ., Torres MI. A Speaker Clustering Algorithm for Fast Speaker Adaptation in Continuous Speech Recognition. In P. Sojka, I. Kopecek and K. Pala Eds., *Proceedings of the 7th International Conference on Text, Speech, and Dialogue (Brno, Czech Republic, Septiembre 2004)*, pp 433-440, LNCS/LNAI 3206, Springer-Verlag, 2004.

[Rodriguez-Fuentes07] Rodriguez-Fuentes L.J., Penagarikano M., Bordel G. A Simple but Effective Approach to Speaker Tracking in Broadcast News. *Martí et al Eds.: IbPRIA 2007, Proceedings, LNCS* 4478, pp 48-55, 2007.

[Rodriguez-Fuentes10] Rodriguez-Fuentes L.J., Penagarikano M., Bordel G., Varona A. The Albayzin 2008 Language Recognition Evaluation. *The Speaker and Language Recognition Workshop*, *ODYSSEY'10*, Brno, Chequia, Junio 2010.

[Rosenberg76] Rosenberg A.E., Automatic Speaker Verification: A Review. In Proceedings of the IEEE, vol 64, pp. 475-187, 1976.

[Rosenberg92] Rosenberg AE., DeLong J., Lee C.H., Juang BH., Soon F.K. The use of Cohort Normalized Scores for Speaker Verification. *Proceedings of ICSLP 1992*, pp 599-602,1992.

[Rosenberg94] Rosenberg AE., Lee CH., Soong FK. Cepstral Channel Normalization Techniques for HMM-based Speaker Verification. *Proceedings of the ICSLP 1994*, pp 1835-1838, Yokohama, Japon, 1994.

[Rosenberg08] Rosenberg A.E., Bimbot F., Parthasarathy S. Overview of Speaker Recognition. Springer Handbook of Speech Processing, Benesty, Sondhi, Huang Edt., Springer 2008, pp. 725-739.

[Rougui06] Rougui JE., Rziza M., Aboutajdine D., Gelgon M., Martinez J. Fast Incremental Clustering of Gaussian Mixture Speaker Models for Scaling Up Retrieval in on-line Broadcast. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'06, Francia, Mayo 2006.

[Ruyter05] Ruyter B., Aarts E., Markopoulos P., Ijsselsteijn W. Ambient Intelligence Research in Homelab: Engineering the User Experience. Ambient Intelligence, in Weber W., Rabay J.M., Aarts E. eds., Springer Verlag, New York, pp. 49-61, 2005.

[Salah08] Salah A.A., Morros R., Luque J., Segura C., Hernando H., Ambekar O., Schouten B., Pauwels E. Multimodal Identification and Localization of Users in a Smart Environment. Journal on Multimodal User Interfaces, vol. 2, no. 2, pp. 75-91, Septiembre 2008.

[Sautrela] Penagarikano M. Sautrela: A Highly Modular Open Source Speech Recognition Framework: http://gtts.ehu.es/TWiki/bin/view/Sautrela

[Scholkopf02] Scholkopf B., Smola A.J. Learning with Kernels. *Cambridge, Massachusetts: The MIT Press*, 2002.

[Scilab] Scilab: The Open Source Platform for Numerical Computation: http://www.scilab.org/

[Siegler97] Siegler M.A., Jain U., Raj B., Stern R.M. Automatic Segmentation, Classification and Clustering of Broadcast News Audio. Proceedings of DARPA Speech Recognition Workshop, Chantilly, VA, Febrero 1997.

[Siohan99] Siohan O., Lee C.H., Surendran A.C., Li Q. Backgroud Model Design for Flexible and Portable Speaker Verification System. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'02, pp. 825-828, Marzo 1999.

[SOAa] Reference Model for Service Oriented Architecture. http://docs.oasis-open.org/soa-rm/v1.0/soa-rm.pdf

[SOAb] Reference Architecture for Service Oriented Architecture. http://docs.oasis-open.org/soa-rm/soa-ra/v1.0/soa-ra-pr-01.pdf

[Solomonoff98] Solomonov A., Mielke A., Schmidt M., Gish H. Clustering Speakers by their Voices. *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, Mayo 1998.

[Solewicz07] Solewicz Y.A., Koppel M. Using Post-Classifiers to Enhance Fusion of Low- and High-Level Speaker Recognition. IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no.7, pp. 2063-2071, Septiembre 2007.

[Sonmez98] Sonmez K., Shriberg E., Heck L., Weintraub M.Modeling Dynamic Prosodic Variation for Speaker Verification. In Proceedings of ICSLP, vol. 7, pp. 3189-3192, 1998

[Soong85] Soong F.K., Rosenberg A.E., Rabiner L.R., Juang B.H. A Vector Quantization Approach to Speaker Recognition, IEEE Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'85), vol.10, pp. 387-390, 1985.

[Stolcke05] Stolcke A., Ferrer L., Kajarekar S., Shriberg E., Venkataraman A. MLLR Transforms as Features in Speaker Recognition, *In Proceedings of European Conference on Speech Communication Technology*, pp. 2425-2428, 2005.

[Sturim01] Sturim D.E., Reynolds D.A., Singer E., Campbell J.P. Speaker Indexing in Large Audio Databases using Anchor Models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 429-432, 2001.

[Sturim05] Sturim D., Reynolds D.A. Speaker Adaptive Cohort Selection for Tnorm in text-independent speaker verification. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP'05, pp. 741-744, Philadelphia, USA, Marzo 2005.

[Tran01] Tran D., Wagner M. A Generalized Normalisation Method for Speaker Verification. *Proceedings of Speaker Odyssey 2001*, pp 73-76, 2001.

[Train04] Tran D. New Background Modelling for Speaker Verification. *Proceedings of Interspeech'04 (ICSLP)*, vol 4, pp 2605-2608, 2004.

[Tranter04] Tranter SE., Reynolds DA. Speaker Diarisation for Broadcast News. *In Proceedings of Speaker and Language Recognition Workshop, ODYSSEY'04*, Toledo, Junio 2004.

[Tranter06] Tranter SE., Reynolds DA. An Overview of Automatic Speaker Diarization Systems. *IEEE Transations on Audio Speech Language Processing*, vol 14, no 5, pp 1505-1512, 2006.

[Tsai01] Tsai W.H., Chang W.W., Huang C.S. Explicit Exploitation of Stochastic Characteristics of Test Utterance for Text-independent Speaker Identification. *In Proceedings of Eurospeech 2001*, Scandinavia, 2001.

[UPnP] UPnP Forum: http://www.upnp.org/

[Valtonen09] Valtonen M., Maentausta J., Vanhala J. TitleTrack: Capacitive Human Tracking using Floor Tiles. In Proceedings of the IEEE International Conference on Pervasive Computing and Communications, PERCOM'09, pp. 1-10, Texas, USA, Marzo 2009.

[VanLeeuwen06] Van Leeuwen DA., Martin AF., Przybocki MA., Bouten JS. NIST and NFI-TNO Evaluation of Automatic Speaker Recognition. *Computer Speech and Language*, vol. 20, pp. 128-158, 2006.

[Vapnik95] Vapnik V.N. The Nature of Statistical Learning Theory. *Springer*, 1995.

[WebServices] Web Services Architecture. http://www.w3.org/TR/ws-arch/

[Weerawarana05] Weeramana S., Curbera F., Leymann F., Storey T., Ferguson D.F. Web Services Platform Architecture. *Prentice Hall*, 2005.

[Weiser91] Weiser M. "The Computer for the twenty-first century". *Scientific American*, pp 94-104, Septiembre 1991.

[Wilcox94] Wilcox L., Chen F., Kimber D., Balasubramanian V. Segmentation of Speech using Speaker Identification. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1994*, 1994.

[WMCampbell02] Campbell W.M. Generalized Linear Discriminant Sequence Kernels for Speaker Recognition. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2002*, Orlando, Florida, Mayo 2002.

[WMCampbell04] Campbell W.M., Campbell J.P., Reynolds D.A., Jones D.A., Leek T.R. High-level Speaker Verification with Support Vector Machines, *in Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol 1., pp. 73-76, 2004.

[WMCampbell06a] Campbell WM., Sturim DE., Reynolds DA. Support Vector Machines Using GMM Supervectors for Speaker Verification. *IEEE Signal Processing Letters*, vol 13, no 5, pp 308-311, 2006.

[WMCampbell06b] Campbell WM, Campbell JP., Reynolds DA., Singer E., Torres-Carrasquillo PA. Support Vector Machines for Speaker and Language Recognition. *Computer Speech & Language*, vol 20, no 2-3, pp 210-229, 2006.

[WMCampbell06c] Campbell W.M., Sturim D.E., Reynolds D.A., Solomonoff A. SVM based Speaker Verification Using GMM Supervector Kernel and NAP Variability Compensation. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP'06, Toulouse, Francia, Mayo 2006.

[WMCampbell07] Campbell W.M., Sturim D.E., Shen W., Reynolds D.A., Navrátil J. The MIT-LL/IBM 2006 Speaker Recognition System: High-Performance Reduced-Complexity Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'07, vol. 4, pp. 217-220, Honolulu, Abril 2007.

[Wu03a] Wu T., Lu L., Chen K., Zhang H. UBM-based real-time speaker segmentation for Broadcasting News. *Proceedings of IEEE ICASSP'03*, Hong Kong, Abril 2003.

[Wu03b] Wu T., Lu L., Chen K., Zhang H. Universal Background Models for Real-time Speaker Change Detection. *Proceedings of International Conference on Multimedia Modelling*, pp 135-149, Tamshui, Taiwan, Enero 2003.

[Yang98] Yang J., Honavar V. Feature Subset Selection using a Genetic Algorithm. *IEEE Intelligent Systems*, vol 13, no 2, pp 44-49, 1998.

[Yang06] Yang H., Dong Y., Zhao X., Zhao J., Wang H. Discriminative Transformation for Sufficient Adaptation in Text-Independent Speaker Verification. Proceedings of ISCSLP 2006, Springer Verlag LNAI 4274, pp. 558-565, 2006.

[Youngblood07] Youngblood G.M., Holder L.B., Cook D.J. Data mining for hierarchical model creation. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 2007.

[Zamalloa06a] Zamalloa M., Bordel G., Rodríguez LJ., Penagarikano M. Feature Selection based on Genetic Algorithms for Speaker Recognition. *Proceedings of IEEE Speaker Odyssey 2006*, San Juan, Puerto Rico, Junio 2006.

[Zamalloa06b] Zamalloa M., Bordel G., Rodriguez LJ., Penagarikano M., Uribe JP. Using Genetic Algorithms to Weight Acoustic Features for Speaker Recognition. *Proceedings of ICSLP 2006*. Pittsburgh, Pensilvania, Septiembre 2006.

[Zamalloa06c] Zamalloa M., Bordel G., Rodriguez LJ., Penagarikano M., Uribe JP. Selección y Pesado de Parámetros Acústicos mediante Algoritmos Genéticos para el Reconocimiento de Locutor. *Proceedings of IVJTH 2006*, Noviembre 2006.

[Zamalloa08a] Zamalloa M., Rodríguez-Fuentes LJ., Penagarikano M., Bordel G., Uribe JP. Feature Selection vs Feature Transformation in Reducing Dimensionality for Speaker Recognition. *Proceedings of VJTH 2008*, Bilbao, Noviembre 2008.

[Zamalloa08b] Zamalloa M., Bordel G., Rodríguez LJ., Penagarikano M., Uribe JP. Feature Dimensionality Reduction through Genetic Algorithms for Faster Speaker Recognition. *Proceedings of Eusipco 2008*, Suiza, Agosto 2008.

[Zamalloa08c] Zamalloa M., Rodríguez-Fuentes LJ., Penagarikano M., Bordel G., Uribe JP. Comparing Genetic Algorithms to Principal Component Analysis and Linear Discriminant Analysis in Reducing Feature Dimensionality for Speaker Recognition. *Proceedings of ACM GECCO 2008*, Atlanta, USA, Julio 2008.

[Zamalloa2008d] Zamalloa M., Penagarikano M., Rodriguez-Fuentes L.J., Bordel G., Uribe J.P. University of the Basque Country +Ikerlan System for NIST 2008 Speaker Recognition Evaluation. *NIST 2008 Speaker Recognition Evaluation Workshop* (véase http://gtts.ehu.es/gtts/NT/fulltext/ZamalloaSRE08.pdf).

[Zamalloa08e] Zamalloa M., Rodriguez L.J., Penagarikano M., Bordel G., Uribe JP. Increasing Robustness to Acoustically Uncovered Signals through Shallow Source Modelling in Speaker Verification. *Proceedings of Eusipco 2008*. Laussane, Suiza, Agosto 2008.

[Zamalloa08f] Zamalloa M., Rodriguez L.J., Penagarikano M., Bordel G., Uribe JP. Improving Robustness in Open Set Speaker Identification by Shallow Source Modelling. *Proceedings of Speaker Odyssey 2008*, Stellenbosch, Sudáfrica, Enero 2008.

[Zamalloa10a] Zamalloa M., Penagarikano M., Rodriguez-Fuentes L.J., Bordel G., Uribe J.P. An Online Speaker Tracking System for Ambient Intelligence Environments, Proceedings of the Second International Conference on Agents and Artificial Intelligence, ICAART'10, Enero 2010.

[Zamalloa10b] Zamalloa M., Rodriguez-Fuentes L.J., Bordel G., Penagarikano M., Uribe J.P. Low-Latency Online Speaker Tracking on the AMI Corpus of Meeting Conversations. To appear in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'10, Texas, USA, Marzo 2010.