Tesis Doctoral

Normalización y Adaptación a Entornos Acústicos para la Robustez en Sistemas de Reconocimiento Automático del Habla.

Luis Buera Rodríguez.

Director de tesis
Eduardo Lleida Solano.

8 de marzo de 2007, last modification 8 de marzo de 2007.

1

Introducción.

La tesis doctoral "Normalización y Adaptación a Entornos Acústicos para la Robustez en Sistemas de Reconocimiento Automático del Habla" versa principalmente sobre el uso de diversas técnicas de normalización, también conocidas como de compensación, de los vectores de características para proporcionar robustez a los sistemas de Reconocimiento Automático del Habla, RAH, frente al ruido propio de los entornos acústicos.

El RAH es una disciplina científica multidisciplinar, cuyo principal objetivo consiste en extraer la secuencia de palabras pronunciadas por un locutor a partir de su señal de voz, que ha sido captada previamente mediante un sensor o micrófono. Bajo ciertas condiciones controladas, que incluyen desde la tarea propia del proceso de reconocimiento, hasta las características acústicas, pasando por todos y cada uno de los parámetros que definen el proceso completo de decodificación, los sistemas de RAH son capaces de proporcionar satisfactorias tasas de reconocimiento. Sin embargo, este hipotético entorno global ideal no deja de ser una utopía en la mayoría de las situaciones reales, que es precisamente cuando más interesante resultarían las aplicaciones basadas en RAH.

Una de las condiciones deseables y, por otra parte menos realista, consiste en que los sistemas de RAH sean independientes del entorno acústico bajo el que se pretende reconocer, de modo que, sea cual sea este último, las tasas de reconocimiento se acerquen idealmente a las obtenidas cuando se decodifica señal limpia. Conseguir este ambiciosos objetivo supondría, como se puede imaginar, abrir enormemente el abanico de posibles aplicaciones de RAH, pudiéndose introducir en ambientes hasta ahora desechados por su hostilidad acústica.

Por todo lo anterior, y a lo largo de los algo más de cuatro años que se ha precisado para completarlo, este trabajo se marcó desde el primer momento como línea de actuación el proporcionar robustez ante entornos acústicos adversos a partir de, principalmente, la normalización de los vectores de características, desarrollando e implementando distintas técnicas que, en su conjunto, se han mostrado efectivas ante diversos y variantes entornos, obteniendo unos resultados sensiblemente superiores a los logrados con los métodos más habitualmente empleados en la actualidad.

1.1. Contexto y Motivación de la Tesis.

Los recientes avances en el ámbito de las Tecnologías de la Información y las Comunicaciones, TIC, han tenido un gran impacto en el modo en que la sociedad vive, trabaja e interactua con

su entorno personal y profesional. De hecho, estas tecnologías ya están permitiendo por ejemplo desarrollar redes distribuidas de sistemas que proporcionan información, comunicación y entretenimiento allá donde el usuario se encuentre. En este contexto, una visión futurista de la Sociedad de la Información enfatiza el desarrollo de entornos en los que las personas interactúan de forma transparente con multitud de dispositivos interconectados para desarrollar las actividades de la vida diaria. Buscando la mayor comodidad para el usuario, los interfaces hombre-máquina, si bien son muchos y variados, tienden a confluir utópicamente en uno solo: la voz; ya que ésta es la manera de comunicación más intuitiva, cómoda y empleada por el hombre, parece lógico pensar que sea también la opción más natural para la comunicación con las máquinas, más allá de los problemas prácticos que ésta acarree.

Desde que, tras la aparición de los primeros ordenadores, se empezara a pensar en la posibilidad de que hombre y máquina pudieran entenderse mediante la voz, se han llevado a cabo muchas mejoras en el ámbito de los interfaces orales. Tantas, que muchas aplicaciones que hace sólo unos pocos años eran tildadas de utópicas ya se han hecho realidad formando parte, en algunos casos, del escenario cotidiano de muchas personas. Así, centralitas telefónicas, sistemas de dictado o de ayuda a minusválidos ya integran en muchas ocasiones interfaces hombre-máquina basados en voz. Sin embargo, hay que tener en cuenta que en todos estos casos comerciales se suele trabajar bajo unas condiciones controladas, y por tanto restrictivas, que hacen que las tasas de error puedan ser aceptables. A partir de lo anterior se puede concluir que aunque la antigua idea de comunicarse con una máquina de un modo natural y familiar, sin limitaciones de vocabulario, temática, contexto, locutor o ambiente esté todavía lejos, cada día lo está un poco menos.

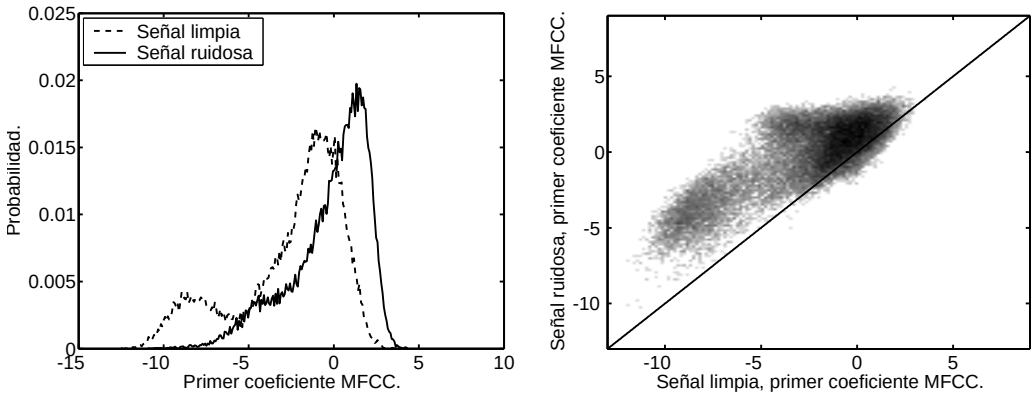
Cuando las condiciones que rodean al interfaz hombre-máquina no están controladas, la tarea de proporcionar una funcionalidad satisfactoria se complica sobremanera debido principal, aunque no únicamente, a dos causas, a saber: la incapacidad de los interfaces para abordar de forma conveniente situaciones imprevistas y el efecto pernicioso de entornos acústicos adversos, que produce generalmente una severa degradación en el comportamiento del sistema de RAH. Para compensar ambas situaciones, típicamente características de condiciones reales, se pueden plantear distintas soluciones que, en general, se agrupan en dos grandes líneas de actuación [Zue, 1997]: flexibilizar el módulo de diálogo, de modo que el usuario alcance el objetivo con el mayor grado de satisfacción global [Kamm *et al.*, 1997] [Gorin *et al.*, 1997], y diseñar sistemas de RAH robustos ante cualquier entorno acústico adverso [Cole *et al.*, 1995] [Moreno, 1996].

Anteriormente se han presentado dos de los sistemas de que se compone un interfaz hombre-máquina basado en el habla: el sistema de RAH y el módulo de diálogo. Sin embargo, y para completarlo, sería necesaria también la inclusión del módulo de comprensión. De todos modos, y a pesar de que para proporcionar una funcionalidad final satisfactoria es necesario que el comportamiento de los tres elementos sea óptimo, es el sistema de RAH sobre el que se sustenta en primera instancia el interfaz por cuanto, al encontrarse a más bajo nivel, los módulos de comprensión y diálogo dependen en gran medidad de las tasas de reconocimiento proporcionadas por él. Es por ello por lo que resulta necesario centrar el máximo esfuerzo en la tarea de mejorar dicho sistema, haciéndolo, en la medida de lo posible, inmune a cualquier tipo de variabilidad. En una parte concreta de esta tarea es en la que se circunscribe la presente tesis.

Que los sistemas de RAH proporcionen un comportamiento satisfactorio bajo condiciones controladas es algo que, hasta cierto punto, se da en la actualidad. Sin embargo, no se puede decir lo mismo cuando dichas condiciones no están acotadas convenientemente. Hay que tener en cuenta, de cara a valorar la complejidad de la tarea del RAH, que cada individuo pronuncia la misma palabra

de distinta manera (variación inter-locutor); y no sólo eso, sino que ni siquiera una misma persona pronuncia de idéntico modo el mismo vocablo en todas las ocasiones (variación intra-locutor). Asimismo, e igualmente crítico, resulta el efecto del entorno acústico, que no solamente aporta ruido que enmascara la señal de voz, sino que, de forma indirecta, hace que el locutor pronuncie de distinto modo a como lo haría en un entorno silencioso. Además de las complicaciones anteriores, que llevan asociadas una cierta causa física, los sistemas de RAH son también muy sensibles a la utilización de palabras que se hallen fuera del vocabulario, a la construcción de frases no permitidas por la gramática de la aplicación, o al uso de abreviaturas, disfluencias... De todas estas problemáticas nombradas, en el presente trabajo únicamente se estudiará como compensar el efecto del entorno acústico, dejando el resto de las mismas al margen.

El entorno acústico afecta a los sistemas de RAH, tal y como ya se ha adelantado, produciendo dos tipos de distorsiones en la señal de voz, a saber, las independientes del locutor, que serán las que se tratarán en este trabajo y que vienen dadas principalmente por el ruido aditivo y la distorsión convolucional característicos del propio entorno, y las dependientes del locutor, que se producen al articular el usuario los vocablos de distinta manera a como lo haría si se encontrara en un entorno limpio debido a la presencia de ruido externo. Para tener una idea aproximada de las distorsiones que un entorno acústico puede producir en los vectores de características con los que posteriormente se llevará a cabo el RAH, se presenta la Figura 1.1, en la que se muestran el *log-scattegram* y el histograma del primer coeficiente *Mel Frequency Cepstral Coefficient*, MFCC, de los vectores de características de voz limpia y degradada por un entorno acústico real grabado en un vehículo cuya SNR media es 8.05 dB. Se puede apreciar como dicho entorno ha producido tanto un desplazamiento de los coeficientes (modificación de la media), como una alteración en la varianza de los mismos, a la vez que la incertidumbre se ha visto incrementada debido a la naturaleza aleatoria del ruido propio del entorno acústico.



(4.a) (4.b)

Figura 1.1: *Log-scattegram* e histograma del primer coeficiente MFCC de los vectores de características de voz limpia y ruidosa para un entorno acústico real grabado en un vehículo cuya SNR media es 8.05 dB. La línea en el *scattegram* representa la función identidad $x = y$.

De cara a compensar las alteraciones en la señal de voz producidas por el entorno acústico se han planteado históricamente tres líneas de actuación, definidas por otras tantas filosofías a la hora de encarar el problema

- Parametrización o extracción de características robusta, esto es, obtener de la señal de voz

aquella información que, siendo útil para el RAH, se vea lo menos afectada posivle por el entorno acústico.

- Adaptación del sistema de RAH a las condiciones acústicas de la señal que se pretende reconocer. De esta manera, el espacio de entrenamiento utilizado para determinar los parámetros que definen la fase de reconocimiento se vería transformado de alguna manera, proyectándolo sobre el de reconocimiento.

- Adaptación de la señal que se pretende reconocer, de modo que se proyecta desde el espacio propio del entorno acústico correspondiente al de entrenamiento, que es el que, como ya se ha indicado, define los parámetros de la fase de reconocimiento.

Cabe destacar que, en general, no se puede asegurar que una línea de actuación proporcione unos resultados más satisfactorios que otra, puesto que esto depende enormemente de la aplicación concreta sobre la que se trabaje, que aporta una serie de condicionantes y limitaciones que pueden hacer, por ejemplo, que aquellos métodos que proporcionarían los mejores resultados de RAH deban ser descartados por cuestiones prácticas. En cualquier caso, este trabajo, enmarcado en el contexto de continua mejora, plantea el ambicioso reto de proporcionar robustez a los sistemas de RAH ante el entorno acústico mediante el desarrollo principalmente de nuevas técnicas de normalización de vectores de características. En este sentido, y dado que la mayoría de este tipo de métodos se basan en la aplicación de una determinada función de compensación, hay que tener en cuenta que la alteración más crítica que el entorno acústico produce en los vectores de características es, como ya se ha podido comprobar, el incremento de la incertidumbre de los correspondientes coeficientes, de modo que, para un valor concreto de un determinado vector acústico limpio, se puede dar, debido a la aleatoriedad del ruido, un amplio margen de valores para los correspondientes vectores ruidosos, y viceversa; lo que, mediante una función, no se puede compensar perfectamente.

1.2. Objetivos de la Tesis.

Tras constatar las limitaciones que los sistemas de RAH tienen a la hora de proporcionar una satisfactoria funcionalidad en entornos acústicos hostiles, se hace absolutamente necesario, si se pretende conseguir en algún momento que la sociedad emplee los interfaces orales en cualquier circunstancia y situación, el proporcionar algún tipo de solución que dote a los sistemas de RAH de la robustez imprescindible.

Después de estudiar las líneas de actuación clásicas más empleadas de cara a proporcionar robustez a los sistemas de RAH, se decidió, desde un primer momento, enfocar la tesis hacia el desarrollo de técnicas de normalización de vectores de características puesto que pueden dotar al sistema de una mayor versatilidad y, en general, necesitan de un menor tiempo computacional sin necesidad, en la mayoría de los casos, de tener que recurrir a información a priori.

Así pues, y una vez centrado el dominio del trabajo, el objetivo final es único: proporcionar la menor tasa de error posible, mejorando, en la medida de lo posible, los trabajos que, en este campo, se hayan desarrollado hasta el momento. Sin embargo, y aunque el objetivo parece claro, siempre hay que tener en cuenta otros parámetros que pueden matizar tanto los resultados obtenidos, como las comparaciones realizadas con ciertas técnicas anteriores. Así pues, en este trabajo, salvo en raras ocasiones, se proponen métodos no supervisados, de manera que no sea precisa la trascripción de los corpora correspondientes empleados en la fase previa de entrenamiento que las técnicas propuestas necesitan. Asimismo, se trabajará principalmente en el ámbito de la normalización de vectores de características empírica o, lo que es lo mismo, asumiendo que los corpora de

entrenamiento representan a entornos acústicos similares a los que posteriormente se darán en la fase de reconocimiento; aunque bien es cierto que se presentan resultados en los que no se da esta circunstancia. Por otra parte, se pretende que las técnicas desarrolladas sean eficientes en entornos acústicos variables y reales, que son los únicos que pueden incluir todo tipo de distorsiones. Por esto último se consideró *SpeechDat Car* en español como base de datos de referencia para realizar la mayor parte de la experimentación. Dicho esto también se creyó oportuno trabajar con el corpus *Aurora2* que, si bien está compuesto por señales degradadas de un modo artificial, actualmente está considerada como un banco de pruebas estándar sobre el que comparar prácticamente cualquier técnica.

Así pues, con la vista puesta en obtener la mínima tasa de error posible y teniendo en cuenta las premisas anteriores, se pueden definir una serie de subobjetivos que, a lo largo del desarrollo de la tesis, se han ido completando

- Revisión bibliográfica.

  Como paso previo a cualquier tipo de investigación es necesario conocer no sólo aquellas líneas ya desarrolladas que buscan el mismo fin bajo premisas similares, sino también aquéllas que, de un modo colateral, pueden proporcionar nuevos enfoques y ayudar así a abrir el campo de visión ante el problema cuya solución se busca. Este subobjetivo, fundamental a la hora de ahorrar tiempo posteriormente, ha quedado plasmado en los Capítulos 2 y 3.

- Estudio de las bases de datos y obtención de resultados de referencia.

  De cara a cotejar posteriormente las distintas técnicas implementadas, es necesario, en primer lugar, definir el marco de experimentación y, seguidamente, obtener unos resultados de referencia. Para la primera cuestión, tal y como se ha justificado anteriormente, se decidió recurrir a las bases de datos *SpeechDat Car* en español y *Aurora2*, que fueron analizadas convenientemente de cara a comprobar que las características de las mismas se adecuaban a los propósitos para las que habían sido seleccionadas. Por su parte, los resultados de referencia se obtuvieron tras considerar previamente distintas opciones tanto para la extracción de los vectores de características como para el modelado acústico. Este subobjetivo se corresponde, en la presente memoria, con el Capítulo 4.

- Implementación de algoritmos afines.

  A partir de la revisión bibliográfica es posible determinar aquellos algoritmos que se adecúan de un modo más estricto a los parámetros que van a regir la investigación y que, por tanto, constituyen las técnicas de referencia con las que se deberá comparar posteriormente los métodos que se vayan desarrollando a lo largo del tiempo. En la primera parte del Capítulo 5 se estudian las técnicas que, en este caso, tratan de proporcionar robustez a los sistemas de RAH cumpliendo los objetivos marcados anteriormente; a su vez se plantean algunas de las posibles limitaciones que estos algoritmos poseen.

- Desarrollo de nuevas técnicas y depurado de las mismas.

  Como consecuencia de un estudio concienzudo de las técnicas afines se puede dar con ciertas limitaciones de las mismas, lo que sirve de pie para desarrollar nuevos métodos que, tratando de mantener las ventajas de las primeras, minimicen sus debilidades. Este proceso se debe realizar con cada nuevo algoritmo desarrollado para, de esta manera, conocerlos en profundidad y poder ir generando extensiones que mejoren su comportamiento. Como se puede apreciar, este subobjetivo, que en cierto modo se basa en los anteriores, es el motor que ha movido todo el trabajo desarrollado, puesto que, desde la parte final del Capítulo 5, en el

que se presenta la primera técnica propia, hasta el Capítulo 8, se van introduciendo continuas modificaciones que tratan de compensar las diversas deficiencias detectadas tras la experimentación y posterior estudio de las mismas.

1.3. Estructura de la Memoria.

La memoria se divide en diez Capítulos que, dejando a un lado el presente e introductorio, se puede considerar que están distribuidos en dos grandes grupos temáticos. El primero de ellos, que tiene como misión establecer las bases teórico-experimentales para todo el trabajo, comprende los Capítulos 2, 3 y 4. Seguidamente a este primer grupo, se presenta el segundo, compuesto por los Capítulos 5, 6, 7, 8, 9 y 10, y en el que se exponen las diferentes técnicas, en su mayoría de compensación de vectores de características, propuestas e implementadas durante el desarrollo de esta tesis, incluyendo convenientemente los resultados obtenidos tras las correspondientes experimentaciones; asimismo, y ya en el Capítulo final, se introducen las conclusiones derivadas de todo el trabajo realizado, así como las futuras líneas de trabajo. A continuación, y de un modo somero, se resume la composición de cada uno de los distintos Capítulos.

- Segundo Capítulo.

  Dentro de este apartado se estudia, desde el punto de vista matemático-estadístico, el sistema de RAH, tratando por separado, y en Secciones diferencidas, cada uno de los distintos bloques de que se compone en su versión más generalizada, a saber: extracción de características, modelado acústico, modelado del lenguaje y procedimiento de búsqueda.

  En la Sección dedicada a la extracción de características se realiza un breve estudio cualitativo sobre los métodos más empleados a tal efecto a lo largo del tiempo, haciendo especial hincapié en los coeficientes MFCC ya que, de entre todos ellos, son los más comúnmente utilizados en la actualidad.

  Por su parte, las técnicas de modelado acústico más habituales en los sistemas de RAH, y que buscan la mejor representación estadística de los vectores de características para cada una de las unidades con que se pretenda decodificar, se presentan en su correspondiente Sección. En este caso se tratan de un modo más profundo los modelos ocultos de Markov, *Hidden Markov Models*, HMMs, por ser prácticamente un estándar de facto actualmente.

  El estudio del modelado de lenguaje también tiene su Sección dedicada correspondiente. En ella se realiza un breve repaso sobre las distintas opciones con que se puede incorporar el conocimiento lingüístico a los sistemas de RAH, incluyendo aspectos como el léxico, la semántica y la gramática. De igual modo que para los bloques anteriores, la Sección se centrará en la opción más utilizada en estos momentos: las N-gramas.

  Los procedimientos de búsqueda constituyen, en cierto modo, el motor del sistema de RAH por cuanto son los encargados de proporcionar, haciendo uso de los modelados acústico y del lenguaje, la secuencia de palabras que más fielmente se adapta al conjunto de vectores de características que se pretende decodificar. En la Sección dedicada a ellos en el Capítulo segundo se tratan aquéllos que, a lo largo del tiempo, han sido los más empleados, deteniéndose brevemente, por ser actualmente un estándar de facto, en el algoritmo de Viterbi.

- Tercer Capítulo.

  Una vez presentadas las bases matemático-estadísticas de los sistemas de RAH, en el tercer Capítulo se aborda, desde un punto de vista taxonómico, las diferentes técnicas que, a lo largo del tiempo, se han venido desarrollando para dotar de robustez a los susodichos sistemas de RAH. Así pues se distinguen, grosso modo, tres grandes líneas de actuación, a saber:

extracción robusta de características, adaptación de modelos acústicos y normalización de vectores de características, cada una de las cuales se trata de modo independiente en una Sección propia dentro del Capítulo.

En la Sección dedicada a la extraccion robusta de características, una vez explicada la filosofía de actuación que subyace bajo esta línea de actuación para proporcionar robustez a los sistemas de RAH, se realiza un breve repaso de aquellos algoritmos que, siendo los más habituales, se hallan enmarcados en ella, prestando especial hincapié tanto a los beneficios como a las limitaciones que cada uno de ellos posee.

La adaptación de modelos acústicos, como segunda opción a la hora de dotar de robustez a los sistemas de RAH, se trata en la siguiente Sección. En ella se enumeran los diversos métodos que, de una forma más o menos continuada, se han venido aplicando entre la comunidad científica. Asimismo se explica breve y principalmente de un modo cualitativo, el funcionamiento de los mismos, así como las posibles deficiencias y ventajas de unos con respecto a otros.

La última Sección perteneciente a este Capítulo versa sobre la normalización de vectores de características. En ella se presentan de una manera somera los algoritmos que, siendo incluidos dentro de esta línea de actuación, son los más empleados cuando se pretende diseñar un sistema de RAH robusto; asimismo, y dejando a un lado los desarrollos matemáticos en los que se basan, se da una idea culitativa de las limitaciones y ventajas que los diferentes métodos poseen.

- Cuarto Capítulo.

  De igual modo que los Capítulos segundo y tercero proporcionan las bases teóricas sobre las que apoyarse a la hora de presentar las distintas técnicas desarrolladas en este trabajo, el Capítulo cuarto define los parámetros de la experimentación con la que comparar de un modo cuantitativo los distintos métodos que, a lo largo del trabajo, se van a ir presentando. Por ello, en la primera Sección se estudian brevemente las dos bases de datos que se van a emplear a tal efecto: *SpeechDat Car* en español y *Aurora2*, si bien es cierto que el grueso de los resultados se obtendrán con la primera de ellas por ser mucho más realista que la segunda, que se genera tras incluir artificialmente ruido aditivo y/o distorsión convolucional a alocuciones limpias procedentes del corpus *TIDigits*.

  Dado que a la hora de cotejar distintas técnicas, que a la postre es lo que se va a determinar la mayor o menor bondad de las mismas, no basta sólo con presentar los resultados de la experimentación y compararlos directamente, sino que es preciso establecer de un modo estadístico hasta qué punto la diferencia de comportamiento es significativa, la segunda Sección de este Capítulo está dedicada a las pruebas de hipótesis estadísticas; de modo que en ella se resumen brevemente los tres algoritmos más ampliamente utilizados en el ámbito del RAH, para, finalmente, centrarse en el denomindado *z-test*, que será el que, a pesar de sus limitaciones, se empleará en este trabajo.

  Finalmente, y como último apartado de este Capítulo, se exponen los parámetros que se van a emplear durante toda la memoria a la hora de realizar las distintas experimentaciones, haciendo especial hincapié en el tipo de parametrización y estructura de los modelados acústico y de lenguaje. Asimismo se incluyen los resultados de referencia obtenidos tanto para la base de datos *SpeechDat Car* en español como para *Aurora2*, y que servirán como base para comparar posteriormente los comportamientos de las distintas técnicas presentadas en este trabajo.

- Quinto Capítulo.

El quinto Capítulo supone la puerta de entrada del segundo gran grupo temático en que se ha dividido esta memoria, y en el que se comienza la exposición propiamente dicha de las técnicas propuestas en este trabajo. Sin embargo, y antes de ello, resulta conveniente, como así se hace, estudiar la problemática que el entorno acústico introduce en el dominio de vectores de caracetrísticas. Así, la primera Sección de este apartado está dedicada al análisis, tanto desde un punto de vista teórico como práctico, de distintos tipos de ruido, pudiéndose apreciar las correspondientes alteraciones que se producen en los vectores acústicos limpios.

Una vez focalizado el problema que se pretende tratar, y tras constatar las dificultades que solucionarlo conlleva, se propone un desarrollo teórico conjunto para distintas técnicas de normalización de vectores de características empíricas basadas en el criterio *Minimum Mean Square Error*, MMSE. De este modo se puede apreciar que algoritmos ampliamente utilizados como CMN, *Cepstral Mean Normalization*, RATZ, *multivariate Gaussian-based cepstral normalization*, o SPLICE, *Stereo based Piecewise LInear Compensation for Environments*, no dejan de estar basados en el mismo principio, y que lo único que los diferencia, más allá de consideraciones conceptuales, son ciertas aproximaciones.

Aprovechando el desarrollo teórico introducido en la Sección anterior, se presenta la técnica de normalización de vectores de características MEMLIN, *Multi-Environment Model-based LInear Normalization*, que trata de compensar alguna de las limitaciones observadas en los métodos CMN, RATZ y SPLICE. Para completar el Capítulo se incluyen, en su correspondiente Sección independiente, los resultados obtenidos a partir de la base de datos *SpeechDat Car* en español con los algoritmos RATZ, SPLICE y MEMLIN, observándose una interesante mejora por parte de este último.

- Sexto Capítulo.

  El Capítulo sexto, al igual que el séptimo y octavo, surge como respuesta a algunas de las deficiencias advertidas en la técnica MEMLIN. En este caso se pretende mejorar el modelo de degradación, que para el método MEMLIN se supone lineal con término dependiente unidad, analizando nuevas posibilidades. Así, en la primera Sección se asume un modelo de degradación lineal en el que el término dependiente puede ser distinto de la unidad, dando lugar de este modo a la técnica *Polynomial Multi-Environment Model-based LInear Normalization*, P-MEMLIN.

  En la segunda Sección se presenta el algoritmo *Multi-Environment Model-based HIstogram Normalization*, MEMHIN, en el que se asume como modelo de degradación una función no lineal estimada a partir de ecualización de histograma. En el fondo esta solución no deja de ser una generalización de las anteriores, cuyos modelos de degradación pueden ser generados a partir de este último.

  Con la idea de usar transformaciones más selectivas de modo que, a su vez, generen vectores de características normalizados que se vean mejor representados por los modelos acústicos limpios, se presenta, ya en la tercera Sección de este Capítulo, la técnica *Phoneme Dependent Multi-Environment Model-based LInear Normalization*, PD-MEMLIN, que es una extensión del algoritmo MEMLIN dependiente de los fonemas.

  Llegados a este punto, y dado que hasta este momento todas las técnicas propuestas necesitan de una fase de entrenamiento previa con señal estéreo, lo que no deja de ser una limitación por cuanto ésta no siempre puede estar disponible, se presenta en la Sección cuarta una fase de entrenamiento "ciega" para la técnica PD-MEMLIN, esto es, que no precisa de señal estéreo. A pesar de que el desarrollo teórico incluido esté destinado al algoritmo PD-MEMLIN, éste es absolutamente aplicable al algoritmo MEMLIN.

Como punto final del Capítulo, la Sección quinta incluye los resultados obtenidos con la base de datos *SpeechDat Car* en español al aplicar los distintos algoritmos presentados en este Capítulo: P-MEMLIN, MEMHIN, PD-MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega". Se puede observar como las dos primeras técnicas, si bien no aportan importantes mejoras con el corpus seleccionado para la experimentación con respecto al algoritmo MEMLIN, sí se adaptan mejor ante las distorsiones que produce el ruido aditivo. Por otra parte, la técnica PD-MEMLIN proporciona una significativa mejora si se comparan sus resultados con los obtenidos con el algoritmo MEMLIN, dándose la circunstancia de que dichas prestaciones no se ven degradas de un modo drástico si se hace uso de la fase de entrenamiento "ciega".

- Séptimo Capítulo.

  Tras desarrollar en el Capítulo anterior nuevos modelos de degradación para compensar de un modo más realista los efectos que los entornos acústicos producen en los vectores de características, en el apartado séptimo se propone una nueva solución para el modelado entre Gaussianas, término este de gran importancia a la hora de calcular la estimación del vector acústico limpio con la mayoría de las técnicas de normalización propuestas en este trabajo.

  Como paso previo a la presentación de la nueva solución propuesta, en la primera Sección del Capítulo se realiza un estudio, tanto cualitativo como cuantitativo, del término de modelado entre Gaussianas para la técnica MEMLIN. Se puede apreciar que el margen de mejora que proporciona dicho término puede ser muy elevado, por lo se concluye que ciertamente merece la pena buscar una nueva solución para estimarlo. Así pues se propone modelar los vectores de características degradados asociados a los distintos pares de Gaussianas, entendiendo por par de Gaussianas, una del modelo que representa el espacio limpio y otra del modelo que hace lo propio con el espacio ruidoso. Esta solución se introduce y se desarrolla matemáticamente en la Sección segunda del Capítulo.

  En la siguiente Sección se plantea como incluir, tanto desde un punto de vista conceptual como matemático, el nuevo modelado propuesto para los pares de Gaussianas en las técnicas MEMLIN y PD-MEMLIN, teniedo en cuenta que, una vez expuesto para el primero, sería directa la aplicación a otros métodos como P-MEMLIN o MEMHIN.

  Al igual que en Capítulos anteriores, éste se finaliza con la Sección dedicada a la experimentación con la base de datos *SpeechDat Car* en español, pudiéndose constatar en esta ocasión como el nuevo modelado propuesto aporta al comportamiento de las técnicas MEMLIN y PD-MEMLIN un salto cualitativo importante en términos de RAH.

- Octavo Capítulo.

  Después de dar con diferentes soluciones para mejorar el comportamiento de la técnica MEMLIN, tanto a la hora de considerar un nuevo modelo de degradación como estimando de un modo más eficiente la probabilidad entre Gaussianas, en el Capítulo Octavo se propone combinar algunas de las técnicas de normalización presentadas hasta el momento con algoritmos que, con el objetivo de compensar la rotación entre los vectores de características normalizados y los limpios, proponen modificar los modelos acústicos. Estas soluciones híbridas se pueden englobar en dos líneas de actuación, según si las opciones propuestas son supervisadas o no. Así, dentro de la segunda línea se presentan dos algoritmos en otras tantas Secciones: las técnicas híbridas basadas en el método MATE, *augMented stAte space acousTic modEl*, y las basadas en el cálculo de matrices de rotación dependientes de GMMs, *Gaussian Mixture Models*.

  Las técnicas híbridas basadas en el método MATE se presentan en la primera Sección del Capítulo, en la que se explica, primeramente, como se decodifican los vectores de características normalizados haciendo uso de los modelos acústicos expandidos; los cuales se generan a

partir de ciertas matrices de rotación que buscan compensar la variabilidad inter-locutor del mismo modo en que lo haría el algoritmo VTLN, *Vocal Track Length*.

La segunda Sección presenta una serie de técnicas que, a nivel de sistema, son similares a las introducidas en la Sección anterior. Sin embargo, conceptualmente hablando, difieren significativamente ya que, en este caso, las matrices de rotación no se estiman con la idea de compensar la variabilidad inter-locutor del mismo modo que lo hace el algoritmo VTLN, sino que se adaptan a las necesidades del entorno; por ello, dichas matrices se calculan dependientes de unas ciertas GMMs que modelan el espacio limpio y el normalizado.

En la tercera Sección se presentan aquellas soluciones híbridas basadas en el entrenamiento de modelos acústicos en el espacio normalizado, para lo que es necesario conocer las trascripciones del corpus de entrenamiento, generando por tanto soluciones híbridas supervisadas. Así pues, en este caso los vectores de características compensados se decodifican directamente con dichos modelos.

Para comparar las diversas técnicas propuestas en este Capítulo, la última Sección se dedica a presentar los resultados obtenidos por todas ellas sobre la base de datos *SpeechDat Car* en español, pudiéndose comprobar que las soluciones híbridas presentadas proporcionan, en todos los casos, una importante mejora si se comparan los resultados obtenidos con los correspondientes a las distintas técnicas de normalización de vectores de características que sirven de base.

- Noveno Capítulo.

  El Capítulo noveno se articula en tres Secciones, en las que se presentan los resultados de RAH obtenidos a partir de la base de datos *Aurora2* tras aplicar las técnicas más representativas propuestas en este trabajo. De este modo, la primera Sección está dedicada al algoritmo MEMLIN, observándose, a pesar de no disponer en el corpus de entrenamiento de todos los tipos de ruidos que posteriormente aparecen en el de reconocimiento, una importante mejora relativa final.

  La segunda Sección está destinada a presentar los correspondientes resultados extraídos con la base de datos *Aurora2* tras aplicar la técnica MEMLIN con el modelado de la probabilidad entre Gaussianas propuesto en el Capítulo séptimo, pudiéndose constatar una cierta mejora en el comportamiento de la técnica si se compara con el del algoritmo MEMLIN bajo las mismas condiciones de experimentación; hecho este que no hace sino afianzar las conclusiones que, tras los experimentos realizados con la base de datos *SpeechDat Car* en español, se presentaron en el Capítulo octavo.

  En la tercera y última Sección se muestran los resultados de RAH obtenidos tras aplicar la solución híbrida no supervisada en la que la técnica de normalización de los vectores de características es el método MEMLIN con el modelado de la probabilidad entre Gaussianas introducido en el Capítulo séptimo, y las matrices de rotación se estiman del modo en que se explica en la segunda Sección del Capítulo octavo, esto es, buscando adaptarse a las necesidades propias del entorno acústico. FALTA DECIR SI SALE BIEN (FALTAN LOS DATOS).

- Décimo Capítulo.

  La memoria finaliza con un Capítulo, el décimo, dedicado a las conclusiones y futuras líneas de trabajo que, apoyándose en las técnicas, resultados e ideas propuestas a lo largo de la memoria, podrían llegar a dar lugar a nuevos algoritmos que, tratando de mantener las ventajas de los propuestos, compensen algunas de sus debilidades.

1.4. Principales Contribuciones.

Durante los algo más de cuatro años que han sido necesarios para completar esta tesis se han publicado en revistas y diversos congresos distintos trabajos, cuyo recorrido cronológico proporciona, si bien no una visión tan compacta como se ha pretendido dar a la memoria, sí una más realista del modo en que se fueron sucediendo los diferentes hitos que, a la postre y conjuntamente, constituyen la presente tesis.

En los primeros trabajos [Buera *et al.*, 2004a] [Buera *et al.*, 2004c], se presentó la técnica MEMLIN, comparando su comportamiento ante variables entornos acústicos adversos con otros métodos de normalización de vectores de características basadas en similares principios y típicamente utilizados por la comunidad científica. Asimismo se empezó a desarrollar una teoría conjunta para todos aquellos algoritmos de compensación basados en el estimador MMSE, de modo que cualquiera de ellos se podría ver como una realización más o menos compleja de una misma idea y supeditada a ciertas aproximaciones.

Posteriormente, y tras analizar la técnica MEMLIN, se llegó a la conclusión de que el modelo de degradación propuesto, lineal con término dependiente unitario, quizás no fuera el más propicio para compensar ciertas degradaciones que los entornos acústicos producen en la señal de voz. Por ello se propuso, por un lado, aplicar un modelo no lineal basado en ecualización de histograma [Buera *et al.*, 2004b], dando lugar así a la técnica MEMHIN, y por el otro, considerar transformaciones dependientes de los fonemas [Buera *et al.*, 2005c] [Buera *et al.*, 2005a], generando el algoritmo PD-MEMLIN. Para el primero de los casos se pudo observar, al compensar las degradaciones producidas sobre la varianza de los vectores de características, una importante mejora en los resultados ante entornos acústicos caracterizados por ruido aditivo. Por su parte, la versión de la técnica MEMLIN dependiente de los fonemas, PD-MEMLIN, se mostró más efectiva desde el primer momento ante cualquier tipo de distorsión de la señal de voz.

Una vez comprobadas las bondades de la técnica PD-MEMLIN en el ámbito del RAH, se abrió una nueva línea de investigación en el dominio de la verificación e identificación de locutor. De este modo se pudo comprobar que el método PD-MEMLIN también proporciona unas importantes mejoras cuando estas nuevas tareas se desarrollan en entornos acústicos hostiles [Buera *et al.*, 2005a] [Buera *et al.*, 2005d] [Buera *et al.*, 2006c]. Sin embargo, y dado que los experimentos propuestos en los distintos trabajos no dejan de ser hasta cierto punto preliminares, no se ha incluido en esta memoria ninguna Sección específica dedicada a la verificación e identificación de locutor, aunque sí es cierto que se pretende retomar esta línea de trabajo en un futuro próximo.

Una de las principales limitaciones que plantean en muchas ocasiones las técnicas de normalización de vectores de características empíricas, como la mayoría de las presentadas en este trabajo, es la necesidad de poseer, de cara a estimar los distintos parámetros que definen los correspondientes métodos, un corpus de entrenamiento estéreo. Esta problemática, que no se ha considerado para algunas de las técnicas empíricas más utilizadas por la comunidad científica, sí se trabajó en [Buera *et al.*, 2005b], donde se presenta una fase de entrenamiento no estéreo para la técnica PD-MEMLIN que es igualmente aplicable al algoritmo MEMLIN. Se pudo observar, a partir de la consiguiente experimentación, que el hecho de hacer uso en la fase de entrenamiento únicamente de la señal ruidosa no supone una importante merma en el comportamiento del método.

La técnica MEMLIN, así como todas aquellas variantes de la misma que se basan en nuevos modelos de degadación, esto es, los algoritmos MEMHIN, PD-MEMLIN y P-MEMLIN se trataron conjuntamente, así como la versión de entrenamiento no estéreo para el método PD-MEMLIN, en

[Buera *et al.*, 2007]. De este modo, y por establecer una relación entre las distintas contribuciones con la estructura de la memoria, las diferentes soluciones presentadas hasta el momento tienen su eco en los Capítulos 5 y 6.

Tras analizar las distintas opciones que el modelo de degradación proporciona, el siguiente término que se analizó, y que está presente no sólo en la técnica MEMLIN, sino también en todas aquellas que se derivaron de ella, fue el modelado de la probabilidad entre Gaussianas. La solución propuesta hasta entonces se basaba en la presunción de que dicho término era independiente del vector de características ruidoso, consideración esta que no deja de ser una aproximación. Como solución más realista se propuso en [Buera *et al.*, 2006b] [Buera *et al.*, 2006a] entrenar un modelo basado en GMMs para los vectores de características ruidosos asociados a cada par de Gaussianas, obteniéndose importantes mejoras tanto si se aplica dicha solución a la técnica MEMLIN [Buera *et al.*, 2006b], o al método PD-MEMLIN [Buera *et al.*, 2006a]. Tanto las ideas y conceptos como los consiguientes desarrollos teóricos necesarios para estimar los nuevos modelos de probabilidad entre Gaussianas se contemplan, junto con la experimentación realizada, en el Capítulo 7.

FALTA HÍBRIDOS. CUANDO SE PUBLIQUE ALGO...

Adicionalmente, y como colaboración con el departamento de Ciencias Computacionales del Tecnológico de Monterrey, campus Monterrey, se presentó el trabajo [Hernández *et al.*, 2007], en el que se emplea la técnica PD-MEMLIN tras aplicar, como paso previo, el método SS, *Spectral Subtraction*, dando lugar al algoritmo PD-MEELIN, *Phoneme Dependent Multi-Environment Enhanced Model based LInear Normalization*. De este modo se trata de compensar en primera instancia los efectos producidos por el ruido aditivo para, posteriormente, actuar sobre la distorsión convolucional. La experimentación presentada en el trabajo se realizó sobre la base de datos *Aurora2*, dando lugar a unos resultados especialmente competitivos, aunque en esta memoria, por tratarse de una línea de trabajo aún emergente, no se han incluido. Actualmente se sigue colaborando para formalizar la técnica PD-MEELIN, buscando sus debilidades e incorporando nuevas mejoras que proporcionen un mejor comportamiento ante entornos acústicos altamente adversos.

2

Sistemas de Reconocimiento Automático del Habla.

El Reconocimiento Automático del Habla, RAH, es una disciplina científica multidisciplinar, cuyo principal objetivo es extraer la secuencia de palabras pronunciadas por un locutor a partir de su señal de voz, que ha sido captada previamente. A pesar de que las primeras aproximacionas serias datan ya de los años 70 [Baum, 1972] [Jelinek, 1976], en la actualidad el RAH no es ni mucho menos un problema resuelto debido principalmente a la variabilidad de la señal de voz y a los factores que, externos a ella, pueden afectarla, como el entorno acústico, el tipo de micrófono, etc.

Cuando las condiciones que rodean a los sistemas de RAH no están controladas, el proporcionar aceptables tasas de reconocimiento se complica sobremanera. Hay que tener en cuenta que cada locutor pronuncia la misma palabra de distinta manera (variación inter-locutor); y no sólo eso, sino que ni siquiera una misma persona pronuncia de idéntico modo el mismo vocablo en todas las ocasiones debido a cambios en sus condiciones físicas o psíquicas, que nunca permanecen constantes (variación intra-locutor). A todo esto hay que añadir el efecto del entorno acústico, que introduce ruido que enmascara la señal de voz a la vez que puede llegar a alterar el propio proceso de producción de la misma mediante el efecto Lombard [Lombard, 1911]. Otros problemas que se pueden dar y que hacen del RAH una disciplina tan compleja son, por ejemplo, la utilización de palabras no contempladas en el vocabulario de la aplicación, la construcción de frases no permitidas por la gramática del lenguaje, el uso de abreviaturas y disfluencias, los escenarios semánticos de las palabras, etc. De todo lo anterior se puede concluir pues que la señal de voz posee una gran variabilidad debido a múltiples causas y que es por ello por lo que se hace extremadamente complejo su modelado y posterior reconocimiento, a no ser, claro está, que el entorno de trabajo se encuentre lo suficientente controlado. Así pues, y atendiendo al acotamiento en mayor o menor medida de los problemas anteriormente comentados, los sistemas de RAH se clasifican atendiendo a distintos criterios [Moore, 1990]

- Dependencia con respecto al locutor. Los sistemas de RAH pueden ser, considerando este criterio, dependientes o independientes del locutor. En el primero de los casos, el sistema se entrena para que sólo lo use una única persona, mientras que si el sistema es independiente del locutor se acondiciona para que lo pueda emplear un gran abanico de usuarios, idealmente cualquiera. En general, los sistemas dependientes del locutor proporcionan unas mayores tasas de reconocimiento a costa, eso sí, de perder generalidad. Estas característias se invierten para el caso de sistemas independientes del locutor. También existen sistemas intermedios, como los multilocutor y los adaptados al locutor. El primero de ellos está pensado para ser empleado por un grupo reducido de usuarios, mientras que los adaptados al locutor parten de

un sistema independiente del locutor y, tras modificar los parámetros necesarios, lo acercan a las prestaciones de uno dependiente del locutor. Por otra parte, también se pueden clasificar los sistemas de RAH atendiendo a los grados de cooperación y experiencia de los usuarios.

- Dependencia con respecto al vocabulario. El incrementar el número de palabras que se pueden reconocer proporciona un sistema más versátil y completo que puede resultar muy útil en gran cantidad de circunstancias y tareas, pero como contraprestación el coste computacional y la tasa de error se resienten. Por todo ello en los sistemas de RAH se debe ajustar al máximo el vocabulario a la tarea que se pretende realizar. Además de por el tamaño, los sistemas de RAH se puede clasificar atendiendo a otros criterios referentes al vocabulario como el grado de discriminalidad del mismo, de dependencia con respecto a la aplicación...

- Dependencia con respecto al tipo de discurso. Atendiendo a este criterio se pueden reconocer desde palabras aisladas hasta habla continua, pasando por cualquier estado intermedio. Cuando se pronuncian palabras aisladas, esto es, con importantes pausas entre ellas, la tasa de reconocimiento es mucho mayor que si se pretende reconocer un discurso continuo. Asimismo, los sistemas de RAH no se comportan del mismo modo si el habla es leída o espontánea, de la misma manera que en ciertas aplicaciones se debe tener en cuenta el nivel de rechazo ante habla extraña.

- Dependencia con respecto a la estructura del diálogo. En este caso se definen las características de un sistema de RAH frente a la capacidad de procesamiento del lenguaje, clasificando los sistemas de forma gradual atendiendo a la perplejidad o a la tarea, pudiendo ir desde el reconocimiento de comandos aislados, que se corresponde con el nivel más bajo, hasta de lenguaje natural, lo que supondría el nivel más complejo.

- Dependencia de las condiciones de trabajo, que hace referencia a la variabilidad del entorno, especialmente acústico, en el que está inmerso el sistema de RAH. De este modo, no se obtienen las mismas tasas de reconocimiento bajo condiciones de laboratorio que en situaciones reales, normalmente más adversas que las primeras.

Este conjunto de descriptores permite, además de clasificar los sistemas de RAH, compararlos en cuanto a prestaciones, a la vez que dan una idea de las posibles fuentas de variabilidad que se pueden presentar a la hora de plantear una cierta aplicación, elemento este de capital importancia ya que la robustez del sistema ante las mismas determinará el rendimiento final del sistema.

En la Tabla 2.1, y para comprobar lo alejados que aún se encuentran los sistemas de RAH con respecto a las capacidades humanas, se presentan los resultados de tasa de error por palabra, *Word Error Rate* WER, obtenidos para distintas tareas tanto por humanos como por un sistema convencional de RAH a la vanguardia del estado del arte [Huang *et al.*, 2001] [Lleida *et al.*, 2002]. Igualmente junto a cada tarea se incluyen el número de palabras que componen el vocabulario correspondiente. Con estos resultados se puede constatar el largo camino que todavía hoy le queda por recorrer a la comunidad científica a la vez que cabe destacar como únicamente en aquella tarea en la que se reconocen secuencias de trigramas carentes de significado, el sistema de RAH proporciona mejores resultados que los humanos, lo que es debido a que éstos hacen un uso mucho más eficiente del contexto léxico.

Este Capítulo, centrado en los sistemas de RAH, se estructura del siguiente modo: en la Sección 2.1 se explicarán los fundamentos matemáticos de los sistemas de RAH basados en un efoque probabilístico bayesiano. Posteriormente, y en las siguientes Secciones, se irán desgranando cada uno de los componentes que constituyen el sistema de RAH completo. Así, en la Sección 2.2 se

tratará la extracción de las características a partir de la señal de voz, que determinarán a la postre los vectores acústicos que se utilizarán a la hora de reconocer. En la Sección 2.3 se estudiarán los distintos tipos de modelado acústico que actualmente se están utilizando. Por su parte, la Sección 2.4 está dedicada al modelado del lenguaje que, junto con el acústico, compone la base estadística de todo sistema de RAH basado en un enfoque estadístico. Finalmente en la Sección 2.5 se explica cómo se realiza la búsqueda entre las palabras del vocabulario hasta dar con la secuencia de las mismas más probable.

Tareas	# Vocabulario	WER Humanos (%)	WER Máquinas (%)
Dígitos conectados	10	0.009	0.72
Deletreo	26	1	5
Conversaciones telefónicas espontáneas	2000	3.8	36.7
WSJ con señal de voz libre de ruido	5000	0.9	4.5
WSJ con señal de voz ruidosa (10-dB SNR)	5000	1.1	8.6
Trigramas de señal de voz libre de ruido	20,000	7.6	4.4

Cuadro 2.1: Tasas de error de palabras, *Word Error Rate*, WER, entre humanos (columna "Humanos WER (%)") y un sistema de RAH a la vanguardia del estado del arte (columna "Máquinas WER (%)") para distintas tareas con diferente número de palabras para los correspondientes vocabularios (columna "# Vocabulario"). WSJ: *Wall Street Journal database*.

2.1. Reconocimiento Automático del Habla.

El Reconocimiento Automático del Habla, RAH, tal y como ya se ha indicado, es una disciplina científica cuyo principal objetivo es extraer la secuencia de palabras pronunciadas por un locutor a partir de su señal de voz captada previamente. Para ello, y a pesar de que los sistemas de RAH a lo largo de su corta historia se han basado en distintas técnicas, en la actualidad las aproximaciones más utilizadas poseen un enfoque probabilístico basado en el teorema de Bayes, la teoría de la información y las técnicas de comparación de patrones y de programación dinámica [Bellman and Kabala, 1965] [Ney, 1990] [Ney, 1993] [Duda *et al.*, 2000].

Desde este enfoque probabilístico, y en una primera y simple aproximación, se puede decir que todo sistema de RAH debe disponer de unos patrones asociados a las distintas partes del habla que se pretende reconocer, de modo que, dado un conjunto de observaciones acústicas como entrada, devolverá la secuencia de patrones que con mayor probabilidad lo represente.

El conjunto de observaciones acústicas, $\mathbf{O}$, consiste en una secuencia de vectores de parámetros que se extraen de la señal de audio captada previamente mediante un sensor o micrófono. Dicha extracción pretende obtener aquellas características de la señal de voz más representativas de cara al RAH. Por todo ello, a las observaciones acústicas también se las conoce como vectores de características

$$\mathbf{O} = (\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_T), \tag{2.1}$$

donde t es el índice temporal asociado a las observaciones acústicas, $t \in [1, T]$. Por su parte, y tal y como se ha comentado con anterioridad, la salida del sistema de RAH será una secuencia de palabras $\mathbf{W}$, que idealmente debería coincidir con las pronunciadas por el locutor

$$\mathbf{W} = (w_1, ..., w_n, ..., w_N), \tag{2.2}$$

donde n es el índice temporal asociado a los vocablos reconocidos, $n \in [1, N]$. Para hallar la secuencia de palabras óptima, siempre desde el enfoque probabilístico, se buscará aquélla que tenga la mayor probabilidad de estar asociada a la secuencia de observaciones acústicas de entrada, $p(\mathbf{W}|\mathbf{O})$, lo que matemáticamente se expresa mediante (2.3). Dado que esta expresión no es directamente calculable salvo para problemas muy sencillos, con un conjunto de observaciones de una dimensionalidad y tamaño muy reducidos, se recurre al teorema de Bayes, de modo que (2.3) queda expresada como (2.4). Así pues, el nuevo problema de estimación se puede ver como la búsqueda de la secuencia de palabras que proporciona la máxima probabilidad a priori, $p(\mathbf{W})$ (modelo de lenguaje) y que además produce la secuencia de observaciones con máxima probabilidad, $p(\mathbf{O}|\mathbf{W})$ (modelo acústico). Es decir, mediante el teorema de Bayes se ha dividido el intratable primer problema en otros dos de más sencilla resolución: un problema de decodificación lingüística y otro de decodificación acústica.

$$\mathbf{W} = \arg\max_{\mathbf{W}} p(\mathbf{W}|\mathbf{O}), \tag{2.3}$$

$$\mathbf{W} = \arg\max_{\mathbf{W}} \frac{p(\mathbf{O}|\mathbf{W})p(\mathbf{W})}{p(\mathbf{O})}, \tag{2.4}$$

donde $p(\mathbf{O})$ es la probabilidad a priori de las observaciones acústicas y que, dado que resulta intrascendente a la hora de estimar la frase pronunciada, no se tiene en cuenta finalmente en la maximización [Rabiner and Juang, 1993]. De esta manera, la expresión que se debe evaluar para obtener la secuencia óptima de palabras pronunciada es (2.5)

$$\mathbf{W} = \arg\max_{\mathbf{W}} p(\mathbf{O}|\mathbf{W})p(\mathbf{W}). \tag{2.5}$$

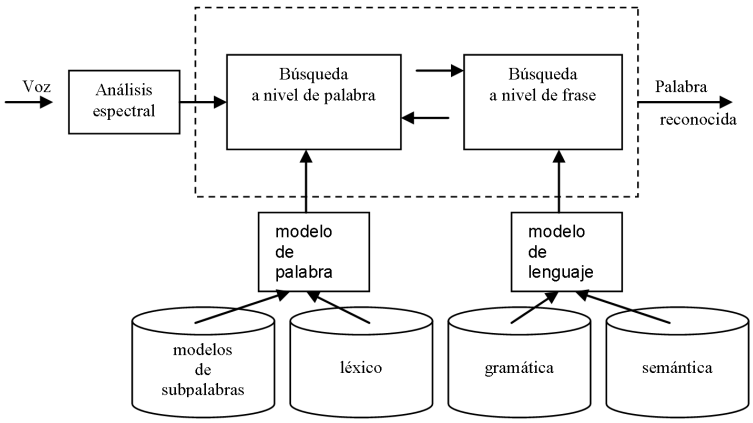


Figura 2.1: Esquema general de un sistema de Reconocimiento Automático del Habla, RAH, basado en un enfoque estadístico Bayesiano.

Para llevar a cabo todo el proceso necesario de RAH según el enfoque estadístico Bayesiano, esto es, para evaluar la expresión (2.5), son necesarios cuatro bloques fundamentales, los cuales se encuentran representados en la Figura 2.1 y son: extracción de las características de la señal de

voz, modelado acústico, modelado del lenguaje y búsqueda de la secuencia de palabras óptima. A continuación se comenta brevemente cada uno de ellos

- Extracción de características. Los sistemas de RAH no utilizan directamente las muestras de audio como entrada, sino que éstas se preprocesan buscando aquellas características óptimas de cara al reconocimiento. De este modo se obtienen unos vectores de parámetros representativos que son los que realmente constituyen la entrada al sistema de RAH. Sobre este bloque se entrará en más detalle en la Sección 2.2.
- Modelado acústico. Este bloque describe la probabilidad de observar un conjunto de vectores acústicos dada una secuencia de palabras, $p(\mathbf{O}|\mathbf{W})$. Típicamente los modelos acústicos de palabras se construyen a partir de modelos de unidades menores haciendo uso de una serie de reglas que rigen como unir estos últimos. En la Sección 2.3 se abordará con más profundidad este bloque.
- Modelado de lenguaje. Esta unidad cubre la sintaxis, semántica y gramática del lenguaje, aspectos estos que quedan reflejados matemáticamente en el cálculo de la probabilidad a priori de las diferentes secuencias de palabras, $p(\mathbf{W})$. La Sección 2.4 trata con más detalle este bloque.
- Procedimiento de búsqueda. Esta unidad tiene como meta encontrar la frase óptima, esto es, aquella que posea la máxima probabilidad a posteriori dada la secuencia de vectores acústicos mediante el teorema de Bayes, tal y como indica la expresión (2.5). La Sección 2.5 proporciona una visión más profunda de este bloque.

2.2. Extracción de Características.

El objetivo de las distintas técnicas de extracción de características es, a partir de la señal de audio previamente grabada mediante un micrófono o sensor, proporcionar unos vectores de parámetros que, idealmente deberían cumplir las siguientes tres características, a saber

- Representar cada segmento de voz mediante un vector compuesto por el menor número de parámetros posible, de modo que se logre un cierto grado de compresión y, por consiguiente, la reducción del tiempo necesario para procesar dicho vector. De cualquier otra manera sería imposible llegar a reconocer en tiempo real en la mayoría de las aplicaciones y tareas.
- Hacer uso sólo de aquellas características de la señal de voz más representativas y que, por su naturaleza, se adecúen óptimamente a cada aplicación concreta, ya que no todas deben ser tenidas en cuenta del mismo modo. Así, por ejemplo, una buena parametrización para sistemas de RAH tendrá en cuenta el tracto vocal, mientras que se desecharán otras cualidades de la voz que puedan generar modelos sesgados, como por ejemplo el pitch, que proporcionaría unos modelos acústicos altamente dependientes del locutor y que, sin embargo, sí podría ser de gran utilidad para sistemas de reconocimiento de locutor.
- Ser robusta, de tal forma que cualquier alteración sobre la señal de voz afecte de la menor forma posible a los vectores de características. De este modo, los distintos sistemas de RAH que utilicen la correspondiente parametrización podrían tener un comportamiento satosfactorio aun cuando los desajustes entre las señales empleadas para obtener los modelos acústicos y las que se pretenden reconocer fueran sensibles.

En la actualidad son dos los sistemas de extracción de características más comúnmente utilizados, [Davis and Mermelstein, 1980]: los coeficientes LPC (*Linear Prediction Coefficients*) y los

MFCC (*Mel-Frequency Cepstral Coefficients*). La parametrización LPC surge al aplicar análisis de predicción lineal a la señal de voz y, en algunas ocasiones, se modifican de cara a obtener una representación más adecuada para los sistemas de RAH, dando lugar a los parámetros cepstrum LPC [Oppenheim and Schafer, 1975] [Rabiner and Juang, 1993]. Por su parte, los coeficientes MFCC se obtienen tras transformar el espectro de los coeficientes cepstrales de la señal de voz a la escala bark mediante una transformación Mel.

QUIZÁS ESQUEMA CON PARAMETRIZACIÓN ETSI COMO MFCC

A pesar de que los coeficientes LPC y MFCC son los más extendidos, a lo largo del tiempo han propuesto diferentes mejoras y técnicas en la extracción de características [Mori, 1997]. Así, por ejemplo, se han desarrollado representaciones basadas en el modelo auditivo humano, como los coeficientes PLP, *Perceptual Linear Prediction* [Hermansky, 1990], el modelo EIH, *Ensemble-Interval Histogram* [Ghitza, 1992] [Ghitza, 1994] [Rabiner and Juang, 1993], o los modelos auditivos asíncronos, como el de Seneff [Jankowski *et al.*, 1995] o el SLP, *Synchronous Linear Prediction* [Junqua and Haton, 1996]. Todas estas representaciones basadas en el modelo auditivo humano proporcionan, especialmente en condiciones acústicas adversas, buenos resultados si se comparan con los coeficientes cepstrum LPC [Junqua and Wakita, 1989] [Junqua and Haton, 1996] [Jankowski *et al.*, 1995]; sin embargo no consiguen una significable mejoría si se comparan con la parametrización basada en los coeficientes MFCC. Si a eso se le añade el hecho de que el tiempo necesario para la extracción de dichos coeficientes basados en el modelo auditivo humano es, en general, mucho mayor, se puede comprender que este tipo de parametrizaciones no sea tan difundida ni se haya empleado apenas en sistemas de RAH en tiempo real, donde mayoritariamente se hace uso de los coeficientes MFCC.

Otras alternativas y mejoras a los sistemas de extracción de características clásicos son *modulation spectrum* [Kingbury *et al.*, 1998] [Greenberg and Kingsbury, 2000], que mediante el uso de filtros paso bajo se trata de eliminar ciertas componentes de la señal que pudieran resultar negativas de cara a la aplicación de RAH concreta, el uso de modelos perceptuales de enmascaramiento, que ocultan el ruido que pudiera afectar a la señal de voz [Usagawa *et al.*, 1994], la utilización de técnicas más robustas que la transformada de Fourier para obtener el espectro, como por ejemplo las transformadas wavelets [Rioul and Vetterli, 1991], [Erzin *et al.*, 1995], el desarrollo de parametrizaciones basadas en operadores no lineales que representen de un mejor modo la generación de la voz en el tracto vocal y sus irregularidades, como el operador TEO, *Teager Energy Operador* [Kaiser, 1990], o la modificación de la escala Mel en el cálculo de los coeficientes MFCC, de tal forma que aquellas bandas frecuenciales más afectadas por el ruido tengan menos peso en el cálculo final de los coeficientes, a la vez que se favorece aquéllas que se encuentran menos contaminadas [Bou-Ghazale and Hansen, 2000].

A la hora de construir el vector de características que finalmente conformará la entrada al sistema de RAH, se suele incluir no sólo los parámetros calculados mediante alguna de las técnicas de extracción de características consideradas anteriormente, sino también otros parámetros, como pueden ser la energía, la frecuencia de los formantes [Holmes *et al.*, 1997], o la velocidad de pronunciación [Morgan *et al.*, 1997]. Por su parte, y para modelar la correlación temporal existente entre los vectores acústicos próximos se suele hacer uso de la primera y segunda derivada [Furui, 1986], o estudiar la información presente en fragmentos de señal más amplios que la ventana de análisis [Hermansky, 1998].

De todos modos, y a pesar de los esfuerzos realizados en este campo, hasta la fecha no se ha dado con ninguna técnica de parametrización que cumpla a la perfección con las tres características básicas que, tal y como se ha comentado con anterioridad, idealmente deberían poseer; de ahí que en muchos casos se requiera de una serie de técnicas que compensen tales limitaciones.

2.3. Modelado Acústico.

Tal y como se ha indicado anteriormente, el modelado acústico tiene como misión el determinar la probabilidad de observar un conjunto de vectores de parámetros dada una secuencia de palabras, $p(\mathbf{O}|\mathbf{W})$. Para ello se emplean diversas técnicas de aprendizaje que hacen uso de amplios corpora de audio. Dado que el problema de modelar probabilísticamente secuencias de una manera completa puede llegar a ser computacionalmente inviable debido a que la complejidad crece exponencialmente con la longitud de la secuencia de observaciones, se han venido considerando distintas aproximaciones de independencia que hacen el modelado acústico más sencillo [Baum, 1972] [Jelinek, 1976] [Rabiner, 1988].

Actualmente los modelos ocultos de Markov, HMMs, *Hidden Markov Model* [Baker, 1975] [Rabiner, 1988], constituyen el modelado acústico más extendido entre los sistemas de RAH. Dichos modelos son unos autómatas de estados finitos en los que cada uno de los estados posee asociada una función de densidad de probabilidad, pdf, *probability density function*, que normalemente suele ser una mezcla de Gaussianas, GMM, *Gaussian Mixture Model*, aunque podría ser cualquier otra. Por otra parte, los estados se relacionan unos con otros mediante probabilidades de transción. Para los HMMs empleados en los sistemas de RAH, las aproximaciones de independencia anteriormente comentadas se materializan mediante dos consideraciones: el proceso estocástico de generación de observaciones sólo depende de un estado del modelo en cada momento y se supone que el transitar entre estados dentro del modelo depende únicamente del estado origen y destino. Al tipo de modelo oculto de Markov que cumple estas estrictas restricciones se le denomina de orden 1 [Ghahramani, 2002], mientras que el entrenamiento de las diversas variables que conforman los HMMs (las probabilidades de transición entre estados y los parámetros de las funciones de densidad de probabilidad de cada estado) se realiza mediante la estimación ML, *Maximum Likelihood*, haciendo uso del algoritmo iterativo EM, *Expectation Maximization* [Dempster *et al.*, 1977].

QUIZÁS HABRÍA QUE PONER UN ESQUEMA DE HMM Y ALGO DE GMM (gentle tutorial).

Tal y como se puede deducir de las aproximaciones de independencia mencionadas, los HMMs de orden 1, a pesar de considerarse un estándar de facto debido a los buenos resultados obtenidos, poseen ciertas debilidades que los alejan de la naturaleza real de la señal de voz, como la interdependencia temporal de las realizaciones sonoras y el carácter no discreto del proceso de producción de la voz. Por todo ello se han desarrollado a lo largo del tiempo distintas extensiones de los HMMs que tratan de compensar dichas limitaciones [Holmes and Huckvale, 1994]. Así, en busca de estos objetivos se desarrolló el modelado basado en la descomposición temporal, previamente empleada en codificación [Atal, 1983], y en la que se ve la señal de voz como una combinación de ciertas funciones base controladas por un parámetro, [Bimbot *et al.*, 1988] [Deleglise, 1990] [Lleida, 1990]. Esta técnica, si bien no se ha continuado desarrollando, mantiene algunas similitudes conceptuales con otras extensiones de los HMMs, como los modelos segmentales y los de trayectorias.

Los modelos segmentales tratan de eliminar alguna de las debilidades de los HMMs de orden 1 mediante la concatenación de submodelos de fragmentos de voz, no necesariamente de la mis-

ma longitud y síncronos como sucede con los HMMs [Ostendorf *et al.*, 1996] [Glass, 2003]. Por su parte, los modelos de trayectorias tratan de parametrizar los descriptores de la señal de voz de manera que el parámetro que controla la forma de la señal o de su evolución se pueda adaptar de una manera más fina a sus cambios [Goldenthal, 1994] [Sun, 1995] [Gish and Ng, 1996].

Otra extensión de los HMMs empleada para el modelado acústico son los Campos de Markov, *Markov Random Fields*. Esta técnica, concebida inicialmente para tratamiento de imagen [Gravier *et al.*, 1999] [Gravier, 2000], trata de modelar el espectro tiempo-frecuencia de la señal de voz modificando el índice temporal de los estados de los clásicos HMMs, que pasa de ser unidimensional a bidimensional ya que las relaciones de dependencia se definen en este caso como vecindades 2-dimensionales [Lafferty *et al.*, 2001] [Zhu and Ghahramani, 2002].

Los modelos de Markov para la generación de observaciones, más conocidos como HMMs2, suponen otra modificación de los HMMs clásicos utilizada para el modelado acústico. En este caso, la función de densidad de probabilidad asociada a cada estado de los modelos clásicos, normalmente, tal y como ya se ha comentado, una mezcla de Gaussianas, se sustituye por un nuevo HMM. Con ello se pretende modelar la variabilidad que existe en la evolución de los formantes en la señal de voz [Weber *et al.*, 2000], aunque hasta el momento no se ha conseguido que los HMMs2 sean lo suficientemente discriminativos [Weber, 2003] como para obtener unos satisfactorios resultados.

Por su parte, también se ha utilizado con éxito el modelado conjunto de varias fuentes de información, *streams*. El fundamento de este método reside en utilizar conjuntamente distintas fuentes de información de modo que los errores de alguna de ellas se puedan subsanar con las otras, para lo que habrá que elegir adecuadamente dichas fuentes de información. Así pues, se distinguen tantas técnicas de modelado conjunto como naturalezas de las señales que se pretenda combinar. De este modo, por ejemplo se puede hablar principalmente de fusión de: información de subbandas frecuenciales, binaural, de distintas parametrizaciones y audiovisual. El incluir un modelado independiente para cada subbanda frecuencial de la señal de voz [Bourlard *et al.*, 1996] [Bourlard *et al.*, 1972] busca dotar al sistema de RAH de robustez ante ruidos de banda estrecha, de modo que prevalezcan aquellas subbandas menos afectadas por el ruido sobre las más contaminadas. Sin embargo, las subbandas frecuencias no son independientes, por lo que se obtienen mejores resultados si la fusión se realiza teniendo en cuenta la correlación entre ellas [Morris *et al.*, 1999] [Hagen, 2000]. Basándose en la idea de la percepción humana, el uso conjunto de señales de voz registradas a partir de dos (binaural) o más sensores también se ha mostrado efectivo bajo ciertas condiciones [Wittkop, 2001] [Kleinschmidt, 2002]. La motivación de la fusión de parametrizaciones reside en intentar aprovechar los distintos puntos fuertes de cada una de ellas a la vez que se trata de eludir sus debilidades [Pujol *et al.*, 2003], asimismo se pueden incorporar igualmente vectores de características asociados a distintas escalas temporales [Hagen and Bourlard, 2000], ya que en esos casos existe información incluso de niveles más altos que el puramente acústico-fonético [Hermansky, 1998] [Segura *et al.*, 2001]. Incorporar información visual a la señal de voz es otra clase de fusión que ya desde los primeros tiempos se planteó [Sumby and Pollack, 1954]; sin embargo hasta el momento no se han obtenido los resultados deseados ya que la lectura de labios aún no proporciona una aceptable tasa de acierto si no es bajo condiciones muy controladas [Potamianos *et al.*, 2003].

Los modelos de deformación elástica son otra de las extensiones del modelado acústico básico basado en HMMs que originariamente se desarrolló para el tratamiento de imágenes [Rueckert *et al.*, 2001], pero que ha sido aplicada satisfactoriamente a sistemas de RAH en el dominio tiempo-frecuencia de la señal de voz. Dicho dominio se ve como una matriz que sufre deformaciones locales debido

a causas fuera del alcance de otros modelos y que tratan de representarse con esta nueva extensión [Uchida and Sakoe, 1998] [Kybic and Unser, 2003]. Parcialmente relacionados con los modelos de deformación elástica, los modelos de normalización del tracto vocal [Wakita, 1977] tratan de normalizar la escala frecuencial de manera que se compensen diferencias entre locutores, que son en muchos casos causantes de importantes errores de RAH. Para llevar a cabo esta idea se han desarrollado a lo largo del tiempo diversas realizaciones a partir de distintos puntos de vista [Andreou *et al.*, 1994] [Lee and Rose, 1998] [Pitz, 2005].

Además de las extemsiones consideradas anteriormente, cabe destacar como el mejor conocimiento de los HMMs [Ghahramani and Jordan, 1997] [Ghahramani and Beal, 2000], unido a la aparición de nuevas y más completas teorías sobre el aprendizaje estadístico, han hecho que el modelado acústico sea actualmente una prometedora línea de investigación que proporciona cada día mejores y más completas extensiones de los clásicos HMMs aprovechando las cualidades de la señal de voz [Jordan and Jacobs, 1994] [Digalakis, 1992] [Patil and Taillie, 2000] [Jacobs *et al.*, 2002].

COMENTAR REDES NEURONALES?

2.4. Modelado de Lenguaje.

El modelado de lenguaje tiene como objetivo incorporar el conocimiento lingüístico a los sistemas de RAH, incluyendo aspectos como el léxico, la semántica y la gramática, de modo que se incluyan las restricciones propias que existen en el modo en que se concatenan las palabras para una determinada tarea de reconocimiento [Lea, 1979] [Lee *et al.*, 1989] [Pieraccini *et al.*, 1992] [Ward and Young, 1993]. Esto tiene su traducción matemática en los sistemas de RAH basados en aproximación estadística Bayesiana en el cálculo de la probabilidad a priori de las distintas secuencias de palabras, $p(\mathbf{W})$.

Al igual que sucedía en el modelado acústico, modelar probabilísticamente una secuencia de palabras de un modo completo puede resultar inviable por cuestiones computacionales, por lo que se limita la dependencia entre vocablos próximos. Así, suponiendo que las secuencias de palabras siguen un proceso de Markov de orden $(n-1)$ [van Kampen, 1992], la probabilidad de una palabra dependerá sólo de las $(n-1)$ anteriores y no de toda la historia previa. A este tipo de modelo de lenguaje se le denomina N-gramas [Jelinek *et al.*, 1975] y actualmente son los modelos de lenguaje más utilizados, aunque no los únicos.

QUIZÁS ESQUEMA DE LAS N-GRAMAS

Las N-gramas pueden entrenarse, de igual modo que el modelado acústico, mediante el criterio de máxima verosimilitud, ML, utilizando la perplejidad como criterio de evaluación [Bahl *et al.*, 1983]. Para ello sólo se requieren bases de datos de texto y no de audio, lo que es una gran ventaja por ser más fácilmente accesibles, aunque si se incrementa el orden del modelo (n), buscando con ello una mayor especificidad del modelado, el tamaño de la base de datos también se debe incrementar considerablemente, pudiéndose dar en muchas ocasiones la ausencia de varias de las posibles agrupaciones de n palabras, lo que repercutiría en el entrenamiento de un modelo erróneo. Para solventar este problema se suelen emplear métodos de suavizado, *smoothing*, [Martin *et al.*, 1999] en los que los parámetros de los modelos de las unidades problemáticas se estiman a partir de los de orden menor $(n-1,\ n-2,...)$. Este procedimiento se puede llevar a cabo de distintos modos: *discounting*, *co-ocurrence*, *backing off*, o categorizando las palabras en clases más amplias y, por

tanto, más comunes [Katz, 1987] [Ney and Essen, 1991] [Ney *et al.*, 1994] [Kuhn *et al.*, 1994].

A su vez, a lo largo del tiempo se han ido desarrollando distintas técnicas para mejorar el comportamiento de los modelos de lenguaje basados en N-gramas. Así, por ejemplo se puede nombrar *language model cache* [Kuhn and de Mori, 1990], que emplea las últimas palabras reconocidas para adaptar el modelo de lenguaje a la tarea de RAH en cuestión, lográndose de este modo una mayor especificidad. Otra mejora consiste en agrupar palabras que aparecen habitualmente en el mismo orden para tratarlas como si de una única unidad se tratara [Jelinek, 1991]. Una idea similar a la anterior, y ya comentada como un método de *smoothing*, consiste en agrupar palabras en clases atendiendo a un criterio concreto de modo que se robustece la estimación de los modelos de lenguaje a la vez que se reduce la cantidad de datos necesarios para entrenarlos [Brown *et al.*, 1992].

Como consecuencia de la incapacidad práctica de las N-gramas para aprender restricciones propias del lenguaje que requieren de una gran memoria, se desarrollaron las gramáticas de estados finitos, en las que se determinan las posibles concatenaciones de las palabras mediante reglas [Ney, 1990] [Wrigley and Wright, 1991] [Fred and Leitao, 1994]. De este modo, este tipo de gramáticas, si bien más potentes que las N-gramas y útiles en entornos restringidos, tienen el serio inconveniente de que ante tareas complejas su manejo puede llegar a ser inviable por la complejidad de los árboles necesarios. A su vez, y aunque las probabilidades de las reglas pueden calcularse de un modo automático, la generación previa de las propias reglas suele ser un proceso manual ya que hasta la fecha los sistemas de generación automática de las mismas no están muy desarrollados [Segarra and García, 1991] [Cerf-Danon and El-Béze, 1991]. Por todo ello las gramáticas de estados finitos no están tan extendidas como las N-gramas entre los sistemas de RAH. Por otra parte, y dado que las gramáticas de estados finitos tampoco son capaces de modelar todos los aspectos del lenguaje natural, se han desarrollado modelos más complejos, como las gramáticas transformativas o de unificación [Morgenthaler and Hansen, 1982] [Pereira and Warren, 1980] [Shieber, 1985] que, al igual que los modelos de lenguaje de estados finitos, necesitan de procesos manuales para generar las reglas, lo que hace que tampoco supongan actualmente una opción muy utilizada en los sistemas de RAH, prefiriéndose emplear en la mayoría de los casos N-gramas como modelado de lenguaje.

2.5. Búsqueda.

La tarea de búsqueda consiste en encontrar la secuencia de palabras que maximice la expresión (2.5) que viene dada, como ya se ha comentado, por el producto de los modelos acústico y de lenguaje. Una primera e hipotética aproximación a la solución consistiría en evaluar dicho producto o verosimilitud para todas las posibles secuencias de palabras y elegir aquella que mayor valor obtuviera. Sin embargo rápidamente se puede comprobar que esta opción, salvo para tareas muy sencillas, debe desecharse por la complejidad de cálculo, que aumenta exponencialmente con el número de las posibles palabras, $\mathbf{W}$.

La complejidad de la optimización de la expresión (2.5) se puede reducir drásticamente mediante programación dinámica [Bellman, 1957], que descompone el problema inicial en una serie de subproblemas de optimizaciones locales aprovechando la estructura matemática del mismo. Dentro de la programación dinámica, dos algoritmos se han hecho populares entre los sistemas de RAH: *stack decoding* [Jelinek, 1969] y el de Viterbi [Viterbi, 1967] [Vintsyuk, 1971]. La búsqueda mediante la primera de las técnicas, *stack decoding*, se suele implementar mediante el uso de una pila que mantiene para cada instante de tiempo una lista ordenada con los estados hipotéticos que podrían haber generado el correspondiente vector de características; una vez obtenida la lista, las proyeccio-

nes desde cada estado se realizan asíncronamente con el tiempo hacia otra pequeña lista de estados elegida heurísticamente, lo que hace que el resultado final de esta técnica dependa en gran medida de dicha estimación heurística, lo que no deja de ser un inconveniente. Por el contrario, en el algoritmo de Viterbi los hipotéticos estados se proyectan síncronamente con el tiempo, lo que permite que para cada vector de características se pueda comparar la verosimilitud para todos los posibles estados, haciendo posible el uso de métodos de *pruning* que minimizan todavía más la complejidad de cálculo de la optimización. Los métodos de *pruning* consisten en reducir el número de estados sobre los que proyectar a la hora de realizar la búsqueda mediante el algoritmo de Viterbi, de modo que únicamente los hipotéticos estados que previsiblemente van a formar parte de la secuencia de palabras óptima se activan (*beam search* [Ney *et al.*, 1987] [Ortmanns and Ney, 1995]). Si bien la reducción de cómputo es clara, también se puede dar el hecho de que la secuencia de palabras más probable se llegue a desechar antes de completar el proceso de decodificación, pero éste no es un hecho muy frecuente si se ajustan adecuadamente los parámetros que rigen los distintos métodos de *pruning*.

QUIZÁS ESQUEMA Y FUNCIONAMIENTO DE VITERBI

Asimismo, cabe destacar que se pueden aplicar diversas técnicas para incrementar la eficiencia de los métodos de *pruning*, como *language model look-ahead* [Steinbiss *et al.*, 1994], en el que la pronunciación del léxico se organiza mediante un árbol, de modo que se acota el final de las posibles palabras en cada nodo del árbol pudiéndose propagar hacia atrás la estimación del modelo de lenguaje. Además, si el sistema de RAH tolera un pequeño desfase temporal, se puede calcular la contribución del modelo acústico para unos pocos vectores de características siguientes al que se está decodificando mediante modelos simplificados [Ney *et al.*, 1992], lo que ayuda a reducir más aún el tiempo de búsqueda.

Además de las técnicas vistas anteriormente, y dado que el cálculo de la verosimilitud asociada a cada estado del modelo acústico suele influir de un modo crucial en el coste computacional final, se han venido desarrollando distintos métodos para agilizar dicho cálculo. Esto se puede conseguir por ejemplo mediante la estructuración del espacio de búsqueda [Fritsch, 1997], la cuantización de los vectores de características [Bocchieri, 1993], o la partición del espacio de los vectores de observaciones acústicas [Nene and Nayar, 1996]. También se consigue una importante reducción del coste computacional paralelizando la cálculo de la verosimilitud mediante instrucciones SIMD, *Single Instruction Multiple Data* [Kanthak *et al.*, 2000].

Por otra parte, la decodificación en varias pasadas, aunque con el inconveniente de no poder ser aplicada en tiempo real, se presenta como una técnica para agilizar el proceso de búsqueda al obtener la secuencia final de palabras reconocida como producto de varias iteraciones. Así, inicialmente se utilizan modelos acústicos y de lenguaje más simplificados, que proporcinan no sólo la secuencia de palabras más probable, sino el conjunto de las N más probables, *N-best*, [Schwartz and Chow, 1990], o bien un grafo de palabras [Schwartz and Austin, 1991]. Posteriormente, las siguientes iteraciones del sistema de RAH se realizarán únicamente sobre estos resultados ya con modelos acústicos y de lenguaje cada vez más restrictivos pero aplicados sobre un espacio de búsqueda mucho más reducido.

3

Robustez en Reconocimiento Automático del Habla.

En general, los sistemas de RAH ofrecen, en cuanto a tasa de reconocimiento se refiere, unos resultados aceptables siempre que se den ciertas condiciones controladas que afectan a todos y cada uno de los ámbitos de los mismos. Una de dichas condiciones deseables consiste en que tanto el modelado acústico como la extracción de características en el momento del RAH se realicen bajo ausencia de ruido. Sin embargo, en una situación ordinaria no se suele dar esta cicunstancia por lo que se ha de recurrir a técnicas de robustez para compensar el correspondiente desajuste.

Si se revisan los elementos de que se compone un sistema de RAH (ver Capítulo 2), y dejando a un lado el modelado del lenguaje, que depende de la tarea en cuestión, y el sistema de búsqueda, que puede considerarse fijo, se puede concluir que principalmente las técnicas de robustez pueden actuar bien sobre el modelado acústico, bien sobre la parametrización o extracción de características. Así, se distinguen tres tipos distintos de métodos de robustez [Gong, 1995] [Bellegarda, 1997], a saber, extracción robusta de características, de modo que los vectores acústicos se vean lo menos afectados posible por el ruido, haciendo por tanto que el desajuste entre ellos y los modelos acústicos entrenados en las condiciones de referencia sea mínimo; la segunda actuación posible se denomina adaptación de modelos acústicos y pretende transformar éstos últimos con la finalidad de acercarlos a las condiciones con que se han extraído los vectores de características que se trata de reconocer; por último la tercera actuación posible es la adaptación de los vectores de características o normalización, que propone la solución inversa a la solución anterior, esto es, adecuar los vectores de características a los modelos acústicos entrenados en las condiciones de referencia. Cabe resaltar llegados a este punto que, en muchas ocasiones, la normalización de los vectores de características puede verse, de un modo general y si el modelado acústico se realiza mediante HMMs con GMMs como funciones de densidad de probabilidad asociadas a cada estado, que es el caso más habitual, como una adaptación de modelos acústicos en la que se recalculan todos los vectores de medias para cada trama, de modo que finalmente la probabilidad a posteriori de las distintas Gaussianas dado el vector acústico es el misma en ambos casos. Además de las tres líneas de actuación nombradas se puede pensar en soluciones híbridas [Nolazco and Young, 1994] [Sankar and Lee, 1996], que son simplemente combinaciones de las anteriores.

QUIZÁS INCLUIR FIGURA/ESQUEMA DE LAS 3 TÉCNICAS.

Por lo general, se suele considerar que la adaptación de modelos acústicos proporciona mejores tasas de reconocimiento que cualquiera de las otros tipos de técnicas de robustez [Neumeyer and Weintraub, 1995]

por cuanto puede representar estadísticamente la aleatoriedad del ruido, que es en último término la causante de la incertidumbre entre las correspondientes realizaciones ruidosas y limpias y, por tanto, también de los errores de RAH; sin embargo, la adaptación de modelos acústicos precisa de más datos y tiempo de computación que otros métodos, por lo que la decisión final de cara a la utilización de un tipo de algoritmo de robustez u otro dependerá en gran medida de las características y limitaciones de la aplicación concreta que se pretenda realizar en cada caso.

Este Capítulo, basado en las distintas técnicas de robustez empleadas en los sistemas de RAH, se estructura del siguiente modo: en la Sección 3.1 se resumen brevemente los métodos más empleados en la extracción robusta de características. Los algoritmos más sobresalientes enmarcados en la adaptación de modelados acústicos se estudian en la Sección 3.2. Finalmente, y ya en la Sección 3.3, se enumeran y comentan brevemente aquellos métodos incluidos en la normalización de vectores de caracetrísticas más utilizados en la actualidad.

3.1. Extracción Robusta de Características.

Como ya se ha adelantado, uno de los puntos fundamentales sobre los que se puede actuar para proporcionar robustez a cualquier sistema de RAH, es la adecuada elección del conjunto de parámetros que compongan los vectores de características empleados para representar la señal de voz. Los algoritmos que forman parte de esta línea de actuación asumen que son inmunes al ruido, de modo que no necesitarían hipotéticamente ningún otro método para proporcionar robustez adicional al sistema final de RAH. Así pues, con objeto de obtener técnicas de extracción de características que se vean lo menos afectadas posible por el ruido, se han investigado soluciones como la utilización de ventanas de *liftering*, distancias basadas en la proyección cepstral, parametrizaciones obtenidas mediante criterios discriminativos o mediante el procesado en sub-bandas. Cabe resaltar que en este apartado no se mencionarán, por considerarlos actualmente estándares de facto en muchos de los sistemas de RAH, aquellas técnicas que, aunque nacidas para proporcionar robustez, están basadas en el modelo auditivo humano o en el cepstrum en escala Mel (ver Sección 2.2). Igualmente, y por la misma razón previamente esgrimida, la inclusión de las características dinámicas en el vector final de características tampoco se menciona en este apartado, aunque sí lo fue igualmente en la Sección 2.2.

La idea del empleo de ventanas de *liftering* se basa en que generalmente el ruido no afecta del mismo modo a todos los coeficientes cepstrales. De esta manera, dichas ventanas pueden realzar aquellos coeficientes menos sensibles al ruido, a la vez que reduce la importancia del resto, lográndose así un mejor comportamiento ante entornos acústicos adversos; por ejemplo, en los coeficientes cepstrum-LPC, son los de orden menor los que, generalmente, más afectados se encuentran por el ruido, de modo que se podría aplicar ventanas de liftering del tipo seno remontado [Juang *et al.*, 1987] o *general exponential lifter* [Junqua and Wakita, 1989] para reducir la importancia final de dichos coeficientes.

El fundamento del uso de las técnicas de proyección cepstral se encuentra en que uno de los efectos más importantes del ruido blanco es la reducción de la norma de los vectores cepstrales. Así, la medida de la proyección cepstral enfatiza los picos espectrales de energía, que se ven menos afectados por el ruido, haciendo que la distancia basada en proyección cepstral sea una medida más robusta que las distancias euclídeas [Carlson and Clements, 1991] [Junqua and Haton, 1996]. Desgraciadamente, si bien este tipo de técnicas son eficaces con ruido blanco, no lo son tanto con otros tipos de ruidos [Herrando and Nadeu, 1994].

Las parametrizaciones discriminativas se caracterizan por enfatizar las características de la señal de voz que sean más útiles para separar en clases, proporcionando así la robustez deseada. De entre todas las técnicas, es el análisis lineal discriminativo, *Discriminative Linear Analysis*, LDA, [Duda and Hart, 1973] [Fukunaga, 1990] el más extendido, habiéndose empleado con éxito en sistemas de RAH mediante distintas aproximaciones y métodos [Yang *et al.*, 2000] [Segura *et al.*, 2001].

Si el ruido es de banda estrecha, hecho este que en algunos entornos se puede asumir, se puede emplear la parametrización mediante procesado de sub-bandas [Hermansky *et al.*, 1996] [Bourlard *et al.*, 1972], que comprende un conjunto de técnicas que están relacionadas con el dominio frecuencial de la señal de voz y que se basan en ponderar en mayor medida aquellas bandas frecuenciales libres de ruido, mientras que las que se ven más afectadas pasan a un segundo plano a la hora de obtener el vector de características final. Desde la aparición de este tipo de parametrizaciones, muchas han sido las contribuciones, la mayoría de las cuales con la intención de combinar la información no corrupta por el ruido [Hagen, 2000] [Hagen and Bourlard, 2000]. Cabe destacar que esta aproximación de procesado en sub-bandas está muy relacionada con las técnicas de *missing data* [Morris *et al.*, 2001] [Cook *et al.*, 2001], en las que se utilizan técnicas de marginalización para reconstruir una función de densidad de probabilidad y generar con ella las observaciones eliminando la dependencia con las variables aleatorias supuestamente corruptas por el ruido [Vizinho *et al.*, 1999] [Josifovski, 2002].

QUIZÁS COMO PARAMETRIZACIÓN ROBUSTA MÁS EMPLEADA: ETSI ADV

3.2. Adaptación de Modelos Acústicos.

Las técnicas de adaptación de modelos acústicos se han usado profusamente hasta la fecha para tratar de compensar la variación estadística que el ruido introduce en la señal de voz. En general, este tipo de algoritmos se emplean cuando no se dispone de la suficiente cantidad de datos como para entrenar los modelos acústicos bajo las condiciones específicas con las que se pretende reconocer. Aunque en la mayoría de los casos es materialmente imposible obtener tal cantidad de datos de una manera natural mediante grabaciones, bien es cierto que se puede contaminar la señal limpia con ruido del entorno acústico concreto [Bippus *et al.*, 1999] [Matassoni *et al.*, 2001] [Moreno, 1996], pero esta última opción, si bien más viable que la grabación del corpus correspondiente, tampoco proporcionaría unos modelos perfectos ya que no quedaría reflejado en ellos ciertas alteraciones de la voz producidas por el estrés, el ruido... (efecto Lombard [Lombard, 1911]). Por todo ello, se suele recurrir en la mayoría de los casos a la adaptación para obtener unos modelos acústicos que representen estadísticamente la señal de voz bajo las condiciones concretas de reconocimiento. En cualquier caso, y del mismo modo que en cualquier fase de entrenamiento, la eficacia de este tipo de algoritmos está supeditada a la similitud estadística entre la señal empleada para adaptar los modelos acústicos y la que posteriormente se reconocerá.

De entre todas las técnicas de adaptación de modelos acústicos, las más empleadas actualmente son MAP, *Maximum A Posteriori*, MLLR, *Maximum Likelihood Linear Regression*, PMC, *Parallel Model Component*, JA, *Jacobian Adaptation*, VTS, *Vector Taylor Series*, para adaptación de modelos acústicos (hay una versión que normaliza los vectores de características) y selección de modelos acústicos. A pesar de que en las siguientes subsecciones se considere cada una de ellas de modo independientemente, es interesante indicar que en muchas ocasiones los distintos métodos se utilizan conjuntamente tratando así de mantener los puntos fuertes de cada una de ellos, a la vez que se procura minimizar los débiles.

3.2.1. MAP, *Maximum A Posteriori.*

En general, tal y como se ha comentado en la Sección 2.3, a la hora de estimar los parámetros que definen los modelos acústicos (en el caso de emplear HMMs, probabilidades de transición entre estados y las correspondientes pdfs asociadas a cada estado), se suele recurrir a la estimación de máxima verosimilitud, ML. Sin embargo, en la técnica MAP, *Maximum A Posteriori*, [Gauvain and Lee, 1994], se emplea la estimación a priori, que consiste en suponer que los parámetros en cuestión son variables aleatorias con una conocida función de densidad de probabilidad, incluyendo de este modo un conocimiento a priori de los parámetros. Cabe reseñar que, para el caso de modelos acústicos basados en HMMs, si esta supuesta pdf se considera homogénea, las expresiones que definen los diferentes parámetros con los criterios ML y MAP coinciden.

A partir de lo anterior se puede concluir que la clave del buen funcionamiento de la técnica MAP recae sobre la adecuada elección de la pdf a priori, que suele tomarse como Normal-Wishart para poder posteriormente reestimar de forma sencilla los distintos parámetros de los modelos acústicos. Con esta suposició se puede comprobar que los vectores de caracterísicas se acaban modelando como una mezcla de Gaussianas, que es el caso más extensivamente usado en los sistemas de RAH.

Cabe resaltar que, en general, los resultados de RAH obtenidos mediante la estimación MAP son mejores que los logrados con el criterio ML, especialmente cuando la cantidad de datos disponibles es aceptable. A su vez, y para solventar algunos de las debilidades que posee la técnica MAP, se han propuesto ciertas aproximaciones, así existe una versión *on line* [Huo and Lee, 1997], o se puede reducir el número de datos necesarios si se emplea el conocimiento a priori de la función de correlación entre los diversos parámetros de los modelos acústicos. A esta última extensión se la denomina EMAP, *Extended* MAP [Stern and Larsy, 1987] [Zavaliagkos *et al.*, 1995].

3.2.2. MLLR, *Maximum Likelihood Linear Regression.*

Suponiendo nuevamente que los modelos acústicos se componen mediante HMMs con GMMs como pdfs asociadas a cada estado, la versión más extendida del algoritmo MLLR, *Maximum Likelihood Linear Regression*, es aquella en la que únicamente se modifican mediante una función afín los vectores de medias de las Gaussianas asociadas a cada estado de los HMMs [Legetter and Woodland, 1995]; de todos modos también existe la posibilidad de compensar igualmente las matrices de covarianzas [Gales, 1997b], lo que en general supone un mayor coste computacional sin que por ello se logre una mejora considerable. En ambos casos los nuevos parámetros de los modelos acústicos se calculan mediante el estimador ML haciendo uso del algoritmo EM [Dempster *et al.*, 1977].

En muchas ocasiones las mezclas de Gaussianas que forman parte de las pdfs asociadas a los estados de los HMMs se suelen agrupar en clases de regresión, de modo que cada una de ellas se adapta haciendo uso de la misma función afín. Esto, que indudablemente es una importante ventaja ya que puede reducir sensiblemente el número de funciones afines que se precisa estimar, se convierte en un inconveniente cuando el número de clases de regresión es extremadamente reducido ya que se pierde especificidad debido a que las funciones afines pasan a ser demasiado generales, resintiéndose así las tasas de RAH. Por tanto, el punto clave para lograr buenos resultados empleando pocos datos con la técnica MLLR viene de la correcta elección de las clases de regresión, lo que no deja de plantear un compromiso.

Cabe destacar que, basándose en la misma transformación lineal propuesta para modificar las medias de las Gaussianas que componen las pdfs de los HMMs, se puede emplear el estimador MAP

en lugar del ML, lo que da lugar a la técnica MAPLR, *Maximum A Posteriori Linear Regression* [Chesta *et al.*, 1999], que proporciona unas tasas de RAH algo mejores que los obtenidos por el método MLLR.

3.2.3. PMC, *Parallel Model Component.*

PMC, *Parallel Model Component*, [Gales and Young, 1993] [Gales, 1995] [Gales, 1997a] es una técnica que obtiene los nuevos modelos acústicos combiando los asociados al espacio de referencia y los correspondientes al ruido que se ha localizado en el entorno acústico concreto en el que se pretende reconocer.

A partir de lo anterior se puede concluir que el punto clave en este método es el modo en que se combinan los modelos, lo que viene definido por la función de desajuste, *mismatch function*. En este sentido, y dado que normalmente se suele suponer que el ruido es aditivo en el dominio temporal, la combinación de los parámetros de los modelos suele realizarse por comodidad en el espacio espectral, ya sea logarítmico o lineal, aplicando la correspondiente función de desajuste inversa.

Este algoritmo, si bien probado con éxito en multitud de circunstancias, se sustenta, como se ha podido apreciar, en la presunción de una determinada función de desajuste, de modo que si ésta no se corresponde con la real, puede desembocar en el cálculo de unos modelos acústicos erróneos. Asimismo, el coste computacional es extremadamente elevado ya que se debe transformar del espacio cepstral al espectral y viceversa. Para subsanar en la medida de lo posible este inconveniente se han venido planteando diversas aproximaciones que pretenden reducir el tiempo necesario para llevar a cabo la adaptación [Gales, 1995].

3.2.4. JA, *Jacobian Adaptation.*

JA, *Jacobian Adaptation*, [Yapanel *et al.*, 2002] es un algoritmo que proporciona una eficiente solución a la hora de adaptar modelos acústicos a un no muy elevado coste computacional bajo la premisa de que las diferencias acústicas entre el espacio de reconocimiento y el de referencia, objetivo deseado, sean pequeñas.

Esta técnica requiere de cuatro pasos, a saber: entrenar los modelos acústicos de referencia, a partir de los cuales se construirán los nuevos ya adaptados a las nuevas condiciones, calcular las matrices jacobianas en el dominio cepstral, para lo que habrá que suponer un determinado modelo de degradación de los vectores de características (igual que para el algoritmo PMC), el tercer paso consiste en obtener una estimación de la diferencia del ruido entre el espacio de reconocimiento y el de referencia, y, ya por último, a partir de las matrices jacobianas y la estimación de la diferencia del ruido, se obtienen los nuevos vectores de medias y matrices de covarianzas mediante una función lineal con término independiente nulo.

De lo anterior se puede concluir que la adaptación Jacobiana se sustenta en tres aproximaciones: reestimar los nuevos parámetros de los modelos acústicos a partir de los del espacio de referencia mediante únicamente una función lineal sin término independiente, presuponer un modelo de degradación de los vectores de características entre el espacio de reconocimiento y el de referencia y, finalmente, suponer que sólo hay una pequeña degradación acústica entre ambos espacios. Todo ello hay que tenerlo en cuenta a la hora de elegir el algoritmo JA para adaptar los modelos acústicos.

Dado que en muchas situaciones no se dan las aproximaciones anteriores, se han propuesto a lo largo del tiempo diversas mejoras a la adaptación JA. Así, en el caso de que el desajuste entre los espacios sea elevado, se puede partir de un *clustering* de modelos acústicos iniciales [Shimodaira *et al.*, 2000], de modo que se elija entre ellos el que mejor se adecúe a la situación que se pretende compensar, o bien se puede descomponer el ruido en tres términos para modelar el efecto aditivo, la distorsión convolucional y la dependencia de cada locutor [Shimodaira *et al.*, 2002]. También para incrementar el rango de acción de la adaptación Jacobiana y reducir el número de matrices jacobianas necesarias se puede enfatizar el espectro del ruido para mejorar la aproximación lineal a la vez que se realiza un *clustering* de matrices Jacobianas [Cerisana *et al.*, 2000] [Sarikaya and Hansen, 2000]. De este modo, para cada una de las mejoras anteriores se cumplen más estrictamente las aproximaciones anteriormente comentadas ampliando de esta manera el rango de acción de la técnica.

3.2.5. VTS, *Vector Taylor Series* para adaptación de modelos acústicos.

La técnica VTS, *Vector Taylor Series*, para adaptación de modelos acústicos [Moreno, 1996] [Kim *et al.*, 1998] [Acero *et al.*, 2000] presupone un modelo de degradación de la señal acústica, normalmente constituido por un filtro más un término aditivo [Acero, 1990], y cuyos parámetros en el dominio cepstral se proximan mediante una serie de Taylor que, generalmente, se suele truncar hasta orden uno. Una vez estimados los parámetros del modelo de degradación, y aplicando la correspondiente función inversa, se procede a modificar los parámetros de los modelos acústicos.

Esta técnica, al igual que todas las que presuponen la existencia de un modelo de degradación, tiene el inconveniente de basar todas sus espectativas de mejora en dicho modelo, de modo que si no se corresponde con el real, los resultados no serán los esperados. Por otra parte, cabe destacar que el método de adaptación Jacobiana anteriormente comentada, JA, no deja de ser en cierto modo un caso especial y más sencillo de la adaptación mediante la técnica VTS con polinomio de orden uno.

3.2.6. Selección de modelos acústicos.

En algunas ocasiones, la variabilidad del espacio de reconocimiento es demasiado elevada como para recurrir a los algoritmos básicos de adaptación tratados anteriormente. En estos casos proporciona mejores resultados poseer diversos modelos acústicos que representen de la mejor manera posible los distintos subentornos básicos que constituyen el espacio de reconocimiento y elegir en cada momento aquél que mejor se adecúe.

Por otra parte, y basándose también en la idea de selección de modelos acústicos, se han desarrollado técnicas basadas en autovoces, *eigenvoices*, en las que se obtienen los modelos acústicos adaptados a la nueva condición a partir de una combinación lineal de los asociados a diversos subentornos básicos [Kuhn *et al.*, 2000] mediante PCA, *Principle Component Analysis*. Como los modelos acústicos poseen generalmente una gran cantidad de parámetros, la elección de los que entrarán en el cálculo del análisis PCA es crítica, considerándose en la mayoría de los casos únicamente los vectores de medias de las Gaussianas que componen las pdfs asociadas a cada estado de los HMMs. Cabe resaltar que este método se emplea principalmente cuando la cantidad de datos de los que se dispone es extremadamente pequeña.

3.3. Normalización de Vectores de Características.

La tercera línea de acción considerada a la hora de proporcionar robustez a un sistema de RAH propone modificar los vectores de características acercándolos estadísticamente a los modelos acústicos con los que se pretende reconocer. En principio, esta opción, tal y como se ha indicado anteriormente, no puede compensar el efecto de la aleatoriedad del ruido, causante de la alta incertidumbre entre las señales limpias y ruidosas, del mismo modo que las técnicas de adaptación de los modelos acústicos. Sin embargo, el menor coste computacional y la reducida cantidad de datos necesarios para proporcionar interesantes resultados hacen actualmente de las técnicas de normalización de vectores de características uno de los modos más empleados a la hora de proporcionar robustez a un sistema de RAH. Los algoritmos incluidos dentro de esta línea pueden agruparse en tres grandes clases [Stern *et al.*, 1997], a saber: filtrado paso alto, *high-pass filtering*, basados en modelos, *model-based*, y empíricos, *empirical*. A continuación se trata por separado cada una de estas clases, haciendo especial hincapié en los algoritmos más representativos en cada caso.

3.3.1. Filtrado paso alto.

Dentro de los métodos de normalización de vectores de características basados en filtrado paso alto se incluyen técnicas por lo general bastante sencillas que, si bien no pueden competir en cuanto a prestaciones con otros algoritmos de normalización, sí pueden hacerlo atendiendo al coste computacional, que es mínimo. Por todo ello, en muchas ocasiones se llegan a considerar como un estándar de facto, incluyéndose en la mayoría de los sistemas de RAH. Así, englobados dentro de esta clase, se pueden encontrar métodos como CMN, *Cepstral Mean Normalization*, también conocido como CMS, *Cepstral Mean Substraction*, el procesamiento RASTA, *RelAtive SpecTral Amplitude*, o técnicas clásicas de filtrado.

El algoritmo CMN [Hanai and Stern, 1994] [Yapanel *et al.*, 2002] [de la Torre *et al.*, 2001] consiste en un filtrado paso alto sobre los coeficientes cepstrales. Para ello, en su versión más sencilla, se sustrae a cada trama el vector de características medio visto hasta el momento. Por su parte, el procesamiento RASTA [Hermansky and Morgan, 1994] consiste en un filtrado paso alto, o paso banda, aplicado en el dominino log-espectral [Hermansky *et al.*, 1991] [Hermansky *et al.*, 1993] o en el cepstral [Mokbel *et al.*, 1993]. En cualquiera de las dos técnicas se pretende compensar principalmente los efectos de la distorsión convolucional, ya que ambas se basan principalmente en dos aproximaciones: considerar que la respuesta impulsional de un filtro afecta de forma aditiva en el dominio cepstral, y suponer que dicha respuesta impulsional es invariante en el tiempo. Conforme estas dos aproximaciones se acerquen a la realidad los algoritmos proporcionarán mejores resultados. Conviene recordar llegados a este punto que en el dominio Mel-cepstrum no se da la primera de las aproximaciones debido al enventanamiento previo que se realiza sobre la señal de voz para proporcionar estacionalidad. Por otra parte, en muchas ocasiones tampoco la invariabilidad temporal de la respuesta impulsional que define la distorsión convolucional puede considerarse una aproximación válida.

Asimismo, y dado que los vectores de características de la señal de voz y de silencio son bastantes distintos entre sí, se ha propuesto una extensión de la versión clásica del método CMN que consiste en calcular independientemente el vector de características medio para cada uno de los dos casos (voz y silencio) y sustraer el que corresponda en cada momento [Acero and Huang, 1995], obteniéndose así una ligera mejora sobre el método clásico.

QUIZÁS NOMBRAR NORMALIZACIÓN DE MÁS ÓRDENES.

Por su parte, las técnicas clásicas de filtrado en el dominio temporal [Meyer and Simmer, 1997] pueden ser muy útiles en aquellas situaciones en las que el ruido predominante es de banda limitada [Mokbel and Cholet, 1995]. De este modo, y para estas condiciones concretas, se pueden utilizan filtros paso banda en los que se adaptan las frecuencias de corte atendiendo en cada momento a la naturaleza del ruido existente [Hoshino, 2001]; también se pueden emplear filtros de Wiener tras captar la señal mediante un array de micrófonos [Meyer and Simmer, 1997].

3.3.2. Técnicas basadas en modelos.

Las técnicas de normalización de vectos de características basadas en modelos se sustentan en considerar que el desajuste entre los espacios de entrenamiento y de reconocimiento se puede representar mediante un modelo de degradación. Una vez definido dicho modelo se estiman los parámetros que lo representan convenientemente y, tras aplicar de un modo apropiado la correspondiente función inversa, se transforman los vectores de características. El éxito de este tipo de técnicas depende de hasta qué punto el modelo propuesto se acerca al real, por lo general siempre más complejo, así como de la precisión con que se logre estimar los parámetros que lo definen. De las técnicas de normalización basadas en modelos desarrolladas casi todas consideran sólo dos tipos de degradación, a saber: suponer que la señal ruidosa es la suma de la limpia y un ruido aditivo, o bien modelar la señal contaminada como la limpia afectada por un ruido aditivo y un filtro [Acero, 1990], aproximación esta algo más completa y realista.

Dentro de los algoritmos de normalización basados en modelos que consideran la degradación entre la señal limpia y la ruidosa como la suma de las componentes del ruido aditivo y de la distorsión convolucional, los más representativos son VTS, *Vector Taylor Series*, para normalización [Kim *et al.*, 1998] [Moreno, 1996] CDCN, *Codeword Dependent Cepstral Normalization*, [Acero, 1990] y VPS, *Vector Polynomial Approximations*, [Stern *et al.*, 1997] [Raj *et al.*, 1996]. Por otra parte, y dentro de aquellos métodos que modelan la señal contaminada como la limpia afectada únicamente por ruido aditivo, se encuentran MMSE, *Minimum Mean Square Error*, [Yapanel *et al.*, 2002] [Matassoni *et al.*, 2002] [Ephraim and Malah, 1985] StatComp, *Statistical Compensation*, [de la Torre *et al.*, 2001], ecualización del ruido, *noise equalization*, [Gelin and Junqua, 1999], o sustracción espectral, *Spectral Subtraction, SS* [Boll, 1979] [Lockwood and Boudy, 1992] [Nolazco and Young, 1994].

QUIZÁS PONER TAMBIÉN Minimum Mean Square Error Log Spectral Amplitude estimator (MMSE-LSA) [Ephraim and Malah, 1985]

La técnica VTS para normalización utiliza los mismos fundamentos básicos que su variante para adaptación de modelos acústicos, y considera además que la señal limpia se puede modelar como una mezcla de Gaussianas, GMM, de manera que a cada una de sus componentes se le asocia una determinada transformación. Dicha transformación viene dada por la serie de Taylor de la función inversa del modelo de degradación propuesto, normalmente truncada hasta orden cero o uno, y que tratará de compensar conjuntamente los dos efectos perniciosos considerados del entorno acústico: el proviniente de la distorsión convolucional y el del ruido aditivo. De esta manera, la estimación final del vector de características limpio se realiza mediante una combinación lineal de todas las transformaciones asociadas a las distintas Gaussianas haciendo uso del estimador MMSE, *Minimum Mean Square Error*.

La técnica CDCN estima los parámetros del modelo de degradación mediante el algoritmo EM, a la vez que supone que la señal limpia se puede modelar mediante una mezcla de Gaussianas. Con todo ello se obtiene un vector de transformación asociado a cada Gaussiana para compensar la correspondiente degradación del entorno acústico. De cara a obtener el vector de características

normalizado se ponderan convenientemente todos los vectores de transformación. Por otra parte, y como modificación de la técnica CDCN, surgió posteriormente el algoritmo ISDCN, *Interpolated Signal to Noise Rate Dependent Cepstral Normalization*, [Acero, 1990], que es la versión interpolada del método de normalización empírico SDCN, *SNR Dependent Cepstral Normalization*, [Acero, 1990]. En la técnica ISCDN se utilizan los mismos principios básicos que en el método CDCN, añadiendo además como elemento diferenciador la relación señal a ruido, SNR, *Signal to Noise Rate*, de modo que los vectores de transformación dependerán tanto de esta relación como del modelo de degradación presupuesto. Así, ante situaciones altamente ruidosas, SNR baja, el vector de transformación pasará a compensar principalmente el ruido aditivo, mientras que si por el contrario la SNR es alta se considerará que el efecto pernicioso predominante, y por tanto el que se tratará de compensar en mayor medida, será la distorsión convolucional.

En el algoritmo VPS se supone nuevamente que la señal limpia puede modelarse mediante una mezcla de Gaussianas y, para cada una de ellas, se obtiene un término de corrección que es la diferencia entre la media de la Gaussiana limpia y la estimada como ruidosa. Esta última se calcula a partir de la mezcla de Gaussianas que modela el espacio limpio y el modelo de degradación previamente considerado. Esto supone, de alguna manera, que el efecto del entorno acústico genera una Gaussiana para el modelo contaminado por cada una del limpio, lo que no es estrictamente correcto. A su vez, y para estimar los parámetros que definen el modelo de degradación se recurre a una transformación lineal, igual que normalmente se suele hacer para el algoritmo VTS, aunque en este caso el algoritmo VPS suele mejorar ligeramente las prestaciones obtenidas con la técnica VTS.

En la técnica MMSE se propone realzar la señal de voz en el dominio log-espectral haciendo uso de las relaciones señal a ruido tanto a priori como a posteriori. En muchas ocasiones este método se combina con otros algoritmos para proporcionar mayor robustez al sistema final de RAH. De este modo se pueden utilizar conjuntamente técnicas de *arrays* de micrófonos, así como otros métodos de normalización, caso por ejemplo del algoritmo CMN, e incluso de adaptación de modelos acústicos, como la técnica MLLR [Yapanel *et al.*, 2002].

En el método StatComp la señal limpia se estima a través de la señal ruidosa que se pretende normalizar mediante la generación de muestras con el método de Monte Carlo [Lin and Chen, 1998], determinando previamente eso sí, las funciones de densidad de probabilidad del ruido, que se suelen suponer Gaussianas, y de la señal limpia, entrenada mediante el correspondiente corpus de entrenamiento. Los experimentos [de la Torre *et al.*, 2001] indican que los resultados obtenidos con este método son superiores a los logrados con otras técnicas como SS o CMN.

El método de la ecualización del ruido se sustenta en asumir que la señal limpia en el dominio temporal se puede obtener mediante una suma ponderada de la señal ruidosa y un ruido artificial. Por su parte, los pesos con que se pondera cada uno de los sumandos se determinan a partir de la relación señal a ruido y el correspondiente nivel de energía. Cabe destacar que el correcto funcionamiento de esta técnica depende en gran medida, además de la aproximación anteriormente comentada, de poseer un buen detector de voz silencio, VAD, *Voice Activity Detection*.

Ya para finalizar, la forma más sencilla para implementar la técnica SS consiste en estimar el espectro del ruido y sustraerlo del espectro de la señal de voz, de modo que en muchas ocasiones es necesario contar con un VAD suficientemente fiable. Esta solución, si bien en muchos casos produce una señal mucho más agradable de escuchar para el oído humano, también puede generar una importante distorsión (ruido musical) que hace que hace que la utilización de este algoritmo en sistemas de RAH no siempre resulte tan satisfactoria. Asimismo hay que tener siempre presen-

te las aproximaciones que se asumen en este método, esto es, considerar que el entorno acústico sólo incluye distorsión propia de ruido aditivo y que la fase de la señal de voz no se ve afectada. Así pues, si estas condiciones no se dan, el sistema final no alcanzará las prestaciones deseadas. Por todo ello se han incluido algunas extensiones para compensar las limitaciones anteriores, dando lugar a técnicas como CSS, *Constrained Spectral Subtraction*, [Korkmazskiy *et al.*, 2000], o la sustracción espectral usando harmónicos espectrales, *spectral subtraction using spectral harmonics*, [Beh and Ko, 2003].

3.3.3. Técnicas empíricas.

Las técnicas basadas en compensación empírica por comparación directa de los vectores de características requieren, en la mayoría de los casos, de señal estéreo, aunque también existen aproximaciones que no la precisan, obteniendo, eso sí, unas tasas de RAH algo menos satisfactorias. En general, este tipo de algoritmos consisten en dos fases; en la primera de ellas, que se puede denominar fase de entrenamiento, se estiman trama a trama todos aquellos parámetros necesarios para la normalización, que se lleva a cabo en la segunda fase, también conocida como de compensación. En ella se normalizan los vectores de características correspondientes haciendo uso de los parámetros anteriormente calculados y de un estimador, normalmente MMSE. Dado que en este tipo de métodos no se realiza suposición alguna sobre el modelo que produce la contaminación de la señal limpia, su éxito se sustenta principalmente en cuál próximos se encuentran los datos ruidosos empleados en la fase de entrenamiento con respecto a los que posteriormente se normalizarán.

Dentro de los métodos de compensación empírica que más frecuentemente se han venido aplicando hasta la fecha destacan SDCN, *SNR-Dependent Cepstral Normalization*, [Acero, 1990], ecualización de histogramas, *histogram equalization* [de la Torre *et al.*, 2005] [Molau, 2003], POF, *Probabilistic Optimum Filtering* [Neumeyer and Weintraub, 1994], RATZ, *multivariate Gaussian-based cepstral normalization*, y sus variantes [Moreno, 1996] y SPLICE y sus extensiones *Stereo based Piecewise LInear Compensation for Environments*, [Droppo *et al.*, 2001].

La técnica SDCN obtiene, mediante señal estéreo y en su fase de entrenamiento, un vector de transformción para cada intervalo de relación señal a ruido, para lo que el rango de la misma se divide en varias bandas. Posteriormente, y ya en la fase de compensación, se determina a qué banda corresponde cada uno de los vectores de características que se pretende normalizar y, en cada caso, se utiliza el vector de transformación correspondiente aplicando una función lineal. Por otra parte, y para reducir el tiempo de cómputo, se comprobó [Acero, 1990] que no es preciso modificar todos los coeficientes de los vectores de características, sino sólo aquellos que sean más significativos y que, en el caso de utilizar la parametrización MFCC, se corresponden con los de orden menor.

La ecualización de histograma es un método que aplica una función de transformación no lineal monótona creciente para normalizar los vectores de características, suponiendo, además de la naturaleza ya comentada de la función de transformación, la independencia de las distintas componentes de los vectores acústicos, lo que no permitirá actuar, por ejemplo, sobre efectos de rotación que el ruido pueda producir en el vector de características. El objetivo que se pretende con esta técnica es modificar la función de densidad de probabilidad de los vectores de características, acercándola a la de la señal limpia o a una predeterminada, siendo éste el criterio utilizado para estimar la correspondiente función de transformación. A partir de este algoritmo se desarrollaron ciertas extensiones, como tratar el silencio de distinta manera que la señal de voz, o añadir una técnica de rotación espacial para contrarrestar ciertos efectos del ruido [Molau, 2003].

El algoritmo POF se basa en filtrar cada vector de características a partir de los anteriores tratando de minimizar el error entre la señal normalizada y la limpia. Por otra parte, el espacio limpio se divide en varias regiones de modo que para cada una de ellas se determina un filtro. Al igual que en otras técnicas ya presentadas, el filtro final que se empleará para la normalización de cada vector acústico se obtendrá a partir de la suma ponderada de todos los asociados a las istintas regiones.

En el algoritmo RATZ se presupone que los vectores de características limpios se pueden modelar mediante una función de densidad de probabilidad Gaussiana, o más genéricamente, mediante una mezcla de Gaussianas, GMM. De esta manera, y en la fase de entrenamiento, se estima el vector de transformación asociado a cada una de las Gaussianas. Posteriormente, y ya en la fase de compensación, se determinan las probabilidades de cada una de las Gaussianas del modelo limpio dado el vector de características ruidoso, *"a posteriori invariance"*, [Moreno, 1996] y, utilizando estos valores como pesos de ponderación, se suman todos los vectores de transformaciones para normalizar de este modo el vector de características correspondiente mediante una función linal. El cálculo de los vectores de transformación puede realizarse con o sin señal estéreo [Moreno, 1996]. Sobre la base teórica de este algoritmo, se desarrollaron posteriormente ciertas modificaciones que dieron lugar a otras tantas extensiones. De este modo apareció el método SNR RATZ [Moreno, 1996], *Signal to Noise Rate multivariate Gaussian-based cepstral normalization*, que trata el coeficiente de la energia del vector de características de distinto modo que el resto, pretendiendo de esta manera introducir en la normalización final la información de la relación señal a ruido. También, y para aquellos casos en los que el espacio ruidoso pueda ser especialmente heterogéneo, se desarrolló el método IRATZ, *Interpolated multivariate Gaussian-based cepstral normalization*, [Moreno, 1996], que representa el espacio degradado mediante varios entornos básicos y, para cada uno de ellos se obtienen los vectores de transformación en la correspondiente fase de entrenamiento de manera independiente, incluyendo de este modo un nuevo parámetro: el entorno básico y logrando un algoritmo considerablemente más robusto.

En la técnica SPLICE se presupone que el espacio ruidoso se puede modelar mediante una GMM y, al igual que en el método RATZ, en la fase de entrenamiento se estima un vector de transformación para cada una de las correspondientes Gaussianas. Por otra parte, la transformación final asociada a cada vector de caraterísticas que se pretende normalizar se obtiene a partir del criterio MMSE, sumando ponderadamente todos los vectores de transformaciones a partir de las probabilidades a posteriori de cada una de las Gaussianas dado el vector de características correspondiente. El comportamiento de esta técnica, en cuanto a tasas de RAH, mejora sensiblemente el logrado con el algoritmo RATZ. Del mismo modo que en el caso anterior, y utilizando como base el método SPLICE, se desarrollaron diversas mejoras como la extensión SPLICE *with model selection* [Droppo *et al.*, 2001], que divide el espacio degradado en varios entornos básicos, dando lugar así a transformacione más específicas y robustas (sería la extensión complementaria a la desarrollada con IRATZ), o el método *dynamic* SPLICE [Droppo *et al.*, 2001], que utiliza la correlación entre la señal que se pretende normalizar basándose en la suposición de que también debe haber una cierta relación temporal entre los vectores finales de transformación empleados.

QUIZÁS AÑADIR SPACE

Tal y como se ha podido observar, cada tipo de técnicas tienen unas determinadas características y limitaciones que las hacen más útiles en unas u otras circunstancias. Por ello no es extraño ver soluciones híbridas para proporcionar una mayor robustez a los sistemas de RAH. De este modo, y por poner sólo unos ejemplos, se pueden combinar técnicas de procesado de *arrays* de micrófonos con métodos de normalización de vectores de características, como CMN, o con algoritmos de

adaptación de modelos acústicos, caso de MLLR, [Yapanel *et al.*, 2002]. Asimismo también se ha propuesto con éxito reentrenar los modelos acústicos en el espacio normalizado definido tras aplicar la técnica SPLICE [Deng *et al.*, 2000] [Droppo *et al.*, 2002], o conjugar la estimación del ruido acústico con el propio algoritmo SPLICE [Deng *et al.*, 2003] para compensar así la limitación ya comentada que poseen los métodos de normalización empíricos cuando la señal empleada en la fase de entrenamiento representa un espacio acústico distinto del de reconocimiento.

4

Bases de Datos y Experimentación.

A la hora de evaluar distintas técnicas, no ya sólo en este trabajo, sino en el ámbito de las tecnologías del habla en general, hay que elegir una o varias bases de datos que se adecúen a la problemática que se pretende solucionar del modo más fiel posible y nunca a la inversa. De esta manera, por ejemplo, no reúnen las mismas características un corpus diseñado para reconocimiento de locutor, que uno pensado para RAH. En el primero de los casos es conveniente grabar distintas sesiones de cada locutor transcurrido un cierto intervalo de tiempo, mientras que esto resulta irrelevante para tareas de RAH. Por otra parte, hay que procurar alejarse siempre de la tentación de emplear aquellas bases de datos que mejor se puedan comportar ante las bases teóricas sobre las que se apoya el algoritmo cuyo comportamiento se trata de estudiar, puesto que los resultados pueden ser engañosos. En este sentido, por ejemplo, hay que ser especialmente cuidadoso a la hora de elegir los corpora sobre los que evaluar técnicas de normalización de vectores de características basadas en modelos (*model-based*), ya que éstos asumen un determinado modelo de degradación que, en el fondo, no deja de ser una aproximación del real.

En este trabajo se pretende estudiar el comportamiento de las distintas técnicas de normalización de vectores de características ante entornos acústicos reales y altamente dinámicos, de modo que queden registradas el mayor número de alteraciones posibles en la voz, tanto las independientes del locutor, caso del ruido aditivo y la distorsión convolucional, como las dependientes del locutor, producidas por el estrés y el propio entorno acústico (efecto Lombard). Por todo ello se eligió para llevar a cabo el grueso de la experimentación la base de datos *SpeechDat Car* en español puesto que fue grabada en diferentes vehículos y situaciones de conducción. Además, el hecho de que el conductor fuera el propio locutor permite que el efecto Lombard se manifieste en mayor medida en las grabaciones. A pesar de todas estas ventajas, la base de datos *SpeechDat Car* en español posee el inconveniente de que no es tan ampliamente utilizada por la comunidad científica como lo pueden ser otras, por lo que las comparaciones con otros trabajos y publicaciones externas no son sencillas. Por ello, se realizaron también experimentos con la base de datos *Aurora 2*, que es muy utilizada por la comunidad científica y posee además gran cantidad de entornos acústicos diferentes, lo que hace de ella un banco de pruebas muy válido. Sin embargo, tiene el gran inconveniente de que la señal ruidosa se genera artificialmente añadiendo ruido aditivo, lo que hace que no se manifiesten algunas importantes alteraciones en la señal de voz tal y como ya se ha comentado anteriormente.

Una vez determinados los corpora sobre los que se va a realizar la experimentación, y cuando ésta se ha llevado a cabo, hay que determinar hasta qué punto los resultados obtenidos con las distintas técnicas presentadas y comparadas son estadísticamente significativos, esto es, si las mejoras logradas son consistentes y producto de las propias técnicas estudiadas, o bien si se deben

únicamente a la naturaleza de la base de datos. Para ello se realizan las pruebas de hipótesis estadística convenientes, que proporcionan, con un cierto intervalo de confianza, la certidumbre de si los distintos algoritmos poseen un comportamiento diferenciado.

En el presente Capítulo se analizan los dos corpora sobre los que se va a desarrollar toda la experimentación del trabajo: *SpeechDat Car* en español y *Aurora* 2 (Sección 4.1). En la Sección 4.2 se presentan las técnicas de hipótesis estadísticas más utilizadas en RAH ya que, dependiendo del tipo de experimentación, deberán poseer unas características u otras; asimismo se considerarán en cada caso las ventajas y limitaciones que presentan, determinando finalmente la que se empleará a lo largo del trabajo. Finalmente, ya en la Sección 4.3 se incluyen los resultados básicos obtenidos tanto con la base de datos *SpeechDat Car* en español como con el corpus *Aurora* 2. Estos resultados serán los que posteriormente servirán de referencia para determinar las mejoras que las distintas técnicas presentadas proporcionan.

4.1. Bases de Datos.

En el presente trabajo, y de cara a obtener unos resultados de RAH lo más fieles y comparables posibles, se ha decidido realizar la experimentación con dos bases de datos distintas. De este modo, la mayor parte de la misma se ha llevado a cabo con el corpus *SpeechDat Car* en español ya que, al ser grabado en condiciones reales, introduce todos los efectos que el ruido puede producir, tanto los independientes del locutor, como el ruido aditivo o la distorsión convolucional, como los dependientes del locutor, manifestados en una diferente pronunciación de las alocuciones debidas al estrés o al mismo ruido (efecto Lombard). Por otra parte, también se ha empleado el corpus *Aurora* 2 que, si bien no reúne las condiciones ideales anteriores, ya que el ruido se introduce artificialmente, tiene la ventaja de que es una de las bases de datos estandarizada más empleada y contrastada, por lo que se pueden realizar fácilmente comparaciones con otras técnicas y trabajos similares. A continuación se presentan las características más relevantes de los dos corpora empleados en este trabajo.

4.1.1. Base de datos *SpeechDat Car*.

Para realizar una experimentación comparativa lo más fiel posible entre todas las técnicas que se van a desarrollar a lo largo de este trabajo, se decidió emplear la base de datos *SpeechDat Car* en español [van den Heuvel *et al.*, 1999] [Moreno *et al.*, 2000]. Dicho corpus fue grabado directamente en varios vehículos en situaciones reales de conducción, por lo que la distorsión de la señal de voz no sólo incluye ruido aditivo y distorsión convolucional, como presuponen ciertos modelos de degradación expuestos habitualmente [Acero, 1990], sino también otro tipo de alteraciones en la voz dependientes del locutor producidas por el estrés o el ruido en general (efecto Lombard). Así pues, y dado que el espacio acústico representado en *SpeechDat Car* es extremadamente variable, se pueden distinguir siete entornos básicos atendiendo a las condiciones de conducción, a saber

- E1: coche detenido con el motor en funcionamiento.
- E2: coche circulando por ciudad con las ventanillas cerradas y el climatizador apagado (condiciones silenciosas).
- E3: coche circulando por la ciudad en condiciones ruidosas (ventanillas abiertas y/o climatizador encendido).
- E4: coche circulando a baja velocidad por pavimento en mal estado en condiciones silenciosas.

- E5: coche circulando a baja velocidad por pavimento en mal estado en condiciones ruidosas.
- E6: coche circulando a alta velocidad por pavimento en buen estado en condiciones silenciosas.
- E7: coche circulando a alta velocidad por pavimento en buen estado en condiciones ruidosas.

Cabe destacar que, si bien las condiciones atmosféricas de las distintas grabaciones están registradas y anotadas en el corpus, en ningún momento se tuvieron en cuenta en la experimentación. Por otra parte, la base de datos *SpeechDat Car* en español está compuesta por cuatro canales grabados simultáneamente a partir de otros tantos micrófonos, uno de los cuales (Shure SM-10A) se coloca junto a la boca del locutor, *CLose Talk*, CLK, y el resto se encuentran distribuidos por la parte delantera del habitáculo del vehículo. Sin embargo, a la hora de realizar los distintos experimentos de RAH se eligieron para este trabajo únicamente dos canales: el CLK o de referencia, que por la proximidad del sensor a la boca del conductor se puede considerar que la señal de voz obtenida está libre de ruido y, por tanto, proporciona el límite de RAH al que se puede aspirar. El segundo canal, elegido entre los tres restantes de la base de datos, es de campo lejano y en este caso se encuentra localizado en el techo del vehículo encima del locutor; el correspondiente sensor (Peiker ME15/V520-1) presenta una respuesta frecuencial de paso alto tratando así de minimizar los efectos del ruido propio de los vehículos, típicamente paso bajo. A este segundo canal se le denominará en lo sucesivo *Hands Free*, HF, y se eligió ya que presentaba los mejores resultados de RAH obtenidos con las señales de los tres distintos canales de campo lejano de la base de datos [Lleida *et al.*, 2002]. Las señales para los dos canales, CLK y HF, se muestrearon a 16 KHz y se codificaron con 16 bits.

La base de datos *SpeechDat Car* en español está dividida, además de por canales, en dos corpora: entrenamiento y reconocimiento, de los que, en este trabajo, se emplearán unas versiones algo reducidas por no disponer del todas las señales. Así, el corpus de entrenamiento estará compuesto en este caso por 16.108 frases por canal (CLK y HF) e incluye las siguientes tareas de la base de datos: dígitos aislados y conectados, deletreo, fechas, comandos y nombres. Por su parte, el corpus de reconocimiento sobre el que se va a trabajar se compone en total de 1.086 frases por canal (CLK y HF), correspondientes únicamente a la tarea de dígitos aislados y conectados (T1) y grabadas por locutores distintos de los empleados en el corpus de entrenamiento. La composición de ambos corpora se puede observar en la Tabla 4.1, donde se incluye tanto el número total de frases (# frases entrenamiento y # frases reconocimiento) como el de palabras (# palabras entrenamiento y # palabras reconocimiento) para cualquiera de los dos canales en función de los entornos básicos. Asimismo, y para dar una idea de cual relacionados en cuanto a la tarea se encuentran los corpora de entrenamiento y reconocimiento, se incluye también en la tabla el número de frases y palabras del corpus de entrenamiento correspondientes a la tarea de dígitos aislados y conectados para cualquiera de los dos canales (# frases entrenamiento T1 y # palabras entrenamiento T1).

La relación señal a ruido, *Signal to Noise Ratio*, SNR, para el canal HF posee un importante rango dinámico según el entorno básico de que se trate, ya que las condiciones de conducción, tal y como se ha indicado con anterioridad, son muy distintas; así, para el más benigno, E1, la SNR es de 14.05±3.89 dB (media ± desviación estándar); mientras que si el vehículo circula a gran velocidad por un pavimento en buen estado (entornos básicos E6 y E7 conjuntamente) la SNR pasa a ser de 5.65±4.35 dB. Por otra parte, y para completar las características del ruido presente en el canal HF, se incluye en la Figura 4.1 la densidad espectral de potencia, *Average Power Spectral Densities* (PSD), media del ruido para dicho canal y los distintos entornos básicos. Para ello se aplicó el método de Welch [Welch, 1967] sobre las correspondientes señales del corpus de entrenamiento. Se puede apreciar como en todos los casos hay un pico en torno a 150 Hz después del cual las componentes espectrales decaen exponencialmente, situación esta muy típica en ruido

	E1	E2	E3	E4	E5	E6	E7	Total
# frases entrenamiento	3.393	3.122	2.356	2.106	2.550	2.038	543	16.108
# frases entrenamiento T1	400	368	272	248	304	240	64	1.896
# frases reconocimiento	199	223	136	152	200	120	56	1.086
# palabras entrenamiento	10.542	9.652	7.160	6.517	7.908	6.265	1.673	49.717
# palabras entrenamiento T1	2.105	1.930	1.431	1.301	1.596	1.249	336	9.948
# palabras reconocimiento	1.049	1.166	715	798	1.049	630	294	5.701

Cuadro 4.1: Número de frases (# frases) y palabras (# palabras) para los canales de los corpora de reconocimiento (CLK o HF) y entrenamiento (CLK o HF) de la base de datos *SpeechDat Car* en español utilizadas en este trabajo en los distintos experimentos de RAH. El corpus de reconocimiento se compone de dígitos continuos y aislados, mientras que el de entrenamiento comprende diferentes tareas de la base de datos, no sólo dígitos. Asimismo se incluyen también los datos de la parte del corpus de entrenamiento correspondientes a la tarea de reconocimiento (T1)

de automóvil [Mokbel and Cholet, 1995]. Por otra parte en la misma figura queda patente de un modo indirecto la relación directamente proporcional existente entre la velocidad del vehículo y la potencia del ruido [Meyer and Simmer, 1997]. A pesar de que los entornos básicos no se han definido atendiendo a dicha velocidad, si se tienen en cuenta únicamente las bajas frecuencias, que en este caso es la banda frecuencial más importante, sí se puede apreciar que la potencia media del ruido para el entorno básico E1, en el que el coche está parado, es la menor, mientras que por su parte, si se consideran conjuntamente los entornos básicos E6 y E7 (vehículo circulando a alta velocidad) la potencia media del ruido es la mayor, siguiéndole la unión de los entornos básicos E4 y E5 y posteriormente la de E2 y E3, como podría haberse supuesto desde un inicio atendiendo a la relación entre la velocidad del vehículo y la potencia del ruido.

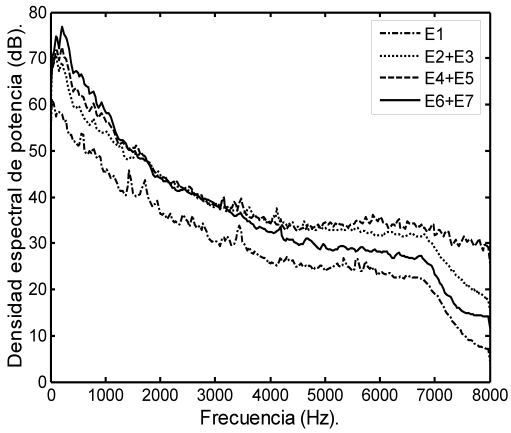


Figura 4.1: Densidad espectral de potencia media, *Average Power Spectral Densities* (PSD), del ruido obtenida a partir del canal HF para los diferentes entornos acústicos básicos definidos para el corpus *SpeechDat Car* en español: E1, que se corresponde con la línea que alterna puntos y rayas, E2 y E3, que se representan con la línea punteada, E4 y E5, que se corresponde con línea discontinua y finalmente E6 y E7, cuya línea representativa es continua.

4.1.2. Base de datos *Aurora* 2.

A pesar de que, como ya se ha comentado, se eligió la base de datos *SpeechDat Car* en español para llevar a cabo la mayor parte de la experimentación por estar grabada en un entorno real en el que se dan todos los efectos posibles del ruido, también se realizaron experimentos de RAH con el corpus *Aurora* 2 [Hirsch and Pearce, 2000] por ser esta una base de datos referente y, por tanto, muy útil a la hora de comparar distintas técnicas y trabajos.

Aurora2 se generó a partir de la base de datos de dígitos aislados y conectados en inglés *TIDigits* [Leonard and Doddington, 1993], añadiendo artificialmente ruido aditivo con diferentes SNRs a la señal limpia. Asimismo, y dado que se pretendía disponer de un corpus que se correspondiera de un modo realista con las características frecuenciales típicas de terminales y equipamiento del área de las telecomunicaciones, las señales limpias del corpus *TIDigits* se submuestreara a 8 KHz para, posteriormente, extraer la señal comprendida entre 0 y 4 KHz. Adicionalmente la señal, ya submuestreada y filtrada, se filtró nuevamente haciendo uso de una de las dos respuestas impulsionales definidas como "estándares" para equipamiento de telecomunicaciones por ITU, *International Telecommunication Union* [ITU, 1996]; dichos filtros "estándar" se denominarán en lo sucesivo G.712 y MIRS. Así pues, la base de datos *TIDigits* submuestreada y doblemente filtrada constituye el corpus limpio de *Aurora* 2, por lo que, utilizando la misma nomenclatura que en la subsección 4.1.1, se corresponde con el canal CLK.

Por su parte, el corpus ruidoso, o canal HF, se genera, como ya se ha indicado con anterioridad, tras añadir artificialmente ruido aditivo a la señal limpia con diferentes SNRs; dichos SNRs se definen tras usar el filtro "estándar" G.712 tanto para la señal no contaminada como para el propio ruido. A la hora de seleccionar el tipo de ruido se recurrió a aquellos que representaran del modo más fiel posible los entornos típicos en los que se suele hacer uso de terminales de telecomunicaciones, definiéndose finalmente ocho escenarios, a saber: metro *subway*, muchedumbre *babble*, coche *car*, salón de exhibiciones *exhibition hall*, restaurante *restaurant*, calle *street*, aeropuerto *airport* y estación de tren *train station*. Las diferentes SNRs comprenden 20dB, 15dB, 10dB, 5dB, 0dB y -5dB.

A la hora de dividir la base de datos *Aurora* 2 en los corpora de entrenamiento y reconocimiento se definen dos tipos de entrenamiento, uno compuesto únicamente por señal del canal CLK y el otro, denominado multi-condición, *multi-condition*, en el que se mezcla señal limpia (canal CLK) con ruidosa (canal HF), seleccionándose para esta última cuatro tipos de ruido diferentes: *subway*, *babble*, *car* y *exhibition hall* con cinco SNRs distintas: 20dB, 15dB, 10dB, 5dB y limpia. En ambos corpora de entrenamiento se hace uso del filtro "estándar" G.712.

Por su parte, el corpus de reconocimiento está dividido en tres *sets* (A, B y C). Los dos primeros se generan a partir de las mismas 4.004 frases limpias provinientes del corpus de reconocimiento de la base de datos *TIDigits*, mientras que el tercer *set* se obtiene únicamente a partir de 2.002 frases limpias del corpus de reconocimiento de la base de datos *TIDigits*. A continuación se indica la composición de los diferentes *sets*

- El *set* A está compuesto por señal ruidosa generada a partir de cuatro tipos de ruido distintos: *subway*, *babble*, *car* y *exhibition hall*, y siete SNRs diferentes: 20dB, 15dB, 10dB, 5dB, 0dB, -5dB y limpia, siendo en todo momento G.712 el filtro "estándar" utilizado. Para cada tipo de ruido y SNR se emplean 1.001 de las 4.004 frases distintas de todo el corpus de reconocimiento seleccionado de la base de datos *TIDigits*. De este modo, el corpus final de reconocimiento está constituido por 28.028 frases ($1.001 \times 7 \times 4$). Se puede apreciar que los tipos de ruido empleados en este caso son los mismos que los que forman parte del entrenamiento multicondición, tal y como quedará patente posteriormente en las correspondientes tasas de RAH;

sin embargo habrá un importante desajuste si se emplean los modelos acústicos obtenidos a partir de la señal del corpus limpio de entrenamiento.

- El *set* B se genera exactamente del mismo modo que el *set* A modificando únicamente los tipos de ruido, que en esta ocasión serán: *restaurant*, *street*, *airport* y *train station*. En esta ocasión existirá un serio desajuste entre la señal de reconocimiento y de entrenamiento incluso para el caso de multi-condición. De este modo, con este experimento se trata de observar la importancia en el RAH cuando se pretende reconocer señales contaminadas con ruido que no se ha visto hasta el momento en la fase de entrenamiento.
- El *set* C utiliza, a diferencia de los dos anteriores, el filtro "estándar" MIRS, incluyendo posteriormente únicamente dos tipos de ruido aditivo: los correspondientes a los entornos *subway* y *street* con las siete SNRs ya consideradas anteriormente: 20dB, 15dB, 10dB, 5dB, 0dB, -5dB y limpia. En este caso para cada clase de ruido y SNR se utilizarán 1.001 frases distintas del corpus de reconocimiento seleccionado de la base de datos *TIDigits*, definiéndose pues un *set* sensiblemente menor que los dos anteriores: 14.014 frases (1.001×7×2). En este caso se pretende estudiar el comportamiento del sistema de RAH cuando la distorsión convolucional es distinta en los corpora de entrenamiento y reconocimiento.

La composición en número de frases tanto para los dos corpora de entrenamiento como para los tres *sets* de reconocimiento se pueden observar en la Tabla 4.2.

	Filtrado	Limpio	*Subway*	*Babble*	*Car*
# frases entrenamiento limpio	G.712	8.440	0	0	0
# frases entrenamiento multi-condición	G.712	1.688	1.688	1.688	1.688
# frases reconocimiento *set* A	G.712	4.004	6.006	6.006	6.006
# frases reconocimiento *set* B	G.712	4.004	0	0	0
# frases reconocimiento *set* C	MIRS	2.002	6.006	0	0

Hall	*Restaurant*	*Street*	*Airport*	*Station*	Total	
0	0	0	0	0	8.440	# frases entrenamiento limpio
1.688	0	0	0	0	8.440	# frases entrenamiento multi-condición
6.006	0	0	0	0	28.028	# frases reconocimiento *set* A
0	6.006	6.006	6.006	6.006	28.028	# frases reconocimiento *set* B
0	0	6.006	0	0	14.014	# frases reconocimiento *set* C

Cuadro 4.2: Número de frases para los dos corpora de entrenamiento (limpio y multi-condición) y los tres *sets* de reconocimiento (A, B y C) de la base de datos *Aurora* 2. Todos ellos están compuestos por dígitos continuos y aislados.

Ya para concluir se presenta en la Figura 4.2 la densidad espectral de potencia media obtenida mediante el método de Welch para los distintos tipos de ruido que se pueden dar en la base de datos *Aurora* 2. Cabe destacar como en todos los casos la mayor parte de la energía del ruido se concentra en baja frecuencia, pudiendo parecer, si únicamente se consideran las PSDs, que varios tipos de ruido son similares. Sin embargo no es así y, por ejemplo, los hay altamente estacionarios, como los correspondientes a *car* y *exhibition*, y los hay que se caracterizan precisamente por su falta de estacionaridad, como *street* o *airport*.

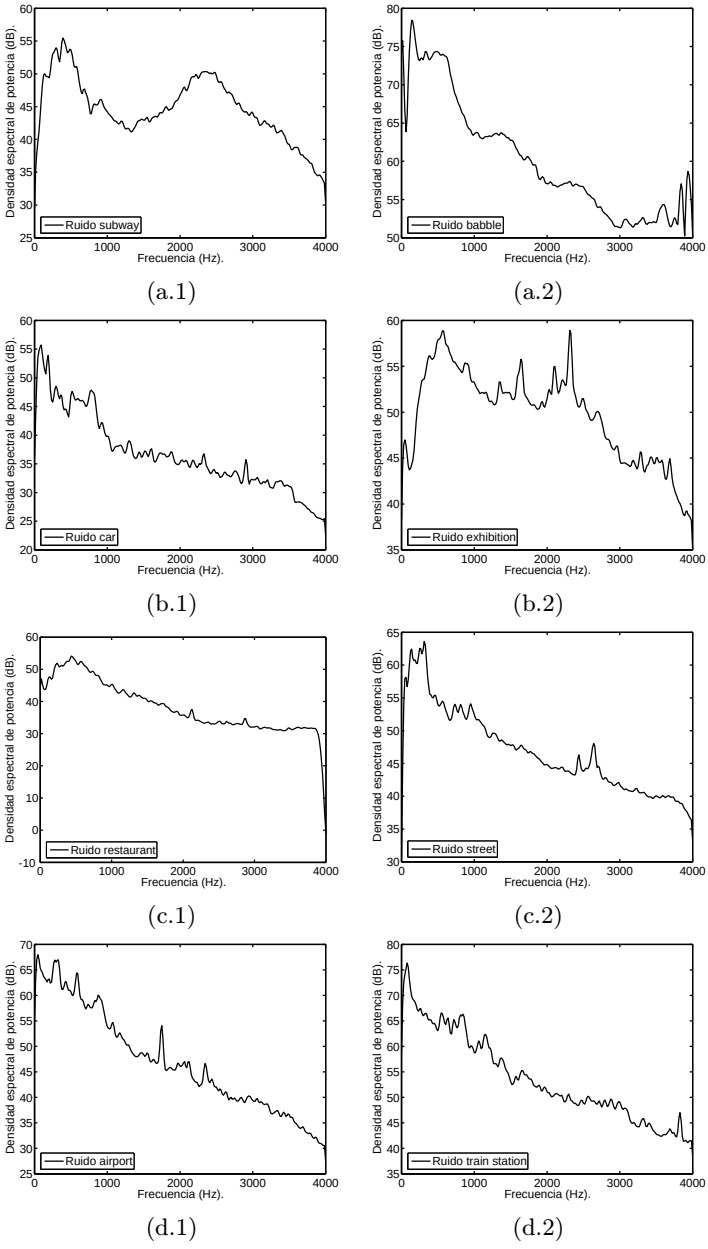


Figura 4.2: Densidad espectral de potencia media, *Average Power Spectral Densities* (PSD), de los distintos ruidos presentes en el corpus *Aurora* 2: *subway* a.1, *babble* a.2, *car* b.1, *exhibition* b.2, *restaurant* c.1, *street* c.2, *airport* d.1, *train station* d.2,.

4.2. Pruebas de Hipótesis Estadística.

A la hora de comparar distintas técnicas, no ya sólo en el ámbito del RAH sino en cualquier disciplina, no basta solamente con presentar los resultados de la experimentación y cotejarlos directamente, sino que se ha de establecer de un modo estadístico hasta qué punto la diferencia de

comportamiento entre las técnicas es significativa [Gillick and Cox, 1989]. Por ello en este trabajo se ha empleado la prueba de hipótesis estadística *z-test* para que a la hora de establecer comparaciones entre los diversos algoritmos se pueda decir, al menos con un cierto intervalo de confianza, que una técnica u otra, independientemente de la base de datos empleada en la experimentación, se comporta de mejor modo.

En el dominio del RAH, tres son las principales pruebas de hipótesis estadística empleadas a la hora de comparar dos técnicas, a saber: la de McNemar [McNemar, 1947] [Gillick and Cox, 1989], *matched-pairs* [Gillick and Cox, 1989] [Pallett *et al.*, 1990] y *z-test* [Gillick and Cox, 1989]; si se deseara estudiar conjuntamente el comportamiento de más de dos algoritmos habría que recurrir a otro tipo de evaluaciones [Lehmann, 1975].

La prueba de hipótesis estadística de McNemar está pensada para evaluar el comportamiento de dos técnicas cuyos resultados se obtienen a partir de variables discretas independientes etiquetadas como correctas o erróneas. Dicha prueba únicamente considera como información relevante el número de variables etiquetadas de distinta manera por ambas técnicas, mientras que desecha el resto. A su vez, toma como hipótesis nula, esto es, que ambos algoritmos no proporcionan diferentes resultados de un modo estadísticamente significativo, el hecho de que, dado que uno de los métodos ha cometido un error, es igualmente verosímil que haya sido uno u otro. Para rechazar la hipótesis nula se presupone que la variable aleatoria definida como errores cometidos por una técnica y no la otra, normalizada en media y varianza asumiendo la hipótesis nula, sigue una densidad de probabilidad normal de media nula y varianza unidad, $\mathcal{N}(0,1)$. De esta manera, se puede calcular mediante tablas la probabilidad de que dicha variable aleatoria sea menor que el valor obtenido a partir de los datos concretos tras aplicar las dos técnicas de estudio. Si dicha probabilidad es menor que una fijada, α, se podrá decir que ambas técnicas presentan resultados estadísticamente significativos con un intervalo de confianza de $1-\alpha$. En el dominio del RAH se podría considerar la palabra como variable discreta independiente, pero eso no es del todo cierto salvo que se reconozcan palabras aisladas ya que, normalmente, los modelos de lenguaje introducen dependencia entre palabras próximas. Por todo ello, en muchas ocasiones se suele considerar a la frase como variable discreta independiente, lo que plantea otro tipo de problema ya que cada frase etiquetada como errónea puede tener cualquier número de palabras erróneas, lo que puede dar lugar a una comparación injusta entre los distintos algoritmos.

La prueba de hipótesis estadística *matched-pairs* se utiliza para comparar el comportamiento de dos técnicas estudiando la diferencia entre el número de errores ocurridos entre los dos algoritmos en unidades de distinta longitud e independientes entre sí, no importando en ningún momento el tipo del error siempre que se recuente de un modo consistente para ambas técnicas. En este caso se introduce la variable aleatoria definida como la media de la diferencia de número de errores por unidad, así como la hipótesis nula, que en este caso consiste en que la media de dicha variable es nula. De este modo, para rechazar la hipótesis nula se presupone que la variable aleatoria definida como la media de la diferencia del número de errores por unidad normalizada en varianza sigue una densidad de probabilidad normal de media nula y varianza unidad, $\mathcal{N}(0,1)$. De esta manera, se puede calcular mediante tablas la probabilidad de que dicha variable aleatoria tome un valor menor que el obtenido a partir de los datos concretos tras evaluar ambas técnicas. Si dicha probabilidad es menor que una fijada, α, se podrá decir que ambas técnicas presentan resultados estadísticamente significativos con un intervalo de confianza de $1-\alpha$. En el ámbito del RAH la elección de las unidades se realiza normalmente fraccionando las frases considerando como límite una palabra correctamente reconocida por los dos sistemas que se pretenden comparar, o bien el inicio y final de la frase; sin embargo en algunas ocasiones, y dependiendo del modelo de lenguaje empleado, se

puede ser más estricto en cuanto al número de palabras seguidas bien reconocidas necesarias para marcar los límites de las unidades. De todos modos, se puede apreciar que esta prueba de hipótesis estadística es bastante dependiente de las unidades que se tomen, de modo que se podrían obtener muy diferentes resultados según cómo se realizara la segmentación, siendo además en muchos casos dificil poder asegurar que los errores cometidos en unidades próximas sean independientes.

Por último, y a pesar de sus limitaciones, como se podrá observar más adelante, el método de prueba de hipótesis estadística más empleado en RAH y del que también se hará uso en este trabajo, es el denominado como *z-test*. En este caso para comparar el comportamiento de dos técnicas, A_1 y A_2, cuyas tasas de error reales, y por tanto desconocidas, son p_1 y p_2 respectivamente, se define la variable aleatoria $d = p_1 - p_2$. De este manera, la hipótesis nula se representa como $H_0 : p_1 = p_2 = p$, o bien como $H_0 : d = p_1 - p_2 = 0$. Por otra parte, y bajo dicha hipótesis nula, la variable d se puede estimar mediante el criterio ML como $\hat{p}_1 - \hat{p}_2$, siendo $\hat{p}_1$ y $\hat{p}_2$ las estimaciones de p_1 y p_2, respectivamente; asimismo la varianza asociada a d, $\sigma_d^2 = var(p_1 - p_2)$, toma, asumiendo que $\hat{p}_1$ y $\hat{p}_2$ son independientes, la forma $\sigma_d^2 = \sigma_1^2 + \sigma_2^2$, donde σ_1^2 y σ_2^2 son las varianzas de p_1 y p_2, respectivamente. De esta manera, la estimación de σ_d^2 será, asumiendo que la hipótesis nula es correcta y que ambas técnicas se evalúan sobre la misma base de datos

$$\hat{\sigma}_d^2 = \frac{2\hat{p}(1-\hat{p})}{n}, \tag{4.1}$$

donde n es el número de palabras de la base de datos y $\hat{p}$ es la estimación de p que, en este caso, y dado que la base de datos sobre las que se evalúan ambos algoritmos es la misma, se obtendrá mediante el estimador ML de la siguiente manera

$$\hat{p} = \frac{\hat{p}_1 + \hat{p}_2}{2}. \tag{4.2}$$

Con todo lo anterior, y asumiendo que la hipótesis nula es cierta, el estadístico empleado por la técnica *z-test*, denominado W, posee la siguiente distribución

$$W = \frac{\hat{d}}{\hat{\sigma_d}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{2\hat{p}(1-\hat{p})}{n}}}, \tag{4.3}$$

que es la estimación de la variable aleatoria d normalizada en media y varianza considerando como correcta la hipótesis nula (la media de d en este caso es nula). Aplicando el teorema central del límite, se puede asumir que el estadístico W tiende a una normal de media 0 y varianza 1, $\mathcal{N}(0,1)$, si n es lo suficientemente elevado (normalmente se suele considerar que basta con que sea mayor de 50). Así, para determinar si la hipótesis nula es incorrecta, esto es, que los dos algoritmos presentan comportamientos diferenciados estadísticamente significativos, bastará con calcular mediante tablas la probabilidad de que la variable W tome un valor menor que el obtenido a partir de los datos concretos tras evaluar ambas técnicas, denominado w. Si dicha probabilidad es menor que una fijada, α, se podrá decir que ambas técnicas presentan resultados estadísticamente significativos con un intervalo de confianza de $1 - \alpha$. Esto es $2p(W \geq |w|) < \alpha$, donde recuérdese que la función de densidad de probabilidad asociada a W responde a una normal de media nula y varianza unidad, $\mathcal{N}(0,1)$.

Esta prueba de hipótesis estadística, si bien muy extendida en RAH, debe utilizarse teniendo siempre en cuenta sus limitaciones. Como ya se ha indicado, para poder utilizar las expresiones anteriormente presentadas es necesario que $\hat{p}_1$ y $\hat{p}_2$ sean independientes, cosa que, desafortunadamente no puede asumirse cuando ambos algoritmos se comparan sobre la misma base de datos. Si no se pudiera asumir la independencia entre $\hat{p}_1$ y $\hat{p}_2$ habría que modificar la expresión (4.1) incluyendo

un nuevo término asociado a la covarianza entre las probabilidades de las dos técnicas. De todos modos, si dicha covarianza fuera negativa y la hipótesis estadística determinara que ambas técnicas poseen un comportamiento estadísticamente diferenciado, el resultado seguiría siendo válido por haberse obtenido en condiciones aún más conservadoras. En el caso contrario, covarianza positiva y que la hipótesis estadística hubiera determinado que las dos técnicas proporcionan resultados estadísticamente diferenciados, no se podría decir lo mismo. De todo lo anterior se puede concluir que siempre que se use el método de prueba de hipótesis estadística *z-test* bajo la misma base de datos, como en este trabajo, hay que tomar los resultados con cierta cautela puesto que se ha obviado el término de covarianza entre las probabilidades de las dos técnicas estudiadas. De todos modos, y llegados a este punto, a la hora de comparar el comportamiento de dos técnicas parece más problemático el uso de corpora diferentes, o el empleo de las pruebas de hipótesis estadísticas anteriormente comentadas, de McNemar o *matched-pairs*, que utilizar la técnica *z-test*, y todo ello a pesar de su limitación.

4.3. Experimentación.

Tal y como se ha justificado con anterioridad, la experimentación realizada en este trabajo conjuga el uso de dos bases de datos: *SpeechDat Car* en español y *Aurora* 2. Igualmente, y tanto por necesidades de algunos de los algoritmos tratados como por proporcionar resultados con varias condiciones de experimentación, se han empleado distintas parametrizaciones, así como diferentes unidades para el modelado acústico. En las siguientes subsecciones se presentan los resultados de referencia, *baselines*, para los dos corpora haciendo uso de las distintas condiciones de reconocimiento cosideradas. Estos resultados permitirán posteriormente comparar el comportamiento de las distintas técnicas presentadas a lo largo del trabajo.

4.3.1. Experimentación con el corpus *SpeechDat Car* en español.

Para la fase de reconocimiento se utilizan dos tipos de parametrizaciones, aunque las diferencias conceptuales entre ellas son escasas. La primera parametrización, que se denominará a partir de ahora como *parametrización UZ* por ser la usada de manera continuada en el grupo de tecnologías del habla de la Universidad de Zaragoza, construye los vectores de características a partir de la parametrización estándar ETSI [ETSI, 2000] con dos pequeñas modificaciones: se realiza una normalización tras aplicar los filtros de escala Mel, y el vector final está compuesto por 37 parámetros: los 12 coeficientes MFCC, la primera y segunda derivada de los mismos y la derivada del logarítmo de la energía. Por su parte, el segundo método es la parametrización estándar ETSI, que proporciona vectores de características de 39 componentes: 12 coeficientes MFCC más el logaritmo de la energía, junto con su primera y segunda derivadas. Independientemente del tipo de parametrización, los vectores acústicos se calculan cada 10 ms utilizando una ventana de Hamming de 25 ms.

En cuanto al modelado acústico, se han considerado dos opciones distintas, representando en cada caso un tipo de unidad distinto, a saber, incontextuales o fonemas y palabras. El modelado acústico de unidades fonéticas se compone de 25 HMMs, de modo que cada uno está asociado a un fonema español; a su vez se utilizan dos modelos de silencio, uno largo y otro corto para representar la pausa entre palabras. Cada HMM, salvo el asociado al silencio corto, está compuesto por tres estados y la función de densidad de probabilidad asociada a cada uno se compone de una GMM de 16 componentes. De esta manera, cada palabra se modela mediante la concatenación de las correspondientes unidades fonéticas. Por su parte, el HMM correspondiente al silencio corto se construye a partir de un único estado al que le corresponde como función de densidad de probabilidad una GMM de 16 componentes.

Por otra parte, el modelado acústico de palabras está compuesto por 12 HMMs, 10 de ellos, los correspondientes a los dígitos que definen la tarea de reconocimiento, compuestos por 16 estados a los que se les asocia una GMM de 3 componentes a cada uno como función de densidad de probabilidad correspondiente. Además el silencio largo se modela con un HMM de 3 estados con una GMM de 6 componentes cada uno, y al silencio entre palabras, o corto, se le asocia un HMM de un estado con una GMM de 6 componentes. Cabe destacar, por cuanto es una diferencia importante en algunos aspectos, que en este caso los modelos acústicos se entrenan únicamente con la parte del corpus de entrenamiento perteneciente a la tarea de dígitos continuos y aislados (ver Tabla 4.1). El modelo de lenguaje en toda la experimentación es muy sencillo, permitiéndose cualquier secuencia de dígitos.

En la Tabla 4.3 se presentan los resultados de referencia en términos de *Word Error Rate*, WER, para los distintos entornos básicos cuando se emplea la *parametrización UZ* y los modelos acústicos de unidades fonéticas. Cabe destacar que MWER se corresponde con el WER medio calculado a lo largo de todos los entornos básicos proporcionalmente al número de palabras de que se dispone para cada uno (ver Tabla 4.1). Por otra parte, la columna marcada como "Entrenamiento" indica el canal de las señales empleadas para estimar los correspondientes modelos acústicos: si se obtuvieron a partir de la señal limpia, la columna se marca con CLK, por el contrario, si la columna se nombra con HF indica que los modelos acústicos se han entrenado con toda la señal ruidosa; por su parte, HF† hace referencia a que las señales de cada entorno básico se reconocen con modelos acústicos específicos, esto es, obtenidos a partir de la señal de entrenamiento del correspondiente entorno básico, lo que no deja de proporcionar unos valores ficticios puesto que en un caso real no se conoce a ciencia cierta el entorno básico al que pertenece la señal que se pretende reconocer. Ya para finalizar, la columna "Reconocimiento" indica que canal se emplea a la hora de reconocer: CLK, que se decodifica la señal limpia, o HF, si se hace lo propio con la señal ruidosa. En toda la experimentación de este trabajo desarrollada sobre la base de datos *SpeechDat Car* en español se aplica la técnica CMN tanto al corpus de entrenamiento como al de reconocimiento.

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)
CLK	CLK	1.90	2.64	1.81	1.75	1.62	0.64	0.35	1.75
CLK	HF	5.91	14.49	14.55	20.17	21.07	16.19	35.71	16.21
HF	HF	6.67	14.24	12.73	12.91	14.97	9.68	8.50	11.81
HF†	HF	2.86	7.12	4.34	4.39	7.63	4.60	4.76	5.30

Cuadro 4.3: Resultados de referencia en términos de WER (%), para los diferentes entornos básicos (E1,..., E7) utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas. Dichos modelos acústicos se pueden generar a partir de la señal limpia (CLK en la columna de Entrenamiento) o la ruidosa (HF en la columna de Entrenamiento); HF† indica que se utilizan modelos acústicos específicos para cada entorno básico. La columna marcada como "Reconocimiento" hace referencia a la señal empleada para reconocer, que puede ser limpia (CLK) o ruidosa (HF).

A partir de la Tabla 4.3 se puede apreciar el negativo efecto en las prestaciones del sistema de RAH que produce el ruido presente en todos y cada uno de los entornos básicos, dando lugar a un incremento significativo del WER en todos los casos (Entrenamiento CLK, Reconocimiento HF) con respecto a las tasas obtenidas cuando es la señal limpia la que se emplea para reconocer (Entrenamiento CLK, Reconocimiento CLK). Por otra parte, utilizar modelos acústicos entrenados

con toda la señal ruidosa, *matched condition*, (Entrenamiento HF, Reconocimiento HF) hace que el valor medio de WER, MWER, decaiga considerablemente con respecto al obtenido con modelos acústicos limpios (Entrenamiento CLK, Reconocimiento HF), aunque esta mejora no es tal en los entornos básicos menos ruidosos, E1 y E2, ya que los modelos acústicos ruidosos obtenidos representan de un modo generalista a toda señal sucia y ésta es muy heterogénea. Debido a que los entornos básicos son muy dispares entre sí, tal y como ya quedó patente con las diferentes SNRs presentadas en la Sección 4.1.1 así como en la Figura 4.1, el emplear modelos acústicos específicos para cada entorno básico (Entrenamiento HF†) proporciona importantes mejoras en todos los casos, hecho este que no se da cuando el reentrenamiento se realiza con toda la señal ruidosa.

Siguiendo la misma nomenclatura de la Tabla 4.3, se presentan en la Tabla 4.4 los resultados de RAH en términos de WER cuando se hace uso de la parametrización ETSI y modelos acústicos de palabras. En este caso se puede apreciar igualmente el pernicioso efecto del ruido que se encuentra en los distintos entornos básicos, (Entrenamiento CLK, Reconocimiento HF). Mediante este resultado también queda patente que la combinación *parametrización UZ*-modelado acústico de unidades fonéticas es sensiblemente más robusta que el empleo de la parametrización estándar ETSI junto con modelado acústico de palabras, obteniéndose una tasa media de error en palabra de 16.21 % para el primero de los casos, sensiblemente inferior al 21.49 % lograda con la segunda combinación. Igualmente representativos son los resultados obtenidos tras entrenamiento con señal ruidosa, *matched condition*, ya que en este caso la combinación parametrización estándar ETSI-modelado acústico de palabras proporciona importantes mejoras con respecto al uso conjunto de la *parametrización UZ* con modelado acústico de unidades fonéticas (4.63 % de MWER comparado con 11.81 %). Ya para finalizar se puede apreciar como los mejores resultados en este caso se obtienen cuando se emplean modelos acústicos ruidosos dependientes de cada entorno básico (3.42 % de MWER), y todo ello a pesar de que los modelos acústicos del entorno básico E7 se ven claramente afectados por una seria falta de datos (64 frases, 336 palabras). Así pues, y a modo de resumen, se puede indicar que el uso de la parametrización estándar ETSI con modelado acústico de palabras, aunque menos robusto, proporciona mejores resultados que los obtenidos con la combinación *parametrización UZ*-modelos acústicos de fonemas en el resto de experimentos de referencia.

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)
CLK	CLK	0.95	2.32	0.70	0.25	0.57	0.32	0.00	0.91
CLK	HF	3.05	13.29	15.52	27.32	31.36	35.56	53.06	21.49
HF	HF	3.81	6.86	3.50	3.76	4.96	4.44	3.06	4.63
HF†	HF	1.14	4.37	1.68	2.13	2.10	2.06	23.13	3.42

Cuadro 4.4: Resultados de referencia en términos de WER (%), para los diferentes entornos básicos (E1,..., E7) utilizando la parametrización estándar ETSI y modelos acústicos de palabras. Dichos modelos acústicos se pueden generar a partir de la señal limpia (CLK en la columna de Entrenamiento) o la ruidosa (HF en la columna de Entrenamiento); HF† indica que se utilizan modelos acústicos específicos para cada entorno básico. "Reconocimiento" hace referencia a la señal empleada para reconocer, que puede ser limpia (CLK) o ruidosa (HF).

4.3.2. Experimentación con el corpus *Aurora* 2.

En la experimentación realizada con la base de datos *Aurora* 2 se utiliza la parametrización estándar ETSI [ETSI, 2000], que proporciona, tal y como se ha indicado con anterioridad, vectores

de características de 39 componentes formados por 13 parámetros estáticos (12 coeficientes MFCC más el logaritmo de la energía), junto con 26 dinámicos (la primera y segunda derivadas). Los vectores acústicos se calculan cada 10 ms haciendo uso de una ventana de Hamming de 25 ms.

Aurora 2 Small Vocabulary	Multicondition training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Absolute word accuracy. If an HTK output is WORD: %Corr=99.14, Acc=98.68 [H=……], the value to enter is 98.68.	Clean	98,34	98,31	98,12	98,98	98,44	98,34	98,31	98,12	98,98	98,44	98,01	97,98	97,99	98,35
	20 dB	98,25	98,04	97,86	98,06	98,05	97,39	97,24	97,43	97,15	97,30	97,79	96,99	97,39	97,62
	15 dB	97,73	97,14	97,29	97,38	97,39	95,78	96,26	96,22	95,50	95,94	97,03	96,00	96,52	96,63
	10 dB	95,87	95,43	95,95	94,22	95,37	92,38	93,65	93,45	92,60	93,02	94,03	93,10	93,57	94,07
	5 dB	90,85	88,94	89,43	87,71	89,23	84,52	85,30	86,99	84,88	85,42	82,89	83,22	83,06	86,47
	0 dB	71,23	65,09	64,88	68,77	67,49	64,66	66,93	69,57	65,06	66,56	47,78	56,41	52,09	64,04
	-5dB	30,50	32,91	22,54	29,07	28,75	35,72	34,52	39,58	34,01	35,96	16,20	27,62	21,91	30,27
	Average	90,79	88,93	89,08	89,23	**89,51**	86,94	87,88	88,73	87,04	**87,65**	83,91	85,14	**84,53**	87,77

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restaurar	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Absolute word accuracy. If an HTK output is WORD: %Corr=99.14, Acc=98.68 [H=……], the value to enter is 98.68.	Clean	99,11	98,91	99,20	99,38	99,15	99,11	98,91	99,20	99,38	99,15	99,14	98,94	99,04	99,13
	20 dB	97,46	94,66	97,32	97,42	96,71	94,90	96,34	95,24	96,34	95,70	96,05	96,68	96,36	96,24
	15 dB	92,82	83,49	92,23	93,24	90,45	85,56	91,31	86,01	89,59	88,12	89,78	91,85	90,82	89,59
	10 dB	81,86	67,64	77,46	81,45	77,10	70,25	76,89	71,41	74,48	73,26	79,45	78,12	78,78	75,90
	5 dB	65,85	51,97	53,35	57,16	57,08	54,46	56,13	54,35	53,54	54,62	58,62	57,04	57,83	56,25
	0 dB	39,32	34,29	23,30	27,56	31,12	35,40	32,95	36,12	28,68	33,29	33,38	31,10	32,24	32,21
	-5dB	16,12	19,36	9,68	10,61	13,94	19,80	14,78	18,54	12,13	16,31	16,52	14,68	15,60	15,22
	Average	75,46	66,41	68,73	71,37	**70,49**	68,11	70,72	68,63	68,53	**69,00**	71,46	70,96	**71,21**	70,04

Figura 4.3: Exatitud por palabra, *word accuracy*, obtenida para la base de datos *Aurora* 2 utilizando la parametrización estándar ETSI y modelos acústicos de palabras. Se incluyen los resultados para los diferentes *sets* de reconocimiento (A, B y C) y las dos condiciones de entrenamiento (multi-condición o *multicondition training* y limpio o *clean training*).

En cuanto al modelado acústico seleccionado, éste utiliza las palabras como unidades, de modo que cada uno de los 11 dígitos (en inglés el dígito 0 tiene dos posibles pronunciacion) se representa con un HMM de 16 estados, asociándoles a cada uno de ellos una GMM de 3 componentes como función de densidad de probabilidad correspondiente. Por otra parte, se consideran también dos silencios, uno largo que se modela con un HMM de 3 estados con una GMM de 6 componentes para cada uno, y un silencio entre palabras, o corto, al que se le asocia un HMM de un estado con una GMM de 6 componentes. El modelo de lenguaje, al igual que para la experimentación realizada mediante el corpus *SpeechDat Car* en español, permite cualquier secuencia de dígitos.

En la Tabla 4.3 se presentan, del modo típico en que se suele hacer, los correspondientes resultados de referencia en términos de exactitud por palabra, *word accuracy*, obtenidos con las condiciones de experientación anteriormente comentadas. Como es de esperar, si se hace uso de los modelos acústicos multicondición, los resultados son sensiblemente más competitivos que si se emplean los modelos acústicos limpios. Esto es debido a que el desajuste entre los espacios de entrenamiento y reconocimiento es menor en el primero de los casos. Del mismo modo, también se aprecia claramente la relación existente entre SNR y el comportamiento del sistema. Así, conforme se reduce la SNR correspondiente, la exactitud por palabra decae hasta llegar a niveles dramáti-

cos. Para finalizar, cabe destacar, tal y como queda patente en la Tabla 4.4, que los resultados presentados en este apartado proporcionan ya de por sí una mejora media del 14.86 % con respecto a los considerados normalmente por la comunidad científica como referencia para esta base de datos. Dichos resultados de referencia se obtienen con el sistema de RAH HTK [Young *et al.*, 2005] mediante las mismas condiciones de experimentación (parametrización y modelos acústicos y de lenguaje) definidas previamente.

Aurora 2 Relative Improvement				
	Set A	**Set B**	**Set C**	**Overall**
Multi	7,96%	-7,22%	-2,67%	**-0,24%**
Clean	28,90%	35,28%	21,43%	**29,96%**
Average	**18,43%**	**14,03%**	**9,38%**	**14,86%**

Figura 4.4: Mejoras relativas logradas para la base de datos *Aurora* 2 utilizando la parametrización estándar ETSI y modelos acústicos de palabras con respecto a los resultados obtenidos bajo las mismas condiciones de experimentación con el sistema de RAH HTK. Se incluyen los resultados para los diferentes *sets* de reconocimiento (A, B y C) y las dos condiciones de entrenamiento (multi-condición o *multicondition training* y limpio o *clean training*).

5

Visión Unificada de las Técnicas de Normalización Basadas en MMSE.

Tal y como se ha indicado en el Capítulo 3, muchas son las técnicas de normalización de vectores de características que se han venido empleado a lo largo del tiempo para proporcionar robustez a los sistemas de RAH; sin embargo, y aunque cada una de ellas aborda el problema de un modo distinto, en el fondo todas se sustentan en un conocimiento o suposición de la degradación que el entorno acústico produce en los distintos coeficientes de los vectores de características de la voz.

Por otra parte, y también en el Capítulo 3, se ha realizado una clasificación pormenorizada de los métodos más comunes de normalización de los vectores de características, separándolos en tres grandes grupos: filtrado paso alto, basados en modelos y empíricos. No obstante, los algoritmos más utilizados en la actualidad no tienen como nexo común el pertenecer a una u otra de estas clases, sino que se caracterizan por tratar de obtener el vector de características limpio mediante el estimador Bayesiano óptimo que minimiza el error cuadrático medio, *Minimum Mean Square Error*, MMSE, para lo cual se ha de suponer una cierta función de densidad de probabilidad a priori de la variable que se pretende obtener, lo que en ciertas situaciones no es sencillo. De esta manera, se puede considerar que algoritmos tan dispares como CMN, CDCN, SDCN, VTS, VPS, RATZ o SPLICE provienen, en primera aproximación, de la misma base teórica.

En este Capítulo se realiza primeramente (Sección 5.1) un estudio sobre los efectos, tanto a nivel estadístico como de incertidumbre, que distintos tipos de entornos acústicos producen sobre los coeficientes MFCC, que a la postre compondrán los vectores de características empleados en las distintas técnicas de normalización propuestas en este trabajo. A continuación (Sección 5.2) se plantea el desarrollo teórico unificado para las técnicas de compensación empírica más utilizadas recientemente: CMN, que aunque es un método de filtrado paso alto, puede verse también como el más sencillo de los algoritmos de normalización empírica, RATZ y SPLICE. Dicho desarrollo servirá posteriormente de base teórica para los distintos métodos propuestos en este trabajo. La Sección 5.3 está dedicada al algoritmo MEMLIN, *Multi-Environment Model-based LInear Normalization*, que proporciona una normalización empírica basada en el criterio MMSE, y que fue desarrollado como respuesta a ciertas limitaciones de las técnicas RATZ y SPLICE. Los resultados de RAH obtenidos tras la aplicación de distintos métodos de normalización empíricos con la base de datos *SpeechDat Car* en español, se incluyen en la Sección 5.4. En ella queda patente el buen comportamiento del algoritmo MEMLIN con respecto a los métodos empíricos basados en el criterio MMSE más utilizados en la actualidad (CMN, RATZ y SPLICE).

5.1. El Efecto del Ruido.

Con el fin de estudiar el efecto que el entorno acústico produce en la señal de voz, se suele proponer normalmente un modelo simplificado de degradación que simule las características del propio entorno acústico real. Así, el modelo más ampliamente utilizado considera que la señal contaminada, en el dominio temporal, se puede aproximar a partir de la correspondiente señal limpia filtrada a la que posteriormente se le añade un ruido aditivo [Acero, 1990]. De este modo, todas las posibles alteraciones que el entorno acústico pueda producir en la señal de voz limpia quedan representadas mediante la combinación de un ruido aditivo y una distorsión convolucional. Tal y como se ha comentado en el Capítulo 2, normalmente los sistemas de RAH no emplean directamente la señal de voz en el dominio temporal, sino que a partir de ella obtienen una serie de vectores de características que finalmente constituyen la entrada al sistema de RAH; así pues, y considerando el mismo modelo simplificado de degradación introducido anteriormente, un vector de características ruidoso en el dominio MFCC, $\mathbf{y}_t$, se puede expresar del siguiente modo

$$\mathbf{y}_t = \mathbf{x}_t + f(\mathbf{x}_t, \mathbf{n}_t, \mathbf{h}_t), \tag{5.1}$$

donde t es el índice temporal de los vectores acústicos correspondientes en el dominio MFCC: $\mathbf{x}_t$, que es el vector de características limpio, $\mathbf{n}_t$, que se corresponde con la trama que representa la contribución del ruido aditivo y, finalmente, $\mathbf{h}_t$, que hace referencia al vector acústico que incluye el efecto de la distorsión convolucional. Asumiendo que el filtro del modelo de degradación es invariante en el tiempo y que el correspondiente ruido aditivo es estacionario e icorrelado con la señal limpia filtrada, la función $f(\mathbf{x}_t, \mathbf{n}_t, \mathbf{h}_t)$ toma la siguiente expresión

$$f(\mathbf{x}_t, \mathbf{n}_t, \mathbf{h}_t) = \mathbf{h}_t + IDFT\{log(1 + e^{DFT\{\mathbf{n}_t - \mathbf{h}_t - \mathbf{x}_t\}})\}, \tag{5.2}$$

donde DFT, *Discrete Fourier Transform*, es la transformada discreta de Fourier, e IDFT, *Inverse DFT*, es la transformada discreta de Fourier inversa. Llegados a este punto, y dejando a un lado la aproximación de invariabilidad temporal asumida anteriormente para la distorsión convolucional, cabe destacar como la naturaleza aleatoria de los dos tipos de alteraciones incluidos en el modelo simplificado de degradación hace que, para cada instante de tiempo, $\mathbf{h}_t$ y $\mathbf{n}_t$ puedan tener distintas expresiones, de manera que se genera lo que se denomina incertidumbre entre los vectores de características limpios y ruidosos, o, dicho de otro modo, que distintos vectores de características ruidosos pueden provenir de la misma trama limpia y viceversa. La incertidumbre es pues, en última instancia, el gran problema que tienen las técnicas de normalización de vectores de características basadas en la aplicación de una función de transformación dependiente del vector de características contaminado, puesto que para cada uno de ellos se asociaría siempre la misma trama como estimación del vector acústico limpio.

En la Figura 5.1 se pueden apreciar los efectos reales que distintos entornos acústicos producen sobre el primer coeficiente MFCC de la señal de voz. La elección de este coeficiente se debe a que es el que posee mayor varianza e importancia de cara al RAH. Asimismo, el hecho de eliminar las tramas de silencio en este estudio tiene por razón el evitar los resultados sesgados que una base de datos con gran cantidad de pausas puede producir. El primer entorno acústico tratado (Figura 5.1.a) está compuesto únicamente por un filtro cuya respuesta impulsional, obtenida a partir de mediciones realizadas dentro del habitáculo de un vehículo, es de menor longitud temporal que la ventana de Hamming utilizada para calcular los vectores de características (25 ms). En este caso se puede apreciar como el histograma de la señal ruidosa (Figura 5.1.a.1) presenta un importante desplazamiento con respecto al de la señal limpia que, para este estudio, se compone del corpus de entrenamiento del entorno básico E4 de la base de datos *SpeechDat Car* en español; más allá de esto, los histogramas son bastante semejantes. A su vez, en el *scattegram* correspondiente (Figura

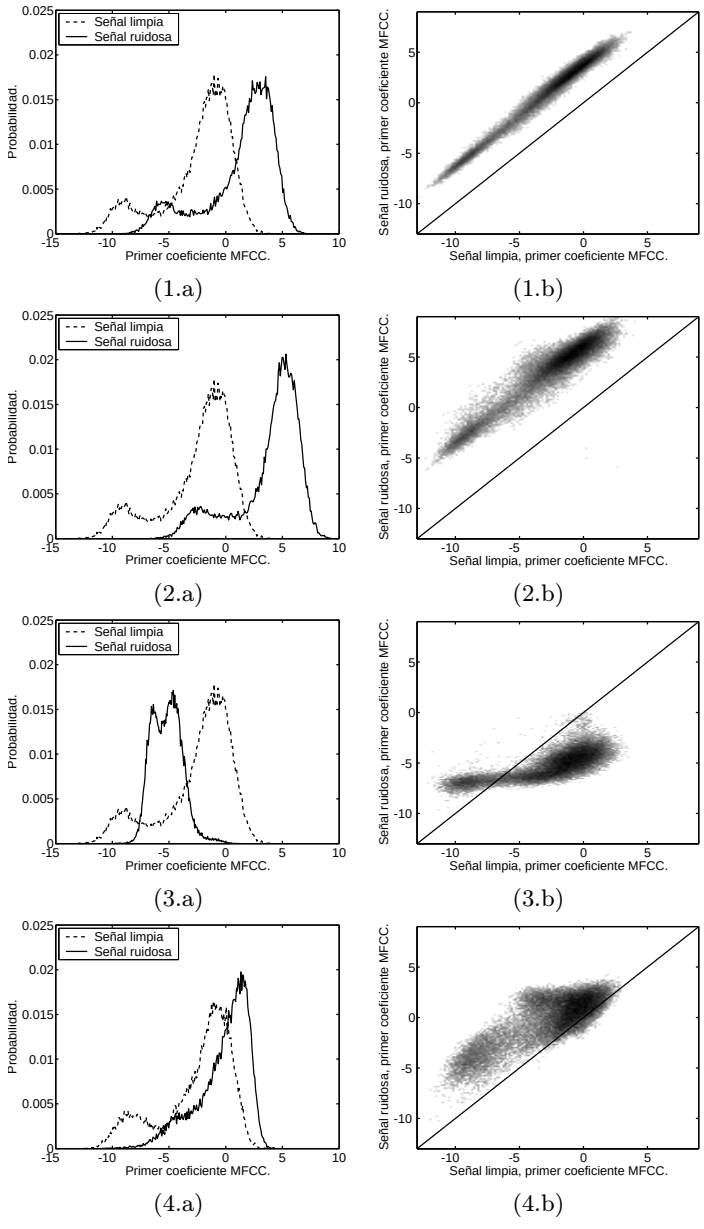


Figura 5.1: *Log-scattegrams* e histogramas del primer coeficiente MFCC de los vectores de características de voz limpia y ruidosa. Las señales limpias se corresponden con el corpus de entrenamiento del entorno básico E4 de la base de datos *SpeechDat Car* en español. Por su parte, las señales contaminadas se obtienen a partir de considerar distintos entornos acústicos: filtro con respuesta impulsional de menor longitud temporal que la ventana de Hamming empleada en el cálculo de los vectores de características (25 ms)(a), filtro con respuesta impulsional de longitud temporal mayor de 25 ms. En ambos casos la respuesta impulsional se obtuvo a partir de medidas en el habitáculo de un vehículo. El tercer entorno acústico considera únicamente ruido aditivo con 0 dB SNR (c). Finalmente el último escenario se corresponde con un entorno acústico real de un vehículo (entorno básico E4) y cuya SNR media es 8.05 dB. La línea en el *scattegram* representa la función identidad $x = y$.

5.1.a.2) queda patente el incremento de la incertidumbre, aunque en este caso es sensiblemente menor que si el entorno acústico se compone del filtro anterior interpolado, de modo que la respuesta impulsional es mayor de 25 ms (Figura 5.1.b.2). En esta ocasión queda patente no sólo que la incertidumbre es más elevada, sino que además el histograma de la señal ruidosa presenta también un mayor desplazamiento con respecto al de la señal limpia, al mismo tiempo que se ven afectadas la varianza y su propia forma (Figura 5.1.b.1). De todo lo anterior se puede concluir que los efectos producidos por la distorsión convolucional en los vectores de características no son independientes de la longitud de la respuesta impulsional y que dicha distorsión no produce, como en muchas ocasiones se asume, únicamente un desplazamiento constante para todos los vectores de características en el dominio MFCC.

En la Figura 5.1.c, la señal contaminada correspondiente se genera añadiando artificialmente ruido aditivo con una relación señal a ruido de 0 dB. En este caso se puede apreciar como el histograma de la señal contaminada, Figura 5.1.c.1, se ha modificado drásticamente de una forma no lineal con respecto al asociado a la señal limpia, reduciendo la varianza y compactando los dos modos que poseía la señal limpia a prácticamente uno. Por su parte, la incertidumbre, Figura 5.1.c.2, en este caso es mucho mayor que en los dos anteriores, ampliando sensiblemente el rango de posibles valores de los coeficientes de las tramas limpias asociados a un mismo valor para el coeficiente del vector de características ruidoso y viceversa, lo que complica enormemente la tarea de RAH.

Ya por último, en un escenario real, donde la señal ruidosa se obtiene en este caso mediante la grabación en un vehículo que circula a baja velocidad por un pavimento en mal estado (corpus de entrenamiento del entorno básico E4 de la base de datos *SpeechDat Car* en español), se puede apreciar que el histograma correspondiente presenta un cierto desplazamiento a la vez que su varianza se ve reducida con respecto a la que se obtendría a partir de los coeficientes de la señal limpia, Figura 5.1.d.1. Por su parte, la incertidumbre también se ve sensiblemente incrementada, Figura 5.1.d.2, aunque sin llegar al nivel del entorno acústico tratado anteriormente, donde la relación señal a ruido, 0dB, era sensiblemente menor a la que se tiene en este caso, 8.05dB en media.

Así pues, de todo lo anterior se puede concluir que la distorsión convolucional produce principalmente y en media un desplazamiento de los coeficientes de los vectores de características en el dominio MFCC, siendo este hecho más definido conforme la longitud temporal de la respuesta impulsional se acorte, lo que trae consigo a la vez una menor incertidumbre. Si la longitud temporal de la respuesta impulsional aumenta, nuevos efectos se unen al desplazamiento de de los coeficientes, como por ejemplo la reducción de la varianza y una mayor incertidumbre. Por su parte, el ruido aditivo afecta en mayor medida y en media a la reducción de la varianza de los coeficientes de los vectores de características en el dominio MFCC, a la vez que se incrementa de un modo importante la incertidumbre. Finalmente y de un modo general, se puede decir que en los entornos reales se dan conjuntamente los efectos ya comentados asociados tanto a la distorsión convolucional como al ruido aditivo.

5.2. Técnicas de Normalización Empíricas Basadas en MMSE.

Independientemente de que una técnica de normalización de vectores de características se pueda incluir en un grupo u otro dentro de la clasificación que se ha presentado en el Capítulo 3 (filtrado

paso alto, basadas en modelos o empíricas), la expresión final que se obtiene para el vector de características normalizado se basa en muchas ocasiones, tal y como ya se ha adelantado, en un estimador Bayesiano. Para ello es preciso suponer una determinada función de densidad de probabilidad a priori de la variable que se pretenda calcular. De esta manera el uso de estimadores Bayesianos, siempre y cuando la pdf supuesta se aproxime a la real, suele proporcionar mejores resultados que si se empleasen las técnicas clásicas de estimación ya que, en cierto modo, se acotan los posibles valores de la variable objeto de estudio.

En muchas ocasiones, de entre los estimadores Bayesianos, se elige aquél que minimiza el error cuadrático medio sobre todas las realizaciones, MSE, *Mean Square Error*. A dicho estimador Bayesiano se le denomina MMSE, *Minimum Mean Square Error* y se puede comprobar que la expresión óptima para el mismo es, en este caso concreto, la media de la pdf del vector de características limpio, variable que se trata de estimar, dado el ruidoso, variable que se considera accesible. Por otra parte, hay que tener siempre presente que los resultados obtenidos con el estimador MMSE, al igual que los proporcionados con cualquier otro estimador Bayesiano, dependerán tanto de las realizaciones de que se disponga como de la elección de la pdf a priori, cosa esta última que a menudo no suele ser sencilla.

Tal y como se ha adelantado, varias son las técnicas de normalización de vectores de características que definen el vector acústico limpio a través de un estimador Bayesiano, siendo el MMSE el más usado. Así, y sólo a modo de ejemplo, se pueden nombrar algoritmos tan dispares a simple vista como CDCN, VTS, VPS, RATZ o SPLICE, e incluso las técnicas CMN y SDCN pueden llegar también a verse de este modo. Dado que el presente trabajo está centrado en el desarrollo de técnicas de normalización de vectores de características empíricas basadas en el estimador MMSE, a continuación se presenta una visión teórica unificada de los algoritmos más empleados actualmente que reúnen estas cualidades: RATZ y SPLICE, añadiendo asimismo la técnica CMN que, aunque en el Capítulo 3 se incluyó como filtro paso alto, también se puede considerar como el algoritmo empírico más sencillo. Posteriormente, en éste y sucesivos Capítulos, y partiendo de la misma evolución teórica que se va a exponer a continuación, se derivarán las distintas técnicas que se han desarrollado a lo largo de esta tesis.

Como se ha indicado anteriormente, el estimador óptimo que minimiza el MSE Bayesiano en este caso es la media de la pdf del vector de características limpio, variable que se trata de estimar, dado el ruidoso, variable de la que se dispone. Sea pues el vector de características ruidoso para un instante de tiempo, t, $\mathbf{y}_t$, y el correspondiente limpio para el mismo instante de tiempo, $\mathbf{x}_t$. De esta manera, el vector acústico estimado obtenido mediante el criterio óptimo MMSE, $\hat{\mathbf{x}}_t$, se calculará del siguiente modo

$$\hat{\mathbf{x}}_t = E[\mathbf{x}|\mathbf{y}_t] = \int_{\mathbf{X}} \mathbf{x}\, f(\mathbf{x}|\mathbf{y}_t) d\mathbf{x}, \tag{5.3}$$

donde $E[\bullet]$ representa la esperanza de $\bullet$ y $f(\mathbf{x}|\mathbf{y}_t)$ es la pdf a priori de $\mathbf{x}$ dado $\mathbf{y}_t$. Llegados a este punto, el modo en que se aproximen tanto $\mathbf{x}$ como la pdf a priori $f(\mathbf{x}|\mathbf{y}_t)$, definirá los diversos algoritmos de normalización de vectores de características basados en el criterio MMSE.

La técnica CMN no asume ninguna expresión específica para la pdf $f(\mathbf{x}|\mathbf{y}_t)$ y el vector de características limpio se aproxima mediante la siguiente transformación $\mathbf{x} \approx \mathbf{\Psi}(\mathbf{y}_t, \mathbf{r}) = \mathbf{y}_t - \mathbf{r}$, donde $\mathbf{r}$ se puede ver como el vector de desplazamiento entre $\mathbf{y}_t$ y $\mathbf{x}$, contemplando pues únicamente un desplazamiento general entre los vectores de características. Así pues, y on estas premisas, la expresión (5.3) para CMN se transforma en

$$\hat{\mathbf{x}}_t \approx \int_{\mathbf{X}} (\mathbf{y}_t - \mathbf{r}) p(\mathbf{x}|\mathbf{y}_t) d\mathbf{x} = \mathbf{y}_t - \mathbf{r}. \tag{5.4}$$

Para estimar el vector de desplazamiento, $\mathbf{r}$, se define el error cuadrático medio, ξ, idealmente sobre todas las realizaciones ($t \in [1, T]$), y se minimiza con respecto a $\mathbf{r}$

$$\xi = \frac{1}{T} \sum_t Tra[(\mathbf{x}_t - \mathbf{\Psi}(\mathbf{y}_t, \mathbf{r}))(\mathbf{x}_t - \mathbf{\Psi}(\mathbf{y}_t, \mathbf{r}))^T], \tag{5.5}$$

$$\mathbf{r} = \underset{\mathbf{r}}{arg\,min}(\xi) = \frac{1}{T} \sum_t \mathbf{y_t} - \mathbf{x}_t = E[\mathbf{y}] - E[\mathbf{x}], \tag{5.6}$$

donde $Tra[\bullet]$ es la traza de $\bullet$. El desarrollo para obtener la expresión (5.6) a partir de (5.5) se encuentra en el Anexo 5.5, situado al final del presente Capítulo. Se puede apreciar que lo que propone esta técnica es suprimir la media de los vectores de características ruidosos y añadir la de las tramas limpias, de modo que al final la media de los vectores de características normalizados coincida con la de los limpios. En algunos casos se suele sustraer a los vectores de características limpios del corpus de entrenamiento su propia media, de modo que los correspondientes modelos acústicos se entrenan con señales ya normalizadas en media, haciéndose por tanto innecesario estimar $E[\mathbf{x}]$ a la hora de normalizar la señal ruidosa mediante el método CMN debido a que el espacio de referencia ya posee media nula; en ese caso el vector de desplazamiento será únicamente $\mathbf{r} = E[\mathbf{y}]$. Por su parte, la estimación de $E[\mathbf{y}]$ en un instante t se suele realizar mediante métodos iterativos considerando los $t-1$ vectores de características ruidosos anteriores, si se pretende proporcionar una solución en tiempo real, o bien se puede hacer uso de toda la frase pronunciada, si se permite dicho retardo. En la actualidad, bien el algoritmo básico CMN o una extensión del mismo se aplica en casi todos los sistemas de RAH ya que es una técnica muy sencilla y de bajo coste computacional que, aunque no soluciona ni de lejos el problema de la robustez, sí que aporta una interesante mejora al comportamiento del sistema ayudando a compensar especialmente la distorsión convolucional. La versión de CMN expuesta en este apartado es la más sencilla de todas las posibles ya que la transformación propuesta no incluye ningún tipo de dependencia con respecto a la naturaleza del vector de características ruidoso. Esta limitación se ve en muchas ocasiones unida a que se considera que la distorsión convolucional es invariable en el tiempo y, por tanto, se puede compensar con un vector de desplazamiento $\mathbf{r}$ fijo, lo que de un modo indirecto supone que el efecto de un filtro afecta únicamente a la media de los vectores de características de una manera constante, cosa que, por otra parte, ya se ha constatado que no es cierta ni siquiera con presencia de ruido convolucional controlado (Sección 5.1). Estas claras limitaciones que posee la técnica CMN las trata de compensar el algoritmo RATZ incluyendo restricciones en la pdf a priori $f(\mathbf{x}|\mathbf{y}_t)$ y modificando igualmente la estimación del vector de características limpio.

El método RATZ considera dos aproximaciones. La primera de ellas consiste en asumir que el espacio limpio se puede modelar mediante una mezcla de Gaussianas, GMM, *Gaussian Mixture Model*

$$p(\mathbf{x}) = \sum_{s_x} p(\mathbf{x}|s_x) p(s_x), \tag{5.7}$$

$$p(\mathbf{x}|s_x) = \mathcal{N}(\mathbf{x}; \mu_{s_x}, \mathbf{\Sigma}_{s_x}), \tag{5.8}$$

donde μ_{s_x}, $\mathbf{\Sigma}_{s_x}$ y $p(s_x)$ son el vector de medias, la matriz diagonal de covarianzas y la probabilidad a priori asociados a la Gaussiana del modelo limpio s_x. La segunda aproximación del algoritmo RATZ es considerar que el vector de características limpio se puede estimar mediante la siguiente transformación $\mathbf{x} \approx \mathbf{\Psi}(\mathbf{y}_t, \mathbf{r}_{s_x}) = \mathbf{y}_t - \mathbf{r}_{s_x}$, siendo $\mathbf{r}_{s_x}$ el vector de desplazamiento entre $\mathbf{y}_t$ y $\mathbf{x}$

asociado a la Gaussiana s_x; de modo que la transformación propuesta consiste nuevamente en un desplazamiento, aunque en este caso dependiente de la Gaussiana del modelo GMM del espacio limpio. Así pues, y haciendo uso de las dos aproximaciones anteriores, la ecuación (5.3) para RATZ se transforma en

$$\hat{\mathbf{x}}_t \approx \int_{\mathbf{X}} \sum_{s_x} (\mathbf{y}_t - \mathbf{r}_{s_x}) p(\mathbf{x}, s_x|\mathbf{y}_t) d\mathbf{x} = \mathbf{y}_t - \sum_{s_x} \mathbf{r}_{s_x} p(s_x|\mathbf{y}_t), \tag{5.9}$$

donde $p(s_x|\mathbf{y}_t)$ es la probabilidad a posteriori de la Gaussiana del modelo limpio, s_x, dado el vector de características ruidoso $\mathbf{y}_t$. Dicha probabilidad se puede estimar a partir de (5.7) y (5.8) asumiendo además que el ruido produce un efecto aditivo en el dominio MFCC [Moreno, 1996]. De este modo se genera un pseudo-modelo GMM que representa el espacio ruidoso, lo que no deja de ser una aproximación que en muchas ocasiones se aleja de la realidad, y que es el que realmente se emplea para estimar $p(s_x|\mathbf{y}_t)$.

Por su parte, el cálculo del vector de desplazamiento, que se realizará en una fase de entrenamiento previa, requiere de señal estéreo (aunque también existe una versión "ciega" [Moreno, 1996] que no la precisa). Sea pues un corpus de entrenamiento estéreo, $(\mathbf{X}^{Tr}, \mathbf{Y}^{Tr}) = \{(\mathbf{x}_1^{Tr}, \mathbf{y}_1^{Tr}); ...; (\mathbf{x}_t^{Tr}, \mathbf{y}_t^{Tr}); ...; (\mathbf{x}_T^{Tr}, \mathbf{y}_T^{Tr})\}$, compuesto por T pares de vectores de características limpio-ruidoso $(\mathbf{x}_t^{Tr}, \mathbf{y}_t^{Tr})$. De este modo, el vector de desplazamiento asociado a la Gaussiana s_x, $\mathbf{r}_{s_x}$, se estima tras minimizar el error cuadrático medio asociado a la Gaussiana correspondiente, ξ_{s_x}, con respecto a $\mathbf{r}_{s_x}$

$$\xi_{s_x} = \frac{1}{T} \sum_t p(s_x|\mathbf{x}_t^{Tr}) Tra[(\mathbf{x}_t^{Tr} - \mathbf{\Psi}(\mathbf{y}_t^{Tr}, \mathbf{r}_{s_x}))(\mathbf{x}_t^{Tr} - \mathbf{\Psi}(\mathbf{y}_t^{Tr}, \mathbf{r}_{s_x}))^T], \tag{5.10}$$

$$\mathbf{r}_{s_x} = \underset{\mathbf{r}_{s_x}}{arg\,min}(\xi_{s_x}) = \frac{\sum_t p(s_x|\mathbf{x}_t^{Tr})(\mathbf{y}_t^{Tr} - \mathbf{x}_t^{Tr})}{\sum_t p(s_x|\mathbf{x}_t^{Tr})}, \tag{5.11}$$

donde $p(s_x|\mathbf{x}_t^{Tr})$ es la probabilidad a posteriori de la Gaussiana del modelo limpio, s_x, dado el vector de características limpio del corpus de entrenamiento, $\mathbf{x}_t^{Tr}$. Dicha probabilidad se puede calcular a partir de (5.7) y (5.8). Por otra parte, el desarrollo teórico para obtener (5.11) a partir de 5.10 se encuentra en el Anexo 5.5 del presente Capítulo.

$$p(s_x|\mathbf{x}_t^{Tr}) = \frac{p(\mathbf{x}_t^{Tr}|s_x)p(s_x)}{\sum_{s_x} p(\mathbf{x}_t^{Tr}|s_x)p(s_x)}. \tag{5.12}$$

Debido a la utilización de transformaciones más específicas, la utilización del algoritmo RATZ mejora sensiblemente las prestaciones de los sistemas de RAH con respecto al empleo de la técnica CMN. Sin embargo, el modo en que se estima $p(s_x|\mathbf{y}_t)$ puede producir serios desajustes ya que asume un modelo de degradación en el dominio MFCC no del todo realista para transformar cada Gaussiana del espacio limpio al ruidoso. Para eliminar este modelo de degradación, el algoritmo SPLICE asume dos aproximaciones distintas: la primera de ellas consiste en modelar mediante una GMM el espacio ruidoso en lugar del limpio, como hace el método RATZ

$$p(\mathbf{y}_t) = \sum_{s_y} p(\mathbf{y}_t|s_y)p(s_y), \tag{5.13}$$

$$p(\mathbf{y}_t|s_y) = \mathcal{N}(\mathbf{y}_t; \mu_{s_y}, \mathbf{\Sigma}_{s_y}), \tag{5.14}$$

donde s_y se corresponde con el índice de la Gaussiana del espacio ruidoso y μ_{s_y}, $\mathbf{\Sigma}_{s_y}$ y $p(s_y)$ son el vector de medias, la matriz diagonal de covarianzas y la probabilidad a priori de la Gaussiana s_y. La segunda aproximación de la técnica SPLICE consiste en asumir que el vector de características limpio se puede expresar mediante la siguiente transformación $\mathbf{x} \approx \mathbf{\Psi}(\mathbf{y}_t, \mathbf{r}_{s_y}) = \mathbf{y}_t - \mathbf{r}_{s_y}$, donde

$\mathbf{r}_{s_y}$ es el vector de desplazamiento entre el vector de características ruidoso, $\mathbf{y}_t$, y el limpio, $\mathbf{x}$, asociado a la Gaussiana del modelo ruidoso s_y. Nuevamente la transformación propuesta incluye únicamente un término de desplazamiento, en este caso dependiente de la Gaussiana del modelo del espacio ruidoso. De este modo, y haciendo uso de las dos aproximaciones anteriores, la ecuación (5.3) para el algoritmo SPLICE se transforma en

$$\hat{\mathbf{x}}_t \approx \int_{\mathbf{X}} \sum_{s_y} (\mathbf{y}_t - \mathbf{r}_{s_y}) p(\mathbf{x}, s_y|\mathbf{y}_t) d\mathbf{x} = \mathbf{y}_t - \sum_{s_y} \mathbf{r}_{s_y} p(s_y|\mathbf{y}_t), \tag{5.15}$$

donde $p(s_y|\mathbf{y}_t)$ es la probabilidad a posteriori de la Gaussiana del modelo ruidoso, s_y, dado el vector de características degradado $\mathbf{y}_t$. Dicha probabilidad se puede calcular a partir de (5.13) y (5.14)

$$p(s_y|\mathbf{y}_t) = \frac{p(\mathbf{y}_t|s_y)p(s_y)}{\sum_{s_y} p(\mathbf{y}_t|s_y)p(s_y)}. \tag{5.16}$$

Por su parte, y a la hora de estimar el vector de desplazamiento $\mathbf{r}_{s_y}$, al igual que para el caso de la técnica RATZ, se precisa de una fase de entrenammiento previa en la que se hace uso de señal estéreo $(\mathbf{X}^{Tr}, \mathbf{Y}^{Tr}) = \{(\mathbf{x}_1^{Tr}, \mathbf{y}_1^{Tr}); ...; (\mathbf{x}_t^{Tr}, \mathbf{y}_t^{Tr}); ...; (\mathbf{x}_T^{Tr}, \mathbf{y}_T^{Tr})\}$. Para ello se minimiza el error cuadrático medio asociado a la Gaussiana correspondiente, ξ_{s_y}, con respecto a $\mathbf{r}_{s_y}$

$$\xi_{s_y} = \frac{1}{T} \sum_t p(s_y|\mathbf{y}_t^{Tr}) Tra[(\mathbf{x}_t^{Tr} - \mathbf{\Psi}(\mathbf{y}_t^{Tr}, \mathbf{r}_{s_y}))(\mathbf{x}_t^{Tr} - \mathbf{\Psi}(\mathbf{y}_t^{Tr}, \mathbf{r}_{s_y}))^T], \tag{5.17}$$

$$\mathbf{r}_{s_y} = \underset{\mathbf{r}_{s_y}}{arg\,min}(\xi_{s_y}) = \frac{\sum_t p(s_y|\mathbf{y}_t^{Tr})(\mathbf{y}_t^{Tr} - \mathbf{x}_t^{Tr})}{\sum_t p(s_y|\mathbf{y}_t^{Tr})}. \tag{5.18}$$

El desarrollo teórico para obtener la expresión (5.18) a partir de (5.17) se encuentra en el Anexo 5.5 del presente Capítulo. A partir de lo anterior se puede apreciar como en esta ocasión no se requiere hacer presunción alguna sobre como el entorno acústico afecta a los modelos GMM del espacio limpio ya que se emplean directamente los del espacio ruidoso. De esta manera, se elimina la aproximación considerada en la técnica RATZ sobre el modelo de degradación necesario para poder estimar $p(s_x|\mathbf{y}_t)$, lo que redunda en un mejor comportamiento a la hora de reconocer por parte del algoritmo SPLICE. Por otra parte, ante espacios ruidosos muy heterogéneos, la técnica SPLICE podría, incrementando el número de Gaussianas del modelo del espacio ruidoso, obtener unos vectores de desplazamiento más selectivos de los que podría proporcionar la técnica RATZ, aunque en ésta se ampliara el número de componentes de las GMMs con que se modela el espacio limpio.

En muchas ocasiones las señales que se pretende normalizar pertenecen a espacios acústicos altamente variables. En estos casos, y para métodos como RATZ y SPLICE con un modelado GMM del espacio ruidoso compuesto por pocas Gaussianas, puede darse el caso de que los vectores de desplazamiento entrenados sean excesivamente genéricos. Para solucionar este inconveniente se suele dividir el espacio acústico ruidoso en varios entornos básicos e con propiedades acústicas similares (relación señal a ruido, componentes espectrales ...) de modo que para cada uno de ellos se estiman los vectores de desplazamiento correspondientes a las distintas técnicas de normalización de forma independiente. Esta variación multi-entorno de las técnicas clásicas de normalización produce unos algoritmos más robustos, como *Interpolate* RATZ, IRATZ, [Moreno, 1996] o SPLICE con selección del modelo de entorno, *SPLICE with environmental model selection* [Droppo *et al.*, 2001]. En ambos casos, a la hora de obtener una estimación del vector de características limpio se puede hacer uso de todos los vectores de desplazamiento ponderándolos por la probabilidad a posteriori de

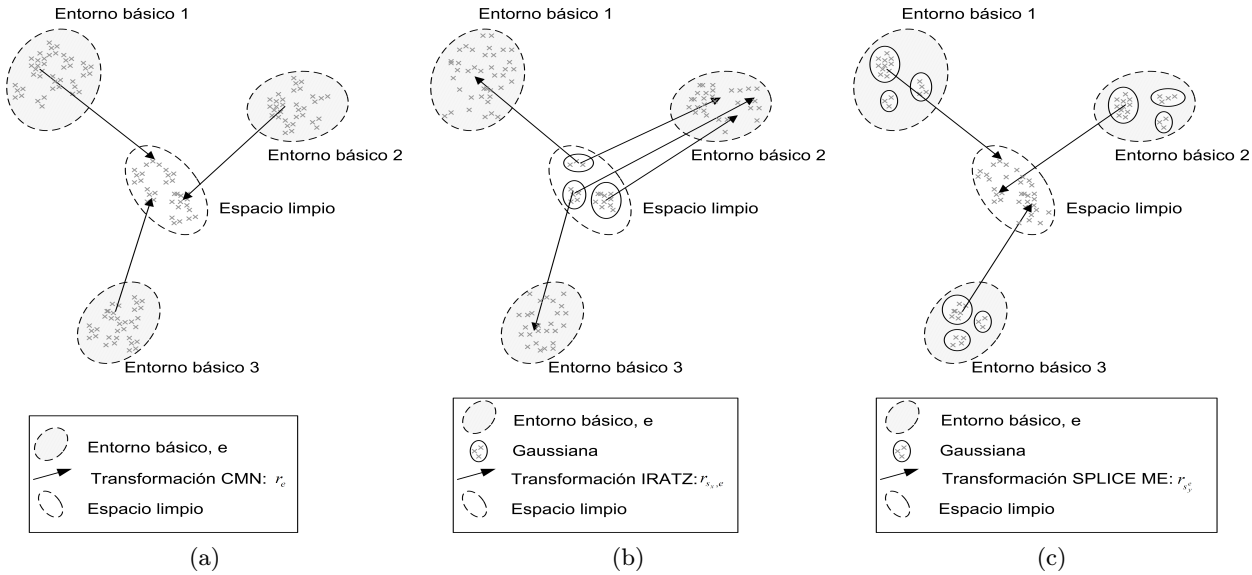


Figura 5.2: Representación gráfica de las técnicas CMN (a), RATZ (b) y SLICE (c), donde $\mathbf{r}_e$, $\mathbf{r}_{s_x,e}$ y $\mathbf{r}_{s_y^e}$ son los vectores de desplazamiento respectivos para cada enorno básico e.

encontrarse en cada entorno básico y Gaussiana dado el vector acústico ruidoso, $p(e, s_x|\mathbf{y}_t)$ para la técnica IRATZ o $p(e, s_y|\mathbf{y}_t)$ para el método SPLICE con selección del modelo de entorno, (decisión *soft*), o bien se puede emplear únicamente aquellos vectores de desplazamiento del entorno básico más probable, $\hat{e}$, (decisión *hard*). Igualmente, y haciendo uso de la misma filosofía, se podría pensar en una versión multi-entorno para la técnica CMN. A modo de resumen se incluye la Figura 5.2, en la que se representan los esquemas gráficos de las extensiones de las técnicas CMN, RATZ y SPLICE para varios entornos básicos, pudiéndose apreciar para cada caso el dominio de actuación de los vectores de desplazamiento correspondientes: desde el más amplio (extensión de la técnica CMN), hasta los más reducidos (algoritmos IRATZ y SPLICE con selección del modelo de entorno).

5.3. Técnica *Multi-Environment Model-based LInear Normalization*, MEMLIN.

Dejando a un margen la técnica CMN por su simplicidad, se ha podido comprobar en la Sección 5.2 que el método RATZ, a pesar de hacer uso de unas transformaciones más selectivas que el algoritmo CMN, tiene una debilidad a la hora de estimar la probabilidad a posteriori de la Gaussiana del modelo limpio dado el vector de características ruidoso, ya que asume un cierto modelo de degradación que en muchos casos se aleja del real. Este problema, tal y como ha quedado igualmente patente en la Sección 5.2, se solventa en el método SPLICE al modelar el espacio ruidoso en lugar del limpio; sin embargo, las transformaciones propuestas en la técnica SPLICE, dependientes de cada Gaussiana del espacio ruidoso, no son todo lo específicas que se podría desear. Si se considera la GMM que representa el espacio ruidoso como un modelo de generación, se podría observar que los vectores de características producidos por una Gaussiana del espacio ruidoso tienen asociados unas tramas limpias que, en general, y debido a la aleatoriedad del ruido del entorno acústico, no se encuentran concentradas en una determinada región del espacio limpio, sino que se distribuyen en

mayor o menor medida por todo él. La misma conclusión se podría extraer si se modelara el espacio limpio mediante una GMM y se supusiera un determinado entorno acústico ruidoso real. Este hecho, si bien se podía intuir de los *scattegrams* presentados en la Sección 5.2, queda totalmente patente en la Figura 5.3, en la que se representa el histograma en dos dimensiones de los pares de Gaussianas más probables obtenidos a partir de la señal estéreo del corpus de entrenamiento del entorno básico E4 de la base de datos *SpeechDat Car* en español; para ello se modeló tanto el espacio limpio como el ruidoso con sendas GMMs compuestas por 16 Gaussianas cada una, generando posteriormente el histograma a partir de los distintos pares de Gaussianas más probables asociados a las parejas de vectores de características. Se puede apreciar como, dada una Gaussiana del modelo limpio como la más probable, pueden ser múltiples las Gaussianas del modelo ruidoso con mayor probabilidad y viceversa. Esto demuestra que modelar únicamente el espacio limpio, como sucede en el método RATZ, o el ruidoso, como se realiza en el algoritmo SPLICE, puede dar lugar a que la señal ruidosa y limpia utilizada para entrenar los vectores de desplazamientos asociados a cada Gaussiana para las técnicas RATZ y SPLICE respectivamente cubran gran espacio, lo que proporcionaría un entrenamiento poco específico de dichos vectores de desplazamiento.

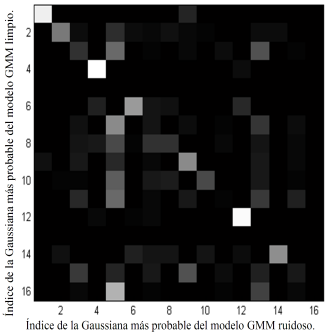


Figura 5.3: Parejas de Gaussianas más probables asociadas a señal estéreo (señal limpia y ruidosa pertenecientes al corpus de entrenamiento del entorno básico E4 de la base de datos *SpeechDat Car* en español). El eje de las abscisas representa el índice de la Gaussiana del modelo ruidoso, y el eje de las ordenadas representa el índice de la Gaussiana del modelo limpio. Ambos modelos constan de 16 Gaussianas. Cuanto más blanca sea la representación, mayor es el número de pares de vectores de características asociados a esa pareja concreta de Gaussianas.

Para solventar el problema de falta de especificidad a la hora de entrenar los vectores de desplazamiento, en este trabajo se propone modelar tanto el espacio limpio como el ruidoso y entrenar de este modo transformaciones asociadas a cada par de Gaussianas, entendiendo por par de Gaussianas la unión de una del modelo del espacio limpio y otra del modelo del espacio degradado. A esta nueva técnica, que también se va a servir del criterio MMSE para estimar el vector de características limpio, se la denomina MEMLIN, *Multi-Environment Model-based LInear Normalization*, [Buera *et al.*, 2004a] y se apoya en tres aproximaciones básicas.

- Para generalizar, y ante la posibilidad de que el espacio ruidoso pueda ser muy heterogéneo, éste se divide en una serie de entornos básicos, e, de modo que los vectores de características degradados $\mathbf{y}_t$ se modelan para cada uno de ellos mediante una mezcla de Gaussianas, GMM

$$p_e(\mathbf{y}_t) = \sum_{s_y^e} p(\mathbf{y}_t|s_y^e)p(s_y^e), \tag{5.19}$$

$$p(\mathbf{y}_t|s_y^e) = \mathcal{N}(\mathbf{y}_t; \mu_{s_y^e}, \mathbf{\Sigma}_{s_y^e}), \tag{5.20}$$

donde s_y^e hace referencia a la correspondiente Gaussiana del modelo ruidoso del entorno básico e, mientras que $\mu_{s_y^e}$, $\mathbf{\Sigma}_{s_y^e}$, y $p(s_y^e)$ son el vector media, la matriz diagonal de covarianzas y la probabilidad a priori asociados a s_y^e.

- Los vectores de características limpios se modelan mediante una GMM: expresiones (5.7) y (5.8).
- Por otra parte, y al igual que en las técnicas CMN, RATZ o SPLICE, el vector de características limpio se aproxima mediante una función lineal del vector de características ruidoso, aunque en este caso dicha función depende del entorno básico y de las Gaussianas de las GMMs limpia y ruidosa: $\mathbf{x} \approx \Psi(\mathbf{y}_t, s_x, s_y^e) = \mathbf{y}_t - \mathbf{r}_{s_x,s_y^e}$, donde $\mathbf{r}_{s_x,s_y^e}$ es el vector de desplazamiento entre las tramas $\mathbf{y}_t$ y $\mathbf{x}$ asociado al par de Gaussianas s_x y s_y^e. Así pues, se puede apreciar como también en el método MEMLIN la transformación propuesta únicamente incluye un factor de desplazamiento.

A partir de todo lo anterior, y haciendo uso de las tres aproximaciones en las que se basa la técnica MEMLIN, la expresión (5.3) se transforma en el caso de la técnica MEMLIN en

$$\hat{\mathbf{x}}_t = \int_{\mathbf{X}} \sum_e \sum_{s_y^e} \sum_{s_x} (\mathbf{y}_t - \mathbf{r}_{s_x,s_y^e}) p(\mathbf{x}, s_x, e, s_y^e|\mathbf{y}_t) d\mathbf{x} = \mathbf{y}_t - \sum_e \sum_{s_y^e} \sum_{s_x} \mathbf{r}_{s_x,s_y^e} p(e|\mathbf{y}_t) p(s_y^e|\mathbf{y}_t, e) p(s_x|\mathbf{y}_t, e, s_y^e), \tag{5.21}$$

donde $p(e|\mathbf{y}_t)$ es la probabilidad a posteriori del entorno básico dado el vector de características degradado $\mathbf{y}_t$; $p(s_y^e|\mathbf{y}_t, e)$ es la probabilidad a posteriori de la Gaussiana del modelo ruidoso del entorno básico e, s_y^e, dado el vector acústico ruidoso, $\mathbf{y}_t$, y el propio entorno básico, e. Estos dos términos se estiman trama a trama en el proceso de normalización de la señal degradada mediante las expresiones (5.19) y (5.20). Por el contrario, el modelo de la probabilidad entre Gaussianas, $p(s_x|\mathbf{y}_t, e, s_y^e)$, que es la probabilidad de la Gaussiana del modelo limpio, s_x, dado el vector de características ruidoso $\mathbf{y}_t$, el entorno básico e y la Gaussiana del modelo degradado del mismo entorno básico, s_y^e, se estima, junto con el vector de desplazamiento, $\mathbf{r}_{s_x,s_y^e}$, haciendo uso de señal estéreo en una fase de entrenamiento previa independiente para cada entorno básico.

La probabilidad a posteriori del entorno básico, $p(e|\mathbf{y}_t)$, se calcula recursivamente aplicando, tal y como ya se ha adelantado, las expresiones (5.19) y (5.20)

$$p(e|\mathbf{y}_t) = \beta \cdot p(e|\mathbf{y}_{t-1}) + (1-\beta)\frac{p_e(\mathbf{y}_t)}{\sum_e p_e(\mathbf{y}_t)}, \tag{5.22}$$

donde β es la constante de memoria ($0 \leq \beta \leq 1$), y $p(e|\mathbf{y}_0)$ se considera uniforme para todos los entornos básicos. Dado que en la mayoría de las situaciones reales se puede asumir que los entornos básicos no se suceden muy rápidamente a lo largo del tiempo, el valor de β ha de ser próximo a 1 (0.98 en todo momento para este trabajo). Por su parte, la probabilidad a posteriori de la Gaussiana del modelo ruidoso, dado el vector de características degradado $\mathbf{y}_t$ y el entorno básico e, $p(s_y^e|\mathbf{y}_t, e)$, se estima igualmente a partir de (5.19) y (5.20) del siguiente modo

$$p(s_y^e|\mathbf{y}_t, e) = \frac{p(\mathbf{y}_t|s_y^e)p(s_y^e)}{\sum_{s_y^e} p(\mathbf{y}_t|s_y^e)p(s_y^e)}. \tag{5.23}$$

Tal y como se ha comentado anteriormente, la estimación del vector de desplazamiento $\mathbf{r}_{s_x,s_y^e}$ y del modelo de la probabilidad entre Gaussianas, $p(s_x|\mathbf{y}_t, e, s_y^e)$, requiere de un proceso de entrenamiento previo independiente para cada entorno básico llevado a cabo con señal estéreo:

$(\mathbf{X}_e^{Tr}, \mathbf{Y}_e^{Tr}) = \{(\mathbf{x}_1^{Tr,e}, \mathbf{y}_1^{Tr,e}); ...; (\mathbf{x}_{t_e}^{Tr,e}, \mathbf{y}_{t_e}^{Tr,e}); ...; (\mathbf{x}_{T_e}^{Tr,e}, \mathbf{y}_{T_e}^{Tr,e})\}$, con $t_e \in [1, T_e]$. De esta manera, a la hora de calcular el vector de desplazamiento se minimiza con respecto a $\mathbf{r}_{s_x,s_y^e}$ el error cuadrático medio asociado a cada par de Gaussianas, ξ_{s_x,s_y^e}, definido del siguiente modo

$$\xi_{s_x,s_y^e} = \frac{1}{T_e} \sum_{t_e} p(s_x|\mathbf{x}_{t_e}^{Tr,e}, e) p(s_y^e|\mathbf{y}_{t_e}^{Tr,e}, e) Tra[(\mathbf{x}_{t_e}^{Tr,e} - \Psi(\mathbf{y}_t^{Tr}, s_x, s_y^e))(\mathbf{x}_{t_e}^{Tr,e} - \Psi(\mathbf{y}_t^{Tr,e}, s_x, s_y^e))^T], \tag{5.24}$$

$$\mathbf{r}_{s_x,s_y^e} = \underset{\mathbf{r}_{s_x,s_y^e}}{arg\,min}(\xi_{s_x,s_y^e}) = \frac{\sum_{t_e} p(s_x|\mathbf{x}_{t_e}^{Tr,e}, e) p(s_y^e|\mathbf{y}_{t_e}^{Tr,e}, e)(\mathbf{y}_{t_e}^{Tr,e} - \mathbf{x}_{t_e}^{Tr,e})}{\sum_{t_e} p(s_x|\mathbf{x}_{t_e}^{Tr,e}, e) p(s_y^e|\mathbf{y}_{t_e}^{Tr,e}, e)}, \tag{5.25}$$

donde $p(s_x|\mathbf{x}_{t_e}^{Tr,e}, e)$ es la probabilidad a posteriori de la Gaussiana del modelo limpio, s_x, dado el vector de características limpio del corpus de entrenamiento, $\mathbf{x}_{t_e}^{Tr,e}$, y el entorno básico, e. Dicha probabilidad se estima haciendo uso de las expresiones (5.7) y (5.8). El desarrollo teórico para obtener la expresión (5.25) a partir de 5.24 se encuentra en el Anexo 5.5 en este mismo Capítulo.

$$p(s_x|\mathbf{x}_{t_e}^{Tr,e}, e) = \frac{p(\mathbf{x}_{t_e}^{Tr,e}|s_x) p(s_x)}{\sum_{s_x} p(\mathbf{x}_{t_e}^{Tr,e}|s_x) p(s_x)}. \tag{5.26}$$

Por su parte, el modelo de la probabilidad entre Gaussianas, $p(s_x|\mathbf{y}_t, e, s_y^e)$, se simplifica eliminando la dependencia temporal proporcionada por el vector de características ruidoso, $\mathbf{y}_t$, de modo que el término que finalmente se debe estimar en la fase de entrenamiento previa es $p(s_x|e, s_y^e)$, que se puede obtener mediante frecuencia relativa, solución *hard*, o bien empleando (5.19), (5.20), (5.7) y (5.8), decisión *soft*. Así pues, la correspondiente expresión para la decisión *hard* es

$$p(s_x|e, s_y^e) = \frac{C_N(s_x|s_y^e)}{N_{s_y^e}}, \tag{5.27}$$

donde $C_N(s_x|s_y^e)$ es el número de veces que el par de Gaussianas más probables es s_x y s_y^e para todas las parejas de vectores de características del corpus de entrenamiento del entorno básico e; por otra parte $N_{s_y^e}$ es el número de veces que la Gaussiana más probable del modelo ruidoso es s_y^e para todos los vectores acústicos degradados del entorno básico e.

La estimación del modelo de la probabilidad entre Gaussianas usando la estimación *soft* se calcula del siguiente modo

$$p(s_x|e, s_y^e) = \frac{\sum_{t_e} p(\mathbf{x}_{t_e}^{Tr,e}|s_x) p(\mathbf{y}_{t_e}^{Tr,e}|s_y^e) p(s_x) p(s_y^e)}{\sum_{t_e} \sum_{s_x} p(\mathbf{x}_{t_e}^{Tr,e}|s_x) p(\mathbf{y}_{t_e}^{Tr,e}|s_y^e) p(s_x) p(s_y^e)}. \tag{5.28}$$

Cuando hay suficientes datos para la estimación del modelo de la probabilidad entre Gaussianas, tanto la solución *soft* como la *hard* obtienen similares resultados en términos de RAH. Sin embargo, en algunas ocasiones puede no haber suficientes datos y en ese caso la opción *soft* proporciona una solución más consistente. En este trabajo todos los experimentos realizados con MEMLIN se llevaron a cabo haciendo uso de la opción *hard*.

A modo de resumen se incluye una representación gráfica del algoritmo MEMLIN (Figura 5.4), en el que se puede apreciar el radio de acción del correspondiente vector de desplazamiento para este método, r_{s_x,s_y}. Nótese igualmente la diferencia entre el espacio de proyección asociado a r_{s_x,s_y}, con mucha menor incertidumbre, y los correspondientes a los distintos vectores de desplazamiento de las técnicas CMN, RATZ y SPLICE (Figura 5.2). Esto supone la principal ventaja del método

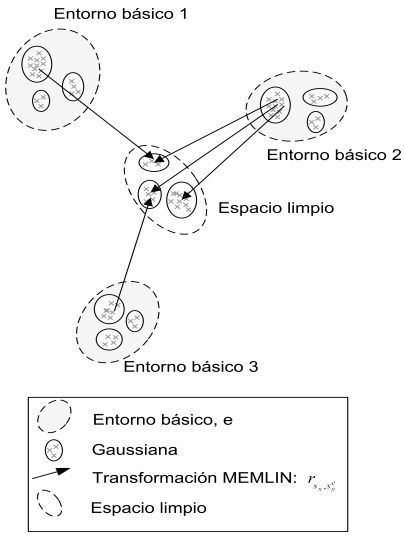


Figura 5.4: Representación gráfica de la técnica MEMLIN, donde $\mathbf{r}_{s_x,s_y^e}$ es el vector de desplazamiento asociado al par de Gaussianas s_x y s_y^e.

MEMLIN si se compara con el resto de técnicas. Este hecho queda patente también en la Figura 5.3, donde el espacio de proyección asociado a cada Gaussiana del modelo limpio (técnica RATZ) o del modelo ruidoso (método SPLICE) ocupa buena parte del espacio ruidoso o limpio, respectivamente, mientras que en el algoritmo MEMLIN el espacio de proyección se circunscribe a nivel de Gaussiana.

5.4. Resultados con la base de datos *SpeechDat Car* en español.

La experimentación comparativa de las técnicas de normalización de vectores de características empíricas tratadas en las Secciones 5.2 5.3 se realizó con la base de datos *SpeechDat Car* en español, que, como ya se indicó en la Sección 4.1 está dividida, además de por canales, en dos corpora: entrenamiento y reconocimiento, que se distribuyen atendiendo a distintos entornos básicos. Así pues, a la hora de realizar el proceso previo de entrenamiento para obtener los correspondientes parámetros necesarios para las distintas técnicas de normalización y entornos básicos, esto es, los vectores de desplazamiento y los modelos de probabilidad cruzada, esto último sólo en el caso del método MEMLIN, se hará uso del corpus de entrenamiento correspondiente a cada entorno básico. Por otra parte, y una vez que se ha llevado a cabo la normalización de los vectores acústicos degradados con las correspondientes técnicas, se aplicará el método CMS. Para esta experimentación se utilizó la *parametrización UZ* y los modelos acústicos de las unidades fonéticas, pudiéndose, de este modo, consultar los resultados de referencia correspondientes en la Tabla 4.3. En la Figura 5.5 se incluyen, de un modo gráfico, los tres pasos precisados para llevar a cabo la experimentación. Así, primeramente es necesario, en una fase de entrenamiento previo ("Entrenamiento"), estimar los diversos parámetros necesarios para las distintas técnicas de normalización, para lo que, tal y como se ha comentado, se hace uso del corpus de entrenamiento estéreo. El segundo paso consiste en estimar los correspondientes vectores de características limpios a partir de los ruidosos ("Normalización"), para, finalmente y en la última fase, decodificar la señal normalizada empleando los modelos acústicos que representan al espacio limpio.

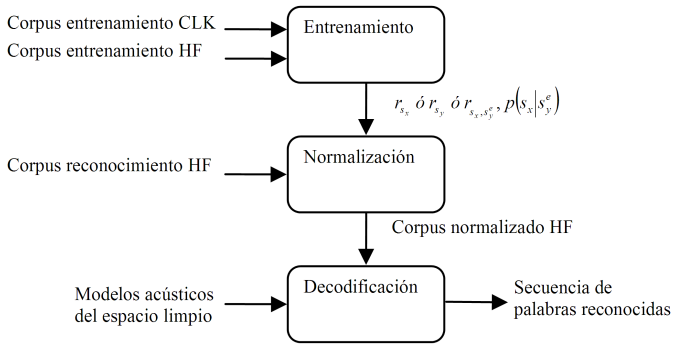


Figura 5.5: Esquema general de la experimentación realizada para las distintas técnicas de normalización de vectores de características empíricas. Se distinguen tres pasos, a saber, la fase previa de entrenamiento de las distintas técnicas de normalización de los vectores de características ("Entrenamiento"), para la que se supone en este caso el uso de señal estéreo. El segundo paso se corresponde con la estimación del vector acústico limpio ("Normalización"). Finalmente, la última fase consiste en la decodificiación de la señal normalizada haciendo uso de los modelos acústicos del espacio limpio ("Decodoficación").

En la Tabla 5.1 se pueden apreciar los mejores resultados para las distintas técnicas de normalización de vectores de características comparadas. Junto al nombre de la técnica (IRATZ, SPLICE con selección de modelo de entorno, que se identifica como SPLICE ME, y MEMLIN), se incluye el número de componentes que conforman las GMMs necesarias para cada algoritmo (se realizó el siguiente barrido del número de componentes: 8, 16, 32, 64 y 128, cuyos resultados completos se pueden observar en el Apéndice 5.6 de este mismo Capítulo). Cabe destacar que, de aquí en adelante, para la técnica MEMLIN, y mientras no se indique lo contrario, el número de Gaussianas empleadas para modelar el espacio limpio será el mismo que el utilizado para representar cada entorno básico. Asimismo se incluye igualmente en la Tabla, además del WER medio, MWER, la mejora media de WER, *Mean IMProvement*, MIMP, en tanto por ciento, y calculada a partir del correspondiente MWER del siguiente modo

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	IRATZ 128	3.74	8.83	6.15	7.77	9.06	7.30	8.50	7.27	61.84
CLK	SPLICE ME 128	2.96	8.06	6.29	6.14	8.77	7.46	9.18	6.75	65.39
CLK	MEMLIN 128	2.30	7.46	4.62	6.39	8.77	5.40	8.16	6.05	70.22

Cuadro 5.1: Mejores resultados con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando las distintas técnicas básicas de normalización de vectores de características. Se ha empleado la *parametrización UZ* y modelos acústicos para las unidades fonéticas. Dichos modelos acústicos se generan a partir de la señal limpia (CLK en la columna de Entrenamiento). La columna marcada como "Reconocimiento" hace referencia a la señal empleada para reconocer, que será la ruidosa normalizada con las técnicas IRATZ, SPLICE con selección de modelo de entorno, que se identifica como SPLICE ME, o MEMLIN. Junto al nombre de las diferentes técnicas aparece el número de Gaussianas con que se modelaron los correspondientes espacios. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

$$MIMP = \frac{100(MWER - MWER_{CLK-HF})}{MWER_{CLK-CLK} - MWER_{CLK-HF}}, \tag{5.29}$$

donde $MWER_{CLK-CLK}$ es el WER medio obtenido en condiciones limpias (1.75 % en este caso), y $MWER_{CLK-HF}$ es el valor de referencia, esto es, el WER medio en condiciones desajustadas (16.21 % para esta experimentación); en ambos casos, tal y como se puede apreciar, se han tomado los valores obtenidos tras aplicar el algoritmo CMS, ya que éste se considera actualmente un estándar de facto para cualquier parametrización. De este modo, y a partir de la expresión anterior, se puede observar que un 100 % de mejora supondría que el MWER conseguido sería el mismo que el obtenido en condiciones limpias, que en principio es el límite al que se debe aspirar. Por otra parte, y para completar la comparativa entre los distintos métodos presentados hasta el momento, se pueden extraer de la Tabla 4.3 los resultados al aplicar únicamente la técnica CMS (Entrenamiento CLK, Reconocimiento HF). A la luz pues de los valores presentados en las Tablas 5.1 y 4.3 se puede concluir que, teniendo en cuenta únicamente los mejores resultados para las distintas técnicas, y para todos y cada uno de los entornos, el método IRATZ proporciona mejores resultados que la técnica CMS; del mismo modo, al aplicar la técnica SPLICE con selección de modelo de entorno, se logra en media un mejor resultado (MWER de 6.75 %) que el obtenido por el algoritmo IRATZ (MWER de 7.27 %); afirmación ésta que se puede repetir si se comparan las técnicas MEMLIN (MWER de 6.05 %) y SPLICE con selección de modelo de entorno.

Por otra parte, y para determinar si se puede afirmar o no que los resultados anteriores son estadísticamente significativos, se recurre a la prueba de hipótesis estadística *z-test*. De este modo, para que dos técnicas presenten comportamientos estadísticamente diferentes independientemente de la base de datos, *SpeechDat Car* en español en este caso, con un intervalo de confianza del 95 %, el valor del estadístico W, w, debe ser mayor que 1.96. Comparando los métodos IRATZ y MEMLIN se puede observar que $w = 2{,}61 > 1{,}96$, por lo que la mejora del algoritmo en este caso sí se puede considerar independiente de la base de datos con el intervalo de confianza elegido. Por otra parte, si se comparan los resultados obtenidos por las técnicas SPLICE ME y MEMLIN, se aprecia que $w = 1{,}53 < 1{,}96$, con lo que no se puede considerar que la diferencia de comportamiento de las dos técnicas sea estadísticamente significativa con un intervalo de confianza del 95 %. De todos modos, a la hora de valorar las conclusiones obtenidas mediante la hipótesis estadística *z-test*, hay que tener en cuenta siempre las limitaciones de la propia prueba, ya comentadas en la Sección 4.2.

En la Figura 5.6 se muestra la mejora media de WER, MIMP, para las distintas técnicas cuando se realiza un barrido del número de componentes de las GMMs necesarias para cada método (8, 16, 32, 64 y 128). A la hora de comparar los distintos algoritmos, se ha representado el MIMP con respecto al número de Gaussianas por entorno básico ya que este parámetro da una idea del coste computacional del proceso de normalización, puesto que la evaluación de las correspondientes exponenciales supone la mayor cantidad de tiempo en dicho proceso. Recuérdese que el método IRATZ, a pesar de modelar el espacio limpio, transforma cada Gaussiana de dicho modelo en otra asociada al espacio ruidoso del entorno básico correspondiente. Se puede apreciar como el algoritmo SPLICE ME proporciona un mejor comportamiento medio que el método IRATZ para todos los casos debido a que el modelo de degradación aplicado en esta última técnica para estimar la probabilidad a posteriori de la Gaussiana del modelo limpio dado el vector de características ruidoso no deja de ser una aproximación que en muchas ocasiones se aleja de la realidad. Por otra parte, el método MEMLIN mejora igualmente los resultados medios obtenidos por el algoritmo SPLICE ME para cualquier número de Gaussianas por entorno básico. De este modo queda patente que el hecho de que el espacio de proyección asociado a los vectores de desplazamiento del método MEMLIN posea menos incertidumbre que los correspondientes a las otras técnicas aquí comparadas, dando

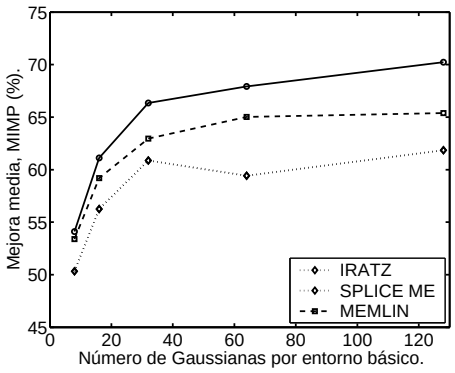


Figura 5.6: Mejora media del WER, MIMP, para las técnicas *Interpolated RATZ* (IRATZ), SPLICE con selección del modelo de entorno (SPLICE MS) y MEMLIN.

lugar por tanto a transformaciones más específicas, se manifiesta en una mejora en las tasas de reconocimiento. Cabe destacar que, aunque el número de vectores de desplazamiento asociados a cada Gaussiana del espacio ruidoso es mayor en el caso del algoritmo MEMLIN que en el del resto de técnicas, el coste computacional en la fase de normalización es casi idéntico puesto que en ella sólo se han de evaluar las Gaussianas asociadas al espacio ruidoso.

Ya para finalizar, las Figuras 5.7.a y 5.7.b presentan los histogramas comparativos y los *log-scattegrams* construidos a partir del primer coeficiente MFCC de los vectores de características de voz provinientes de las señales limpia y ruidosa (a) y limpia y normalizada (b) para el corpus de reconocimiento del entorno básico E4. La señal normalizada se obtuvo a partir del algoritmo MEMLIN con 128 Gaussianas por entorno básico. A pesar de las diferencias, tanto conceptuales como en cuanto a las tasas de RAH que se han comentado con anterioridad, los *log-scattegrams* e histogramas asociados a las técnicas IRATZ y SPLICE ME son visualmente muy similares a los obtenidos con el método MEMLIN, de ahí que no se presenten. Comparada con la Figura 5.7.a.2, la incertidumbre tras el proceso de normalización (Figura 5.7.b.2) se ha visto sensiblemente reducida, acercando además el *log-scattegram* correspondiente hacia la función identidad $x = y$, que define la normalización perfecta. Por su parte, en la Figura 5.7.b.1 queda patente como el histograma de la señal normalizada es muy similar al de la señal limpia salvo por un importante pico que aparece en torno a -5 y que se debe a la transformación de gran número de vectores de características ruidosos hacia el silencio del espacio limpio. Este problema se podría solucionar mediante del uso de un eficiente VAD, *Voice Activity Detector*, en el proceso de normalización, tanto en la fase previa de entrenamiento como en la posterior transformación del vector de características ruidoso. Para asegurar esta afirmación se normalizó la señal ruidosa mediante la técnica MEMLIN con 128 Gaussianas por entorno básico donde, en este caso, la estimación de los vectores de desplazamiento y los modelos de probabilidad de las Gaussianas (proceso de entrenamiento previo) se llevaron a cabo únicamente con los vectores de características etiquetados como de voz. La Figura 5.7.c representa el *log-scattegram* y el histograma construidos a partir del primer coeficiente MFCC de las tramas de voz de la señal limpia y normalizada con la técnica MEMLIN bajo las nuevas condiciones de entrenamiento, pudiéndose observar como el pico en el histograma ha desaparecido.

A pesar del satisfactorio comportamiento en MWER de la técnica MEMLIN, cuyos valores son menores que los obtenidos con los algoritmos CMS, IRATZ y SPLICE ME, se puede decir, a modo de conclusión, que, si se analiza dicho método en profundidad, se observan principalmente

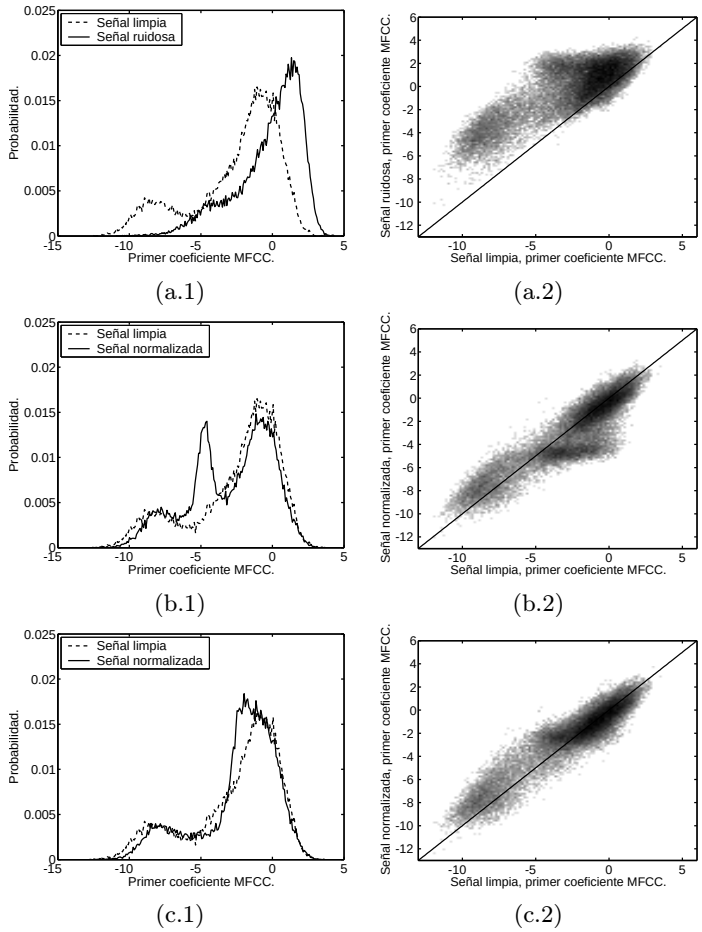


Figura 5.7: *Log-scattegrams* e histogramas realizados entre el primer coeficiente MFCC de las tramas de voz de la señal limpia (eje de abscisas) y la señal ruidosa (a) (eje de ordenadas), o normalizada usando la técnica MEMLIN con 128 Gaussianas por entorno básico (b) (eje de ordenadas). En la figura (c) se representa el *log-scattegram* y el histograma obtenidos a partir de la señal normalizada con la técnica MEMLIN con 128 Gaussianas por entorno básico modificando previamente la fase de entrenamiento, que en este caso se llevó a cabo únicamente con tramas de voz. Todas las representaciones se realizaron a partir del corpus de reconocimiento del entorno básico E4 de la base de datos *SpeechDat Car* en español. La línea en los *log-scattegrams* representa la función $x = y$.

dos aproximaciones que pueden afectar en gran medida al comportamiento final de la técnica: por una parte la elección del modelo de $\mathbf{x}$ ($\mathbf{x} \approx \mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e) = \mathbf{y}_t - \mathbf{r}_{s_x,s_y^e}$), que presupone una transformación lineal del vector de características ruidoso con pendiente unidad, esto es, se asume que el efecto del entorno acústico asociado a cada par de Gaussianas se puede compensar únicamente con un vector de desplazamiento. La segunda aproximación consiste en presuponer que el modelo de la probabilidad entre Gaussianas es independiente del vector de características ruidoso ($p(s_x|\mathbf{y}_t, e, s_y^e) \approx p(s_x|e, s_y^e)$), lo que hace que la probabilidad de una determinada Gaussiana del modelo GMM limpio dada otra del modelo GMM ruidoso del entorno básico correspondiente sea en todo momento la misma, independientemente del vector acústico, lo que no deja de ser algo irreal. La primera limitación se estudiará de un modo directo en el Capítulo 6, mientras que en el Capítulo 8 se analizará cómo compensar ciertos efectos, como las rotaciones [Molau, 2003], que el

entorno acústico puede producir y que, para su normalización, es necesario definir un modelo de $\mathbf{x}$ que considere que los coeficientes de los vectores de características no son independientes, hecho éste no tratado hasta el momento en este trabajo. Por último, y para proporcionar un modelo de la probabilidad entre Gaussianas más realista, se propondrá en el Capítulo 7 una solución basada en GMMs.

5.5. Anexo A.

Dado que la técnica MEMLIN es, de los cuatro algoritmos de normalización de vectores de características empíricos tratados en este Capítulo, el más complejo, en este Anexo se incluirá únicamente el desarrollo teórico necesario para estimar el correspondiente vector de desplazamiento, $\mathbf{r}_{s_x,s_y^e}$, a partir de la minimización del error cuadrático medio asociado a cada par de Gaussianas, ξ_{s_x,s_y^e}. La generalización de este desarrollo teórico para los algoritmos CMN, RATZ o SPLICE es directamente una simplificación del propuesto en este Anexo.

Sea pues un corpus de entrenamiento estéreo para el entorno básico e $(\mathbf{X}_e, \mathbf{Y}_e) = \{(\mathbf{x}_1^e, \mathbf{y}_1^e); ...; (\mathbf{x}_{t_e}^e, \mathbf{y}_{t_e}^e); ...; (\mathbf{x}_{T_e}^e, \mathbf{y}_{T_e}^e)\}$, con $t_e \in [1, T_e]$; nótese que, por simplificar la notación se ha eliminado el índice Tr para indicar que se trata del corpus de entrenamiento, tal y como sí estaba recogido en la Sección 5.3. De este modo, el error cuadrático medio asociado a cada par de Gaussianas para la técnica MEMLIN, ξ_{s_x,s_y^e}, se define como

$$\xi_{s_x,s_y^e} = \frac{1}{T_e}\sum_{t_e} p(s_x|\mathbf{x}_{t_e}^e, e)p(s_y^e|\mathbf{y}_{t_e}^e, e)Tra[(\mathbf{x}_{t_e}^e - \Psi(\mathbf{y}_{t_e}^e, s_x, s_y^e))(\mathbf{x}_{t_e}^e - \Psi(\mathbf{y}_{t_e}^e, s_x, s_y^e))^T], \quad (A.1)$$

donde $\Psi(\mathbf{y}_t, s_x, s_y^e) = \mathbf{y}_t - \mathbf{r}_{s_x,s_y^e}$. Teniendo en cuenta esto último, así como ciertas propiedades del cálculo matricial, se puede observar, antes de llevar a cabo la minimización de ξ_{s_x,s_y^e}, que

$$\begin{aligned}(\mathbf{x}_{t_e}^e - \Psi(\mathbf{y}_{t_e}^e, s_x, s_y^e))(\mathbf{x}_{t_e}^e - \Psi(\mathbf{y}_{t_e}^e, s_x, s_y^e))^T = \mathbf{x}_{t_e}^e(\mathbf{r}_{s_x,s_y^e})^T - \mathbf{x}_{t_e}^e(\mathbf{y}_{t_e}^e)^T + \mathbf{x}_{t_e}^e(\mathbf{y}_{t_e}^e)^T \\ -\mathbf{y}_{t_e}^e(\mathbf{r}_{s_x,s_y^e})^T + \mathbf{y}_{t_e}^e(\mathbf{y}_{t_e}^e)^T - \mathbf{y}_{t_e}^e(\mathbf{x}_{t_e}^e)^T \\ +\mathbf{r}_{s_x,s_y^e}^T(\mathbf{r}_{s_x,s_y^e})^T - \mathbf{r}_{s_x,s_y^e}(\mathbf{y}_{t_e}^e)^T + \mathbf{r}_{s_x,s_y^e}(\mathbf{x}_{t_e}^e)^T. \quad (A.2)\end{aligned}$$

A la hora de estimar el vector de desplazamiento $\mathbf{r}_{s_x,s_y^e}$ se procede a la minimización de la expresión (A.1) con respecto a $\mathbf{r}_{s_x,s_y^e}$ haciendo uso de (A.2). Para ello es necesario

$$\begin{aligned}\mathbf{0} = \frac{\delta\xi_{s_x,s_y^e}}{\delta\mathbf{r}_{s_x,s_y^e}} = \frac{1}{T_e}\sum_{t_e} p(s_x|\mathbf{x}_{t_e}^e, e)p(s_y^e|\mathbf{y}_{t_e}^e, e) \\ \frac{\delta}{\delta\mathbf{r}_{s_x,s_y^e}}[Tra[\mathbf{x}_{t_e}^e(\mathbf{r}_{s_x,s_y^e})^T - \mathbf{x}_{t_e}^e(\mathbf{y}_{t_e}^e)^T + \mathbf{x}_{t_e}^e(\mathbf{y}_{t_e}^e)^T \\ -\mathbf{y}_{t_e}^e(\mathbf{r}_{s_x,s_y^e})^T + \mathbf{y}_{t_e}^e(\mathbf{y}_{t_e}^e)^T - \mathbf{y}_{t_e}^e(\mathbf{x}_{t_e}^e)^T \\ +\mathbf{r}_{s_x,s_y^e}^T(\mathbf{r}_{s_x,s_y^e})^T - \mathbf{r}_{s_x,s_y^e}(\mathbf{y}_{t_e}^e)^T + \mathbf{r}_{s_x,s_y^e}(\mathbf{x}_{t_e}^e)^T]]. \quad (A.3)\end{aligned}$$

O, lo que es lo mismo

$$\mathbf{0} = \frac{1}{T_e}\sum_{t_e} p(s_x|\mathbf{x}_{t_e}^e, e)p(s_y^e|\mathbf{y}_{t_e}^e, e)(\mathbf{x}_{t_e}^e - \mathbf{y}_{t_e}^e + 2\mathbf{r}_{s_x,s_y^e} + \mathbf{x}_{t_e}^e - \mathbf{y}_{t_e}^e). \quad (A.4)$$

Finalmente, se obtiene la expresión óptima para $\mathbf{r}_{s_x,s_y^e}$ despejando convenientemente

$$\mathbf{r}_{s_x,s_y^e} = \frac{\sum_{t_e} p(s_x|\mathbf{x}_{t_e}^e, e)p(s_y^e|\mathbf{y}_{t_e}^e, e)(\mathbf{y}_{t_e}^e - \mathbf{x}_{t_e}^e)}{\sum_{t_e} p(s_x|\mathbf{x}_{t_e}^e, e)p(s_y^e|\mathbf{y}_{t_e}^e, e)}. \tag{A.5}$$

5.6. Anexo B.

6

Mejoras de modelado para la técnica MEMLIN.

Tal y como se ha adelantado en el Capítulo 5, uno de los puntos sobre el que se puede actuar para mejorar el comportamiento de la técnica MEMLIN es la elección del modelo de $\mathbf{x}$ ($\mathbf{x} \approx \mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e) = \mathbf{y}_t - \mathbf{r}_{s_x,s_y^e}$), que presupone para este método, tal y como se puede apreciar, una transformación lineal del vector de características ruidoso compuesta por un término de pendiente unidad y un vector de desplazamiento, $\mathbf{r}_{s_x,s_y^e}$. Este tipo de transformaciones compensa las modificaciones que el entorno acústico pudiera generar en la media de los vectores de características asociados a cada par de Gaussianas, s_x y s_y^e, pero no así las alteraciones de la correspondiente varianza. Asimismo, tal y como se han modelado los espacios ruidoso y limpio, y teniendo en cuenta la definición de los vectores de desplazamiento, todos los sonidos se tratan de la misma manera, de modo que siempre se dispone de un vector de desplazamiento capaz de proyectar desde cualquier Gaussiana del modelo del espacio ruidoso hacia todas y cada una de las del modelo del espacio limpio. Esto puede producir, si las tramas de silencio quedan representadas con un número elevado de Gaussians, que muchos vectores de características ruidosos que que se corresponden con segmentos de voz acaben siendo transformados al silencio del espacio limpio, efecto este que ya se apreció en la Sección 5.4.

Para compensar las limitaciones mostradas por el modelado de $\mathbf{x}$ considerado en la técnica MEMLIN, se propone, por un lado modificar $\mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e)$, dando lugar a un nuevo modelo de transformación más complejo y real, y por el otro lado definir transformaciones dependientes de los sonidos, lo que reducirá sensiblemente el espacio de proyección para cada una de las Gaussianas de los modelos con que se representan los diferentes entornos básicos.

Considerar un patrón más realista de $\mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e)$ pretende compensar las alteraciones que el entorno acústico produce, no sólo en la media, sino también en la varianza de los vectores de características asociados a cada par de Gaussianas, s_x y s_y^e. Por ello se propone, tanto incluir un término de pendiente que pudiera ser distinto de la unidad, lo que da lugar a la técnica P-MEMLIN, *Polynomial Multi-Environment Model-based LInear Normalization* [Buera *et al.*, 2005b], como hacer uso de una transformación no lineal para cada par de Gaussianas, lo que generaría el método MEMHIN, *Multi-Environment Model-based HIstogram Normalization* [Buera *et al.*, 2004b].

Por otra parte, la segunda alternativa propuesta para mejorar el modelo de $\mathbf{x}$ consiste en hacer uso de transformaciones dependientes de los sonidos, de tal manera que únicamente se definan éstas entre Gaussianas de los modelos limpio y ruidoso asociadas a un mismo fonema, acotando de

este modo el rango de acción de las propias transformaciones. Para eso es necesario dividir ambos espacios, limpio y ruidoso, en fonemas, representando cada uno de ellos mediante una GMM. A esta nueva técnica se la denomina PD-MEMLIN, *Phoneme Dependent Multi-Environment Model-based LIneal Normalization* [Buera *et al.*, 2005c] y pretende además, como se podrá a preciar, acercar el efecto de las transformaciones al dominio de los modelos acústicos.

En este Capítulo se presenta primeramente el algoritmo P-MEMLIN haciendo uso de la misma base teórica empleada para explicar las distintas técnicas expuestas en el Capítulo 5 (Sección 6.1). A continuación (Sección 6.2) se plantea el desarrollo teórico de la técnica MEMHIN, cerrando de esta manera la primera línea de actuación propuesta para modificar el modelo de $\mathbf{x}$. El método PD-MEMLIN se analiza convenientemente en la Sección 6.3, donde se podrá observar que, en el fondo, constituye una generalización de la técnica MEMLIN. Hasta el momento todas las técnicas de normalización de vectores de características presentadas precisan, en su fase de entrenamiento previa, señal estéreo. Para evitar esta limitación, en la Sección 6.4 se presenta una fase de entrenamiento para el algoritmo PD-MEMLIN en la que no es necesario señal estéreo. Los resultados de RAH obtenidos tras la aplicación de los distintos métodos de normalización propuestos en este Capítulo con la base de datos *SpeechDat Car* en español, se incluyen en la Sección 6.5. En ella queda patente el buen comportamiento del algoritmo PD-MEMLIN, no sólo con respecto a los métodos empíricos basados en el criterio MMSE más utilizados en la actualidad (CMN, RATZ y SPLICE), sino también si se compara con la técnica MEMLIN. Por su parte, los métodos P-MEMLIN y MEMHIN no proporcionan importantes mejoras con respecto al algoritmo MEMLIN en la experimentación con la base de datos *SpeechDat Car* en español realizada, aunque su comportamientos ante ruido aditivo sí es considerablemente más satisfactorio.

6.1. Técnica *Polynomial* MEMLIN, P-MEMLIN.

La utilización de un modelado de $\mathbf{x}$ más complejo que el tratado hasta el momento, introduciendo un término de pendiente distinto de la unidad para compensar la varianza de los vectores de características, ya se ha utilizado anteriormente para proporcionar robustez a los sistemas de RAH. Así, por ejemplo, incluir esta modificación en la técnica CMN se puede ver como aplicar conjuntamente dicho método con normalización cepstral de varianza, CVN, *Cepstral Variance Normalization* [Viikki and Laurila, 1998]. Por su parte, la técnica SPLICE posee igualmente una extensión en la que se introduce un término de pendiente en el modelado de $\mathbf{x}$ [Droppo *et al.*, 2005]. En ambos casos las mejoras, aunque no sobresalientes, sí aportan algo más de robustez al sistema final, sobre todo ante ruido aditivo, tal y como se podría esperar tras el estudio de los efectos que el ruido introduce en los vectores de características (Sección 5.1). A continuación se trata la correspondiente extensión del algoritmo MEMLIN en la que se considera un modelado de $\mathbf{x}$ compuesto por un polinomio de orden uno del vector de características ruidoso, permitiéndose que el término de pendiente sea distinto de la unidad. A dicha técnica se la denomina *Polynomial Multi-Environment Model-based LInear Normalization*, P-MEMLIN, [Buera *et al.*, 2005b].

Tal y como se ha adelantado, la principal diferencia de la técnica P-MEMLIN, con respecto al algoritmo MEMLIN reside en el modelado de $\mathbf{x}$, que en este caso se asume que es lineal con respecto al vector de características ruidoso, pero posibilitando que el término de la pendiente sea distinto de la unidad. Lo que se pretende con ello es modificar la pdf de la señal ruidosa asociada a cada par de Gaussianas, s_x y s_y^e, acercándola no sólo en media, como es el caso de la técnica MEMLIN, sino también en términos de varianza a la correspondiente pdf de la señal limpia. De este modo, la nueva expresión propuesta para $\mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e)$ para el método P-MEMLIN será

$$\mathbf{x} \approx \mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e) = \mathbf{A}_{s_x,s_y^e}\mathbf{y}_t - \mathbf{b}_{s_x,s_y^e}, \tag{6.1}$$

donde $\mathbf{A}_{s_x,s_y^e}$ y $\mathbf{b}_{s_x,s_ye}$ son la matriz diagonal asociada al término de pendiente y el vector que representa el término independiente del nuevo modelo de $\mathbf{x}$, respectivamente, siendo ambos función de los distintos pares de Gaussianas de los modelos limpio y ruidoso, s_x y s_y^e, *respectivamente*. Nótese que en la definición de $\mathbf{A}_{s_x,s_y^e}$ se encuentra implícita la consideración de que los distintos coeficientes de los vectores de caracteísticas son independientes, hecho este no del todo cierto a pesar de la DCT, *Discrete Cosine Transform*, empleada en las dos parametrizaciones consideradas en este trabajo (*parametrización UZ* y parametrización estándar ETSI). Una vez asumida la expresión (6.1), las otras dos aproximaciones sobre el modelado de los espacios limpio y ruidoso consideradas para la técnica MEMLIN y expuestas en la Sección 5.3 siguen siendo igualmente válidas (expresiones (5.19), (5.20), (5.7) y (5.8)). A partir de todo lo anterior, y haciendo uso de las tres suposiciones en las que se basa la técnica P-MEMLIN, la expresión (5.3) se transforma en este caso en

$$\begin{aligned}\hat{\mathbf{x}}_t &= \int_{\mathbf{X}} \sum_e \sum_{s_y^e} \sum_{s_x} \mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e) p(\mathbf{x}, s_x, e, s_y^e|\mathbf{y}_t) d\mathbf{x} \\ &= \sum_e \sum_{s_y^e} \sum_{s_x} (\mathbf{A}_{s_x,s_y^e}\mathbf{y}_t - \mathbf{b}_{s_x,s_y^e}) p(e|\mathbf{y}_t) p(s_y^e|\mathbf{y}_t, e) p(s_x|\mathbf{y}_t, e, s_y^e),\end{aligned} \tag{6.2}$$

El cómputo de la probabilidad a posteriori del entorno básico, $p(e|\mathbf{y}_t)$, la probabilidad a posteriori de la Gaussiana del modelo ruidoso dado el vector de características degradado y el entorno básico, $p(s_y^e|\mathbf{y}_t, e)$, y el modelo de la probabilidad entre Gaussianas $p(s_x|\mathbf{y}_t, e, s_y^e)$, se pueden obtener del mismo modo que para el método MEMLIN empleando las expresiones (5.22), (5.23) y (5.27) o (5.28), según la decisión seleccionada para el modelo de probabilidad entre Gaussianas, *hard* o *soft*, respectivamente. Por su parte, los parámetros que definen el nuevo modelado de $\mathbf{x}$, esto es $\mathbf{A}_{s_x,s_y^e}$ y $\mathbf{b}_{s_x,s_y^e}$, se estiman mediante señal estéreo para cada entorno básico en un proceso de entrenamiento previo, $(\mathbf{X}_e^{Tr}, \mathbf{Y}_e^{Tr}) = \{(\mathbf{x}_1^{Tr,e}, \mathbf{y}_1^{Tr,e}); ...; (\mathbf{x}_{t_e}^{Tr,e}, \mathbf{y}_{t_e}^{Tr,e}); ...; (\mathbf{x}_{T_e}^{Tr,e}, \mathbf{y}_{T_e}^{Tr,e})\}$, con $t_e \in [1, T_e]$. Dado que, tal y como se ha comentado, el objetivo último de la técnica P-MEMLIN es modificar la pdf de la señal ruidosa asociada a cada par de Gaussianas, s_x y s_y^e, acercándola a la correspondiente pdf de la señal limpia, el criterio que se seguirá en esta ocasión para estimar $\mathbf{A}_{s_x,s_y^e}$ y $\mathbf{b}_{s_x,s_y^e}$ consiste en que tanto la media como la desviación típica de las pdfs asociadas a s_x y s_y^e de la señal limpia y de la obtenida mediante el modelo de $\mathbf{x}$ coincidan. Cabe destacar que dicho criterio coincide con el MMSE cuando el modelo lineal de $\mathbf{x}$ consta únicamente de un vector de desplazamiento, caso de la técnica MEMLIN, por ejemplo. Con todo esto, se puede observar que las correspondientes expresiones óptimas para las variables $\mathbf{A}_{s_x,s_y^e}$ y $\mathbf{b}_{s_x,s_y^e}$ para el algoritmo P-MEMLIN son

$$\mathbf{A}_{s_x,s_y^e} = \sqrt{\mathbf{\Sigma}^x_{s_x,s_y^e}}(\sqrt{\mathbf{\Sigma}^y_{s_x,s_y^e}})^{-1}, \tag{6.3}$$

$$\mathbf{b}_{s_x,s_y^e} = \sqrt{\mathbf{\Sigma}^x_{s_x,s_y^e}}(\sqrt{\mathbf{\Sigma}^y_{s_x,s_y^e}})^{-1}\mu^y_{s_x,s_y^e} - \mu^x_{s_x,s_y^e}, \tag{6.4}$$

donde el operador $\sqrt{\bullet}$ realiza la raíz cuadrada elemento a elemento de la matriz $\bullet$; por su parte, $\mathbf{\Sigma}^x_{s_x,s_y^e}$ y $\mathbf{\Sigma}^y_{s_x,s_y^e}$ son las matrices diagonales de las covarianzas de los vectores de características limpios y ruidosos, respectivamente, asociados al par de Gaussianas s_x y s_y^e. Asimismo, $\mu^x_{s_x,s_y^e}$ y $\mu^y_{s_x,s_y^e}$ son las medias de los vectores de características limpios y ruidosos, respectivamente,

asociados igualmente al par de Gaussianas s_x y s_y^e. Estas cuatro últimas variables se calculan matemáticamente del siguiente modo, donde z puede ser x o y

$$\mu^z_{s_x,s_y^e} = \frac{\sum_{t_e} p(s_x|\mathbf{x}_{t_e})p(s_y^e|\mathbf{y}_{t_e})\mathbf{z}_{t_e}}{\sum_{t_e} p(s_x|\mathbf{x}_{t_e})p(s_y^e|\mathbf{y}_{t_e})}, \tag{6.5}$$

$$\mathbf{\Sigma}^z_{s_x,s_y^e} = diag[\frac{\sum_{t_e} p(s_x|\mathbf{x}_{t_e})p(s_y^e|\mathbf{y}_{t_e})(\mathbf{z}_{t_e} - \mu^z_{s_x,s_y^e})(\mathbf{z}_{t_e} - \mu^z_{s_x,s_y^e})^T}{\sum_{t_e} p(s_x|\mathbf{x}_{t_e})p(s_y^e|\mathbf{y}_{t_e})}], \tag{6.6}$$

donde el operador $diag[\bullet]$ hace nulos todos los elementos de la matriz $\bullet$ distintos de la diagonal. Cabe destacar que el desarrollo teórico completo para obtener las expresiones de $\mathbf{A}_{s_x,s_y^e}$ y $\mathbf{b}_{s_x,s_y^e}$ se puede consultar en el Anexo 5.6 de este mismo Capítulo. Nótese asimismo que, si se asume que el entorno acústico no modifica la varianza de los vectores de características limpios asociados a cada par de Gaussianas s_x y s_y^e, esto es, que los términos de desviación estándar son idénticos para todos los coeficientes y pares de Gaussianas ($\sqrt{\mathbf{\Sigma}^x_{s_x,s_y^e}} = \sqrt{\mathbf{\Sigma}^y_{s_x,s_y^e}} \forall s_x, s_y^e$), las matrices $\mathbf{A}_{s_x,s_y^e}$ se corresponderían con la identidad y las expresiones para el vector $\mathbf{b}_{s_x,s_y^e}$ coincidirían con las del vector de desplazamiento del algoritmo MEMLIN, $\mathbf{r}_{s_x,s_y^e}$, de modo que, en dicho caso, las técnicas P-MEMLIN y MEMLIN proporcionarían exactamente los mismos resultados.

6.2. Técnica *Multi-Environment Model-based HIstogram Normalization*, MEMHIN.

A pesar de que la técnica P-MEMLIN, al emplear como transformación de $\mathbf{x}$ un polinomio de orden uno en el que el término independiente puede ser distinto de la unidad, ya asume que el entorno acústico puede modificar tanto la media como la varianza de las pdfs de los vectores de características limpios asociados a cada par de Gaussianas, s_x y s_y^e, en algunas ocasiones puede ser necesario considerar un modelo de transformación más completo que pueda compensar órdenes estadísticos mayores, igualando así en última instancia las formas de las pdfs de los vectores de características limpios y normalizados para cada par de Gaussianas, s_x y s_y^e. Con esta intención surge la técnica MEMHIN, *Multi-Environment Model-based HIstogram Normalization*, [Buera *et al.*, 2004b] en la que se propone una nueva transformación de $\mathbf{x}$ que, manteniendo la dependencia con cada par de Gaussianas, esté basada en ecualización de histograma, *histogram equalization*.

La ecualización de histograma, tomando como patrón métodos inicialmente utilizados en realce de imagen [González and Wintz, 1987] y tal y como se ha adelantado en el Capítulo 3, ya se ha aplicado con anterioridad en sistemas de RAH para proporcionar robustez. De este modo, considera una función de transformación no lineal monótona creciente para normalizar los vectores de características, suponiendo además que los coeficientes de los mismos son independientes. Dicha transformación tiene como objetivo que la pdf de las tramas normalizadas se aproxime a una considerada de referencia [Molau, 2003] [de la Torre *et al.*, 2005]. A diferencia de esta realización básica de la técnica de ecualización de histograma, el método MEMHIN introduce una ecualización de histograma independiente para cada par de Gaussianas, s_x y s_y^e, aunque se siguen manteniendo, eso sí, las dos grandes aproximaciones ya comentadas: considerar que el efecto del entorno acústico se puede modelar mediante una función no lineal monótona creciente, lo que, debido a la incertidumbre que introduce el entorno acústico, no se ajusta con exactitud a la realidad, tal y como se puede apreciar en la Figura 5.1; y la segunda aproximación es considerar que los coeficientes de los vectores de características son independientes, cosa tampoco del todo cierta a pesar de la DCT incluida en las distintas parametrizaciones consideradas en este trabajo.

Así, el elemento diferenciador de la técnica MEMHIN con respecto al algoritmo MEMLIN reside en el nuevo modelo propuesto para $\mathbf{x}$, de manera que en esta ocasión se deberá aprender una transformación no lineal asociada a cada par de Gaussianas, s_x y s_y^e, que transforme la pdf de los vectores de características ruidosos asociados a dicho par de Gaussianas, acercándola a la pdf de las tramas limpias correspondiente igualmente a s_x y s_y^e. De este modo, el nuevo modelo de transformación para MEMHIN que cumple el criterio anteriormente comentado será

$$\mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e) = \mathbf{C}^{-1}_{x,s_x,s_y^e}(\mathbf{C}_{y,s_x,s_y^e}(\mathbf{y}_t)), \tag{6.7}$$

donde $\mathbf{C}_{x,s_x,s_y^e}$ es el histograma acumulativo del vector de características limpio asociado al par de Gaussianas s_x y s_y^e, $\mathbf{C}^{-1}_{x,s_x,s_y^e}$ es la correspondiente función recíproca y $\mathbf{C}_{y,s_x,s_y^e}$ es el histograma acumulativo del vector de características ruidoso asociado a s_x y s_y^e. Por otra parte, las dos aproximaciones para el modelado de los espacios limpio y ruidosos consideradas para las técnicas MEMLIN y P-MEMLIN y expuestas en la Sección 5.3 siguen siendo válidas (expresiones (5.19), (5.20), (5.7) y (5.8)). A partir de todo lo anterior, y haciendo uso de las tres suposiciones en las que se basa la técnica MEMHIN, la expresión (5.3) se transforma en este caso en

$$\begin{aligned}\hat{\mathbf{x}}_t &= \int_{\mathbf{X}} \sum_e \sum_{s_y^e} \sum_{s_x} \mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e) p(\mathbf{x}, s_x, e, s_y^e|\mathbf{y}_t) d\mathbf{x} \\ &= \sum_e \sum_{s_y^e} \sum_{s_x} (\mathbf{C}^{-1}_{x,s_x,s_y^e}(\mathbf{C}_{y,s_x,s_y^e}(\mathbf{y}_t))) p(e|\mathbf{y}_t) p(s_y^e|\mathbf{y}_t, e) p(s_x|\mathbf{y}_t, e, s_y^e),\end{aligned} \tag{6.8}$$

La probabilidad a posteriori del entorno básico, $p(e|\mathbf{y}_t)$, la probabilidad a posteriori de la Gaussiana del modelo ruidoso dado el vector de características degradado y el entorno básico, $p(s_y^e|\mathbf{y}_t, e)$, y el modelo de la probabilidad entre Gaussianas $p(s_x|\mathbf{y}_t, e, s_y^e)$, se pueden obtener del mismo modo que para el método MEMLIN haciendo uso de las expresiones (5.22), (5.23) y (5.27), o (5.28), según si se emplea la decisión *hard* o *soft* para el modelado de la probabilidad entre Gaussianas, respectivamente. Por otra parte, $\mathbf{C}_{x,s_x,s_y^e}$ y $\mathbf{C}_{y,s_x,s_y^e}$ se estiman haciendo uso de señal estéreo para cada entorno básico en un proceso de entrenamiento previo, $(\mathbf{X}_e^{Tr}, \mathbf{Y}_e^{Tr}) = \{(\mathbf{x}_1^{Tr,e}, \mathbf{y}_1^{Tr,e}); ...; (\mathbf{x}_{t_e}^{Tr,e}, \mathbf{y}_{t_e}^{Tr,e}); ...; (\mathbf{x}_{T_e}^{Tr,e}, \mathbf{y}_{T_e}^{Tr,e})\}$, con $t_e \in [1, T_e]$. Dado que se considera que no hay dependencia alguna entre las componentes de los vectores de características, la estimación de las funciones $\mathbf{C}_{x,s_x,s_y^e}$ y $\mathbf{C}_{y,s_x,s_y^e}$ se puede realizar coeficiente a coeficiente. Para ello se calculan primeramente los histogramas de n bandas de cada componente de los vectores de características de entrenamiento limpios y ruidosos asociados a s_x y s_y^e, lo que se realiza ponderando cada uno de ellos por el producto de las probabilidades a posteriori $p(s_x|\mathbf{x}_{t_e}^{Tr,e}, e)$ (5.26) y $p(s_y^e|\mathbf{y}_{t_e}^{Tr,e}, e)$ (5.23). Una vez obtenidos los histogramas, $\mathbf{C}_{x,s_x,s_y^e}$ y $\mathbf{C}_{y,s_x,s_y^e}$ se calculan mediante la suma acumulada de las distintas bandas de los correspondientes histogramas. El número de bandas, n, determina la flexibilidad de la transformación, siendo necesario un valor lo suficientemente elevado como para poder corregir las no linealidades que el entorno acústico introduce, pero teniendo en cuenta que un incremento excesivo en el número de bandas repercute sensiblemente en el coste computacional.

6.3. Técnica *Phoneme Dependent* MEMLIN, PD-MEMLIN.

Ya se ha podido apreciar anteriormente en el Capítulo 5 que el hecho de que los espacios limpio y ruidosos se modelen mediante GMMs, unido a que se definan vectores de desplazamiento para todos los pares posibles de Gaussianas, s_x y s_y^e, puede producir que, por ejemplo, innumerables vectores de características de voz ruidosos se proyecten hacia el silencio del espacio limpio, especialmente si éste queda modelado con un número elevado de Gaussianas. Así sucede en las técnicas

MEMLIN, P-MEMLIN y MEMHIN y para solventarlo, obteneniendo una serie de transformaciones más específicas a la vez que se trata de reducir el desajuste entre los vectores de características normalizados y los modelos acústicos que se van a emplear en decodificación, se propone entrenar transformaciones de forma independiente para cada fonema. Con este objetivo nace la técnica PD-MEMLIN, *Phoneme-Dependent Multi-Environment Model-based LInear Normalization.*

El uso de técnicas de normalización empíricas dependientes del fonema no es del todo novedoso, ya que anteriormente se han empleado algoritmos como *Phone-Dependent Cepstral Normalization*, PDCN, [Liu *et al.*, 1994] cuya filosofía es similar a la planteada en el método de normalización igualmente empírico denominado *Fixed Codeword-Dependent Cepstral Normalization*, FCDCN, [Acero and Stern, 1990] y cuyas transformaciones dependen de una serie de *codebooks* entrenados para distintos fonemas y SNRs. Asimismo, cabe destacar que la técnica PDCN, que en concepto también es similar al trabajo presentado en [Beattie, 1992], necesita, a la hora de normalizar los distintos vectores de características ruidosos, una hipótesis del fonema al que pertenece cada uno de ellos; estimación esta que se lleva a cabo tras un reconocimiento previo realizado mediante el correspondiente sistema de RAH.

La técnica PD-MEMLIN, por su parte, considera alguna de las premisas ya presentadas para el algoritmo MEMLIN, considerando nuevamente tres aproximaciones, a saber

- El espacio ruidoso se divide en una serie de entornos básicos, e, cada uno de los cuales se encuentra compuesto por un conjunto de fonemas, ph. Con esto, los vectores de características ruidosos asociados a cada fonema y entorno básico se modelan mediante una GMM

$$p_{e,ph}(\mathbf{y}_t) = \sum_{s_y^{e,ph}} p(\mathbf{y}_t|s_y^{e,ph})p(s_y^{e,ph}), \tag{6.9}$$

$$p(\mathbf{y}_t|s_y^{e,ph}) = \mathcal{N}(\mathbf{y}_t; \mu_{s_y^{e,ph}}, \mathbf{\Sigma}_{s_y^{e,ph}}), \tag{6.10}$$

  donde $s_y^{e,ph}$ se corresponde con la Gaussiana asociada al fonema ph y el entorno básico e, mientras que $\mu_{s_y^{e,ph}}$, $\mathbf{\Sigma}_{s_y^{e,ph}}$, y $p(s_y^{e,ph})$ son el vector de media, la matriz de covarianza diagonal y la probabilidad a priori asociados a $s_y^{e,ph}$.

- Asimismo, los vectores de características pertenecientes al espacio limpio y asociadas a cada fonema se modelan iguamente mediante una GMM

$$p_{ph}(\mathbf{x}) = \sum_{s_x^{ph}} p(\mathbf{x}|s_x^{ph})p(s_x^{ph}), \tag{6.11}$$

$$p(\mathbf{x}|s_x^{ph}) = \mathcal{N}(\mathbf{x}; \mu_{s_x^{ph}}, \mathbf{\Sigma}_{s_x^{ph}}), \tag{6.12}$$

  donde s_x^{ph} hace referencia a la Gaussiana correspondiente al modelo del espacio limpio del fonema ph, y $\mu_{s_x^{ph}}$, $\mathbf{\Sigma}_{s_x^{ph}}$, y $p(s_x^{ph})$ son el vector de media, la matriz de covarianza diagonal y la probabilidad a priori asociados a s_x^{ph}.

- Finalmente, el vector de características limpio se puede aproximar mediante una función lineal del ruidoso, siendo además ésta dependiente del entorno básico y de las Gaussianas asociadas a los distintos fonemas tanto para el espacio limpio como para el degradado. Esto, de un modo matemático, puede expresarse del siguiente modo: $x \approx \mathbf{\Psi}(\mathbf{y}_t, s_x^{ph}, s_y^{e,ph}) = \mathbf{y}_t - \mathbf{r}_{s_x^{ph},s_y^{e,ph}}$, donde $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$ es el vector de desplazamiento entre las tramas limpias y ruidosas asociado a cada par de Gaussianas del mismo fonema, s_x^{ph} y $s_y^{e,ph}$.

A partir de las tres aproximaciones anteriores, la expresión (5.3) se transforma para la técnica PD-MEMLIN de la siguiente manera

$$\hat{\mathbf{x}}_t = \int_{\mathbf{X}} \sum_e \sum_{ph} \sum_{s_y^{e,ph}} \sum_{s_x^{ph}} \mathbf{\Psi}(\mathbf{y}_t, s_x^{ph}, s_y^{e,ph}) p(\mathbf{x}, s_x^{ph}, e, s_y^{e,ph}, ph|\mathbf{y}_t) d\mathbf{x}$$

$$= \mathbf{y}_t - \sum_e \sum_{ph} \sum_{s_y^{e,ph}} \sum_{s_x^{ph}} \mathbf{r}_{s_x^{ph}, s_y^{e,ph}} p(e|\mathbf{y}_t) p(ph|\mathbf{y}_t, e) p(s_y^{e,ph}|\mathbf{y}_t, e, ph) p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^e), \quad (6.13)$$

donde $p(e|\mathbf{y}_t)$ es la probabilidad a posteriori del entorno básico; $p(ph|\mathbf{y}_t, e)$ es la probabilidad a posteriori del fonema ph, dado el vector de características ruidoso $\mathbf{y}_t$ y el entorno básico e; $p(s_y^{e,ph}|\mathbf{y}_t, e, ph)$ es la probabilidad a posteriori de la Gaussiana dependiente del fonema del modelo degradado $s_y^{e,ph}$, dado el vector de características ruidoso $\mathbf{y}_t$, el entorno básico e y el fonema ph. Finalmente, $p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph})$ es el modelo de probabilidad entre Gaussianas, esto es, la probabilidad a posteriori de la Gaussiana del modelo limpio asociada al fonema ph, s_x^{ph}, dado el vector de características degradado, el entorno básico e, el fonema ph y la Gaussiana del modelo del espacio ruidoso $s_y^{e,ph}$. Este último término, unido al vector de desplazamiento, $\mathbf{r}_{s_x^{ph}, s_y^{e,ph}}$, se estiman haciendo uso de señal estéreo en un proceso de entrenamiento previo, mientras que el resto de términos presentados se obtienen durante el proceso de normalización haciendo uso de las expresiones (6.9) y (6.10).

La probabilidad a posteriori del entorno básico, $p(e|\mathbf{y}_t)$, se calcula, de la misma manera que en el método MEMLIN, iterativamente aplicando en este caso las expresiones (6.9) y (6.10)

$$p(e|\mathbf{y}_t) = \beta \cdot p(e|\mathbf{y}_{t-1}) + (1-\beta) \frac{\sum_{ph} p_{e,ph}(\mathbf{y}_t)}{\sum_e \sum_{ph} p_{e,ph}(\mathbf{y}_t)}, \quad (6.14)$$

donde se recuerda que β es la constante de memoria y que durante este trabajo se mantendrá fija, adquiriendo el valor de 0.98. Por otra parte, $p(e|\mathbf{y}_0)$ se considera uniforme para todos los entornos básicos.

La probabilidad a posteriori del fonema ph, dado el vector de características ruidoso, $\mathbf{y}_t$, y el entorno básico e, esto es $p(ph|\mathbf{y}_t, e)$, se puede estimar haciendo uso de las expresiones (6.9) y (6.10)

$$p(ph|\mathbf{y}_t, e) = \frac{p_{e,ph}(\mathbf{y}_t)}{\sum_{ph} p_{e,ph}(\mathbf{y}_t)}. \quad (6.15)$$

Ya para acabar con los términos que se han de obtener en la fase de normalización, la probabilidad a posteriori de la Gaussiana dependiente del fonema ph del modelo degradado asociado al entorno básico e, $s_y^{e,ph}$, dado el vector de características ruidoso $\mathbf{y}_t$, el entorno básico e y el fonema ph, esto es $p(s_y^{e,ph}|\mathbf{y}_t, e, ph)$, se calcula empleando (6.9) y (6.10) del siguiente modo

$$p(s_y^{e,ph}|\mathbf{y}_t, e, ph) = \frac{p(\mathbf{y}_t|s_y^{e,ph}) p(s_y^{e,ph})}{\sum_{s_y^{e,ph}} p(\mathbf{y}_t|s_y^{e,ph}) p(s_y^{e,ph})}. \quad (6.16)$$

Tal y como se ha comentado anteriormente, hay dos variables que se deben estimar en el proceso de entrenamiento previo con un corpus de señal estéreo, el vector de desplazamiento y el modelo de probabilidad entre Gaussianas. En este caso el corpus de entrenamiento será independiente para cada entorno básico y fonema: $(\mathbf{X}_{e,ph}^{Tr}, \mathbf{Y}_{e,ph}^{Tr}) = \{(\mathbf{x}_1^{Tr,e,ph}, \mathbf{y}_1^{Tr,e,ph}); ...; (\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}, \mathbf{y}_{t_{e,ph}}^{Tr,e,ph}); ...; (\mathbf{x}_{T_{e,ph}}^{Tr,e,ph}, \mathbf{y}_{T_{e,ph}}^{Tr,e,ph})\}$, con $t_{e,ph} \in [1, T_{e,ph}]$. Cabe destacar que la asignación de cada par de vectores de características

de un determinado entorno básico, e, a un fonema concreto, ph, se realiza a partir de un proceso de segmentación forzada en términos de fonemas sobre la señal limpia mediante el algoritmo de Viterbi. Así pues, a la hora de estimar el vector de desplazamiento $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$, se sigue un proceso similar al realizado para la técnica MEMLIN, obteniendo aquella expresión que minimiza el error cuadrático medio dependiente del par de Gaussianas s_x^{ph} y $s_y^{e,ph}$, y que se define como $\xi_{s_x^{ph},s_y^{e,ph}}$ del siguiente modo

$$\xi_{s_x^{ph},s_y^{e,ph}} = \frac{1}{T_{e,ph}} \sum_{t_{e,ph}} p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}, e, ph) p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, e, ph)$$
$$Tra[(\mathbf{x}_{t_{e,ph}}^{e,ph} - \mathbf{\Psi}(\mathbf{y}_t, s_x^{ph}, s_y^{e,ph}))(\mathbf{x}_{t_{e,ph}}^{Tr,e,ph} - \mathbf{\Psi}(\mathbf{y}_t, s_x^{ph}, s_y^{Tr,e,ph}))^T], \tag{6.17}$$

$$\mathbf{r}_{s_x^{ph},s_y^{e,ph}} = \underset{\mathbf{r}_{s_x^{ph},s_y^{e,ph}}}{arg\,min}(\xi_{s_x^{ph},s_y^{e,ph}}) = \frac{\sum_{t_{e,ph}} p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}, e, ph) p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, e, ph)(\mathbf{y}_{t_{e,ph}}^{Tr,e,ph} - \mathbf{x}_{t_{e,ph}}^{Tr,e,ph})}{\sum_{t_{e,ph}} p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}, e, ph) p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, e, ph)}, \tag{6.18}$$

donde $p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}, e, ph)$ es la probabilidad a posteriori de la Gaussiana dependiente del fonema ph del modelo limpio, s_x^{ph}, dado el vector de características limpio de corpus de entrenamiento $\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}$, el entorno básico e y el fonema ph. Dicho término se puede calcular a partir de las expresiones (6.11) y (6.12) como (6.19). Por otra parte, el desarrollo teórico completo para obtener (6.18) a partir de (6.17), que no deja de ser una extensión directa del presentado en el Anexo 5.5 para la técnica MEMLIN, se puede consultar en el Anexo 6.6 de este mismo Capítulo.

$$p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}, e, ph) = \frac{p(\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}|s_x^{ph}) p(s_x^{ph})}{\sum_{s_x^{ph}} p(\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}|s_x^{ph}) p(s_x^{ph})}. \tag{6.19}$$

El modelo de probabilidad entre Gaussianas, del mismo modo que ya se comentó para la técnica MEMLIN, se puede simplificar eliminando la dependencia con respecto al vector de características ruidoso. Teniendo en cuenta esta aproximación, $p(s_x^{ph}|\mathbf{y}_t, e, s_y^{e,ph}, ph) \simeq p(s_x^{ph}|e, s_y^{e,ph}, ph)$, y al igual que para el método MEMLIN, el modelo de probabilidad entre Gaussianas se puede estimar de dos maneras: mediante frecuencia relativa, solución *hard*, cuya expresión es

$$p(s_x^{ph}|\mathbf{y}_t, e, s_y^{e,ph}, ph) \simeq p(s_x^{ph}|s_y^{e,ph}, e, ph) = \frac{C_N(s_x^{ph}|s_y^{e,ph})}{N_{s_y^{e,ph}}}, \tag{6.20}$$

donde $C_N(s_x^{ph}|s_y^{e,ph})$ es el número de veces que el par de Gaussianas más probable para el corpus estéreo de entrenamiento es s_x^{ph} y $s_y^{e,ph}$ para el entorno básico e y el fonema ph. A su vez, $N_{s_y^{e,ph}}$ es el número de veces que la Gaussiana más probable para los vectores de características ruidosos del corpus estéreo de entrenamiento asociado al entorno básico e y el fonema ph es $s_y^{e,ph}$.

La segunda opción posible para estimar el modelo de probabilidad entre Gaussianas, solución *soft*, precisa de las expresiones (6.9), (6.10), (6.11) y (6.12) y se calcula de la siguiente manera

$$p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph}) \simeq p(s_x^{ph}|s_y^{e,ph}, e) = \frac{\sum_{t_{e,ph}} p(\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}|s_x^{ph}) p(\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}|s_y^{e,ph}) p(s_x^{ph}) p(s_y^{e,ph})}{\sum_{t_{e,ph}} \sum_{s_x^{ph}} p(\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}|s_x^{ph}) p(\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}|s_y^{e,ph}) p(s_x^{ph}) p(s_y^{e,ph})}. \tag{6.21}$$

A la hora de estimar el modelo de probabilidad entre Gaussianas, es posible que no haya suficientes datos en el corpus de entrenamiento como para que la solución *hard* proporcione un modelo suficientemente representativo para algunos fonemas, especialmente cuando éstos se modelan con un número elevado de Gaussianas. Por ello en los distintos experimentos llevados a cabo con la técnica PD-MEMLIN en este trabajo se hará siempre uso de la solución *soft*. Por otra parte, y a modo de resumen, se incluye una representación gráfica del método PD-MEMLIN, Figura 6.1. Obsérvese el rango de acción de los vectores de desplazamiento en este caso es bastante distinto del de la técnica MEMLIN, ya que en el primer caso no se permite la proyección desde una determinada Gaussiana del modelo ruidoso a cualquier otra que forme parte de la representación del espacio limpio. De este modo se pretende reducir el impacto que una descompensación en el corpus de entrenamiento puede generar, ya que de esta manera cada unidad fonética se puede forzar a que esté represetado con el mismo número de Gaussianas, hecho que en otras técnicas como MEMLIN, P-MEMLIN o MEMHIN no se podía asegurar, y de hecho no sucedía puesto que el silencio solía representarse con un mayor número de componentes que cualquier otra unidad fonética. Por otra parte, con este nuevo modelado de los espacios se pretende reducir el desajuste entre la señal normalizada y los modelos acústicos que posteriormente se emplearán en decodificación.

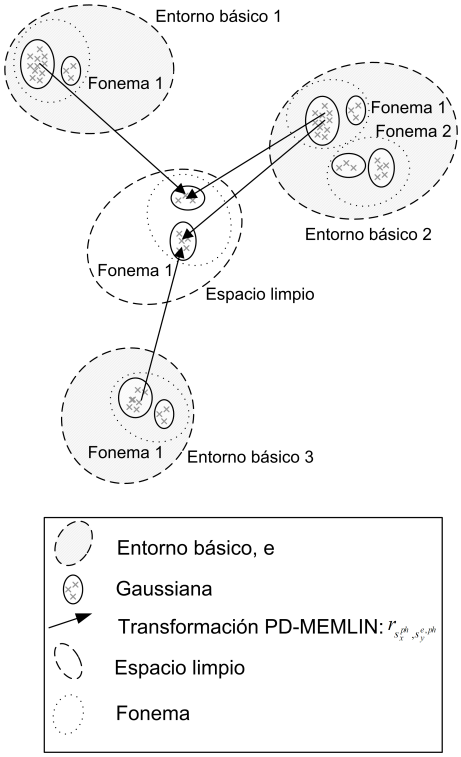


Figura 6.1: Representación gráfica de la técnica PD-MEMLIN, donde $\mathbf{r}_{s_x^{ph},s_y^{ph}}$ es el vector de desplazamiento asociado al par de Gaussianas dependientes del fonema ph de los modelos limpio y del entorno básico ruidoso e: s_x^{ph} y $s_y^{e,ph}$, respectivamente.

6.4. Técnica PD-MEMLIN con Fase de Entrenamiento "Ciega".

En muchas ocasiones, no es posible disponer de señal estéreo para llevar a cabo la fase de entrenamiento previa que se ha visto que es necesaria para los diversos algoritmos de normalización de vectores de características empíricos tratados hasta el momento. En dichos casos es preciso desarrollar un procedimiento para obtener las distintas variables necesarias con un corpus de entrenamiento compuesto únicamente por señal ruidosa, dando lugar a las técnicas que se suelen denominr "ciegas". Así se definió por ejemplo la versión "ciega" del método RATZ [Moreno, 1996]; mientras que la correspondiente al algoritmo SPLICE no se ha desarrollado hasta la fecha.

Dado que la técnica MEMLIN se puede ver como una versión simplificada del algoritmo PD-MEMLIN, a continuación se presenta el procedimiento de entrenamiento "ciego" desarrollado para este último método cuando se dispone únicamente de un corpus de entrenamiento compuesto por señal ruidosa. La obtención de las correspondientes expresiones para la técnica MEMLIN es inmediata considerando que los espacios limpio y ruidoso constan de un único fonema.

Se asume pues que se dispone de un corpus de entrenamiento compuesto por vectores de características ruidosos para cada entorno básico e y fonema ph, $(\mathbf{Y}^{Tr}_{e,ph}) = \{(\mathbf{y}_1^{Tr,e,ph}; ...; \mathbf{y}_{t_{e,ph}}^{Tr,e,ph}; ...; \mathbf{y}_{T_{e,ph}}^{Tr,e,ph}\}$, con $t_{e,ph} \in [1, T_{e,ph}]$. En este caso el fonema asociado a cada vector acústico se obtiene mediante segmentación forzada en términos de fonema de la señal ruidosa a través del algoritmo de Viterbi. Asimismo se dispone de las GMMs con las que se modelan los fonemas para los distintos entornos básicos ruidosos, y para el espacio limpio (expresiones (6.9), (6.10), (6.11) y (6.12)). De este modo, las únicas variables que se han de estimar son el vector de desplazamiento, $\mathbf{r}_{s_x^{ph}, s_y^{e,ph}}$, y el modelo de probabilidad entre Gaussianas, que se sigue considerando independiente del vector de características ruidoso, $p(s_x^{ph}|\mathbf{y}_t, s_y^{e,ph}, e, ph) \simeq p(s_x^{ph}|s_y^{e,ph}, e, ph)$. Para ello se propone un procedimiento de entrenamiento iterativo que consta de dos fases: una primera de inicialización y otra posterior de ajuste.

Durante el proceso de inicialización se calculan en primera aproximación las dos variables anteriormente comentadas, obteniéndose así $p_0(s_x^{ph}|s_y^{e,ph}, e, ph)$ y $\mathbf{r}_{0, s_x^{ph}, s_y^{e,ph}}$. La expresión para $p_0(s_x^{ph}|s_y^{e,ph}, e, ph)$ se calcula a partir de la distancia de Kullback-Liebler [Kullback and Leibler, 1951] modificada a tal efecto, y que proporciona una medida de similitud entre las Gaussianas s_x^{ph} y $s_y^{e,ph}$. Dado que se pretende cuantificar cuan parecidas son s_x^{ph} y $s_y^{e,ph}$ sin tener en cuenta el efecto que el ruido haya podido tener sobre ellas, en el cálculo de la distancia de Kullback-Liebler modificada no se tendrán en cuenta los vectores de medias, puesto que se supone que es sobre ellos donde más afecta el ruido que pudiera darse en los distintos entornos básicos. Así, la distancia de Kullback-Liebler modificada, $d_{KL}(s_y^{e,ph}, s_x^{ph})$, se calculará únicamente en términos de las probabilidades a priori y las matrices de covarianza de las Gaussianas correspondientes del siguiente modo

$$d_{KL}(s_y^{e,ph}, s_x^{ph}) = \frac{p(s_y^{e,ph})}{2} \sum_i [log(\frac{\mathbf{\Sigma}_{s_x^{ph}}(i,i)}{\mathbf{\Sigma}_{s_y^{e,ph}}(i,i)}) + \frac{\mathbf{\Sigma}_{s_y^{e,ph}}(i,i)}{\mathbf{\Sigma}_{s_x^{ph}}(i,i)} - 1] + p(s_y^{e,ph}) log(\frac{p(s_y^{e,ph})}{p(s_x^{ph})}), \quad (6.22)$$

donde $\mathbf{\Sigma}_{s_x^{ph}}(i,i)$ y $\mathbf{\Sigma}_{s_y^{e,ph}}(i,i)$ son el término i^{th}-ésimo de las matrices de covarianzas diagonal de las Gaussianas s_x^{ph} y $s_y^{e,ph}$ respectivamente. En el Anexo 6.7 de este mismo Capítulo se puede observar el desarrollo teórico para obtener la expresión de la distancia Kullback-Liebler para dos Gausianas.

Tal y como se puede apreciar, la distancia de Kullback-Liebler modificada no es simétrica ni proporcional a la verosimilitud entre s_x^{ph} y $s_y^{e,ph}$, que es lo que se pretende medir; por todo ello, se define una nueva variable que tiene en cuenta estos dos hechos y que se denomina pseudo-verosimilitud entre dos Gaussianas $pl_{KL}(s_y^{e,ph}, s_x^{ph})$

$$pl_{KL}(s_y^{e,ph}, s_x^{ph}) = \frac{1}{d_{KL}(s_y^{e,ph}, s_x^{ph}) + d_{KL}(s_x^{ph}, s_y^{e,ph})}, \tag{6.23}$$

Así pues, y con todo lo anterior, se estima finalmente $p_0(s_x^{ph}|s_y^{e,ph}, e, ph)$ de la siguiente manera

$$p_0(s_x^{ph}|s_y^{e,ph}, e, ph) = \frac{pl_{KL}(s_y^{e,ph}, s_x^{ph})}{\sum_{s_x^{ph}} pl_{KL}(s_y^{e,ph}, s_x^{ph})}. \tag{6.24}$$

Por otra parte, $\mathbf{r}_{0,s_x^{ph},s_y^{e,ph}}$ se obtiene sustituyendo $\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}$ por $\mu_{s_x^{ph}}$ en (6.18)

$$\mathbf{r}_{0,s_x^{ph},s_y^{e,ph}} = \frac{\sum_{t_{e,ph}} p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, e, ph)(\mathbf{y}_{t_{e,ph}}^{Tr,e,ph} - \mu_{s_x^{ph}})}{\sum_{t_{e,ph}} p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, e, ph)}. \tag{6.25}$$

Llegados a este punto, y con el proceso de inicialización ya concluido, se repitieron los experimentos expuestos en la Sección 5.4 para comprobar la validez de dicho proceso en términos de RAH. Así pues, se empleó la base de datos *SpeechDat Car* en español, *parametrización UZ* y modelos acústicos fonéticos. Los vectores de características, por su parte, se normalizaron haciendo uso de la técnica PD-MEMLIN utilizando únicamente el proceso de inicialización "ciego" para obtener los vectores de desplazamiento y el modelo de probabilidad entre Gaussianas. A su vez, y por simplicidad, cada fonema se modeló con cuatro Gaussianas, tanto para el espacio limpio como para cada uno de los entornos ruidosos básicos. Con todo lo anterior, la mejora media en términos de WER, MIMP, obtenida fue de 20.2 %; aún lejana, tal y como se verá más adelante, de la lograda con la técnica PD-MEMLIN con proceso de entrenamiento con señal estéreo, pero indicativa del correcto funcionamiento del proceso de inicialización de la fase de entrenamiento "ciega" propuesta.

Una vez calculado $\mathbf{r}_{0,s_x^{ph},s_y^{e,ph}}$, en la fase de ajuste posterior se obtiene $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$ de forma iterativa mediante el algoritmo EM [Dempster *et al.*, 1977]. En este caso, la correspondiente expresión para la iteación m-ésima, $\mathbf{r}_{m,s_x^{ph},s_y^{e,ph}}$, con $m > 0$, es (6.26). Cabe destacar que en el Anexo 6.8 de este mismo Capítulo se ha incluido el desarrollo teórico completo que da lugar a la expresión final de $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$ tras aplicar el algoritmo EM.

$$\mathbf{r}_{m,s_x^{ph},s_y^{e,ph}} = \frac{\sum_{t_{e,ph}} p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, e, ph) p(s_x^{ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, s_y^{e,ph}, m-1)(\mathbf{y}_{t_{e,ph}}^{Tr,e,ph} - \mu_{s_x^{ph}})}{\sum_{t_{e,ph}} p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, e, ph) p(s_x^{ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, s_y^{e,ph}, m-1)}, \tag{6.26}$$

$$p(s_x^{ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, s_y^{e,ph}, m-1) = \frac{\mathcal{N}(\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}; \mu_{s_x^{ph}} + \mathbf{r}_{m-1,s_x^{ph},s_y^{e,ph}}, \mathbf{\Sigma}_{s_y^{e,ph}}) p(s_y^{e,ph})}{\sum_{s_x^{ph}} \mathcal{N}(\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}; \mu_{s_x^{ph}} + \mathbf{r}_{m-1,s_x^{ph},s_y^{e,ph}}, \mathbf{\Sigma}_{s_y^{e,ph}}) p(s_y^{e,ph})}, \tag{6.27}$$

Una vez estimado el vector de desplazamiento en la fase de ajuste mediante el algoritmo EM, se realizaron los mismos experimentos anteriormente comentados, normalizando los vectores de características ruidosos en este caso con la técnica PD-MEMLIN aplicando $p_0(s_x^{ph}|s_y^{e,ph}, e, ph)$ y $\mathbf{r}_{m,s_x^{ph},s_y^{e,ph}}$. Las correspondientes MIMPs fueron de 41.03 % si $m = 1$, y 46.90 % si $m = 10$. Con esto se muestra el importante impacto que tiene la nueva estimación del vector de desplazamiento

obtenida tras la aplicación del algoritmo EM.

A la hora de mejorar la estimación de $p_0(s_x^{ph}|s_y^{e,ph},e,ph)$ en su correspondiente fase de ajuste, se utiliza señal pseudo-estéreo. Dicha señal se obtiene normalizando los vectores de características ruidosos del corpus de entrenamiento mediante la técnica PD-MEMLIN utilizando para cada uno de ellos únicamente aquellas variables del fonema estimado como correcto, $\hat{ph}$, entendiendo por fonema correcto aquél que proporciona la segmentación forzada realizada anteriormente mediante el algoritmo de Viterbi sobre la señal ruidosa. De esta manera, los vectores acústicos limpios que completan la señal de entrenamiento pseudo-estéreo, $\hat{\mathbf{x}}_{t_e^{ph}}^{Tr,e,ph}$, se obtendrán como

$$\hat{\mathbf{x}}_{t_e^{ph}}^{Tr,e,ph} = \mathbf{y}_{t_e^{ph}}^{Tr,e,ph} - \sum_{e}\sum_{s_y^{e,\hat{ph}}}\sum_{s_x^{\hat{ph}}} \mathbf{r}_{s_x^{\hat{ph}},s_y^{e,\hat{ph}}} p(e|\mathbf{y}_{t_e^{ph}}^{Tr,e,ph}) p(s_y^{e,\hat{ph}}|\mathbf{y}_{t_e^{ph}}^{Tr,e,ph},e,\hat{ph}) p(s_x^{\hat{ph}}|\mathbf{y}_{t_e^{ph}}^{Tr,e,ph},e,\hat{ph},s_y^e). \tag{6.28}$$

De esta manera, y una vez definida la señal de entrenamiento pseudo-estéreo $(\hat{\mathbf{X}}_{e,ph}^{Tr},\mathbf{Y}_{e,ph}^{Tr}) = \{(\hat{\mathbf{x}}_1^{Tr,e,ph},\mathbf{y}_1^{Tr,e,ph});...;(\hat{\mathbf{x}}_{t_{e,ph}}^{Tr,e,ph},\mathbf{y}_{t_{e,ph}}^{Tr,e,ph});...;(\hat{\mathbf{x}}_{T_{e,ph}}^{Tr,e,ph},\mathbf{y}_{T_{e,ph}}^{Tr,e,ph})\}$, se puede estimar una nueva iteración para $p(s_x^{ph}|s_y^{e,ph},e,ph)$ usando (6.20), o (6.21), según si emplea la decisión *hard* o *soft* respectivamente. La utilización de señal pseudo-estéreo se puede aplicar iterativamente tantas veces como se considere preciso.

Repitiendo el mismo experimento anteriormente comentado, y haciendo uso de una única iteración del algoritmo EM para estimar $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$, esto es, utilizando $\mathbf{r}_{1,s_x^{ph},s_y^{e,ph}}$ como vector de desplazamiento, y calculando $p(s_x^{ph}|s_y^{e,ph},e,ph)$ con señal de entrenamiento pseudo-estéreo obtenida a partir de la aplicación de la técnica PD-MEMLIN con $\mathbf{r}_{0,s_x^{ph},s_y^{e,ph}}$ y $p_0(s_x^{ph}|s_y^{e,ph},e,ph)$, se puede apreciar que se consigue una mejora media en términos de WER, MIMP, de 50.23 %, lo que certifica la potencia del uso de señal pseudo-estéreo. A raíz de este resultado se decidió usar la señal pseudo-estéreo también para mejorar la estimación de $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$. De este modo, si $p(s_x^{ph}|s_y^{e,ph},e,ph)$ se estima con tres iteraciones haciendo uso de señal pseudo-estéreo y, por otra parte, la primera iteración del vector de desplazamiento obtenida mediante el algoritmo EM, $\mathbf{r}_{1,s_x^{ph},s_y^{e,ph}}$, se ajusta con dos iteraciones adicionales con señal pseudo-estéreo, la MIMP para el experimento considerado anteriormente alcanza el 58.68 %, valor este muy cercano del que se obtendría si se aplicara la fase de entrenamiento con señal estéreo. Estos resultados muestran que la combinación del algoritmo EM y el uso de señal pseudo-estéreo conjuntamente en la fase previa de entrenamiento pueden proporcionar estimaciones de los vectores de desplazamiento y de los modelos de probabilidad entre Gaussianas bastante satisfactorios. Tal y como ya se ha comentado, las expresiones de las variables obtenidas en la fase de entrenamiento "ciega" expuestas en esta Sección se pueden generalizar para la técnica MEMLIN considerando que los espacios limpio y los asociados a los distintos entornos básicos están compuestos únicamente por un fonema.

A modo de resumen, en la Figura 6.2 se presenta el esquema gráfico de la fase de entrenamiento "ciega" para la técnica PD-MEMLIN que se va a utilizar en este trabajo.

6.5. Resultados con la base de datos *SpeechDat Car* en español.

La experimentación de las técnicas de normalización empíricas tratadas en las Secciones 6.1, 6.2, 6.3 y 6.4 del presente Capítulo, esto es, P-MEMLIN, MEMHIN y PD-MEMLIN, esta última

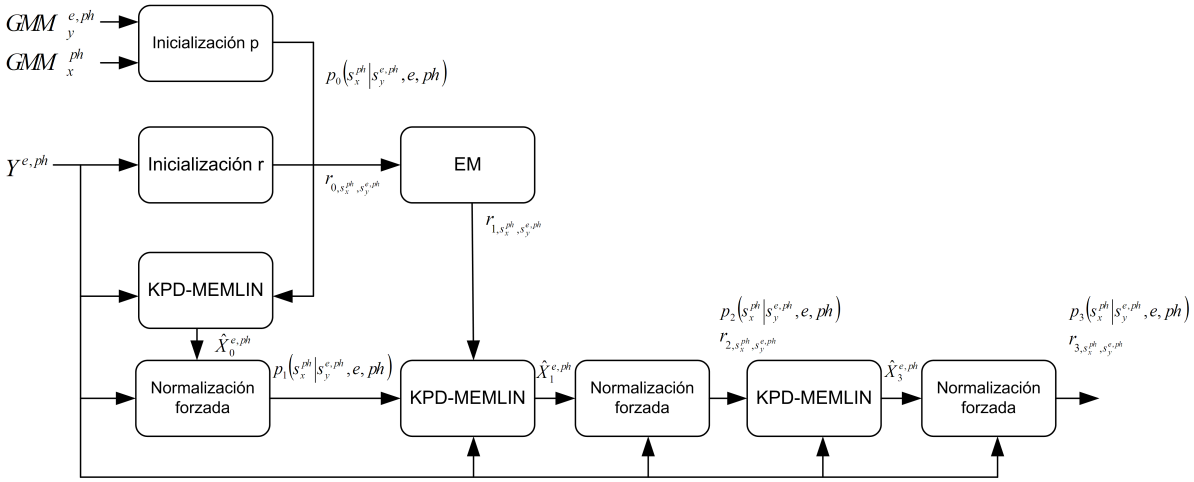


Figura 6.2: Representación gráfica del proceso de entrenamiento "ciego" de la técnica PD-MEMLIN que se va a emplear en este trabajo. A partir del bloque "Inicialización r" se obtiene la primera estimación del modelo de probabilidad entre Gaussian, $p_0(s_x^{ph}|s_y^{e,ph}, e, ph)$ (6.24), del mismo modo que "inicialización r" hace lo propio para el vector de deplazamiento, $r_{0,s_x^{ph},s_y^{e,ph}}$ (6.25). Por su parte, el bloque "EM" proporciona la primera iteración de ajuste para el vector de desplazamiento (6.26). La obtención de la señal pseudo-estéreo a partir de los vectores de características ruidosos se realiza mediante el sistema identificado como "Normalización forzada", que hace uso de la expresión (6.28). Finalmente, el bloque "Entrenamiento estéreo" obtiene los modelos de probabilidad entre Gaussianas y los vectores de desplazamiento, si es el caso, haciendo uso de las señal pseudo-estéreo (6.21) (6.18).

utilizando fase de entrenamiento tanto con señal estéreo, como en su versión "ciega", se realizó con la base de datos *SpeechDat Car* en español. A la hora de realizar la fase de entrenamiento previa para estimar los diversos parámetros necesarios para las distintas técnicas de normalización y entornos básicos, esto es, los vectores de desplazamiento y los modelos de probabilidad entre Gaussianas, se hará uso de los distintos corpora de entrenamiento correspondientes a cada entorno básico, utilizando bien señal estéreo, bien únicamente los vectores acústicos ruidosos, según el caso. Por otra parte, y una vez que se ha llevado a cabo la normalización de los vectores acústicos degradados con las correspondientes técnicas, se aplicará el método CMS. Para esta experimentación se utilizó la *parametrización UZ* y los modelos acústicos de las unidades fonéticas, pudiéndose, de este modo, consultar los resultados de referencia correspondientes en la Tabla 4.3. Se puede apreciar que todos los parámetros que definen la experimentación en este caso coinciden con los aplicados en la Sección 5.4, de modo que los resultados son totalmente comparables. Asimismo la Figura 5.5 sigue siendo válida para explicar los tres pasos precisados para llevar a cabo la experimentación.

6.5.1. Resultados de las técnicas P-MEMLIN y MEMHIN

En la Tabla 6.1 se pueden apreciar los mejores resultados para las técnicas de normalización de vectores de características MEMLIN, cuyos resultados ya se presentaron en la Sección 5.4 y ahora se repiten a modo de comparación, P-MEMLIN y MEMHIN. En todos los casos, junto al nombre de la técnica, se incluye el número de componentes que conforman las GMMs con que se modelan los entornos básicos ruidosos y el espacio limpio (se realizó un barrido con 4, 8, 16, 32, 64 y 128 componentes, cuyos resultados completos se incluyen entre los Anexos 6.9 y 5.6). Cabe destacar que de aquí en adelante para todas las técnicas tratadas en este Capítulo, y mientras no se indique lo contrario, el número de Gaussianas empleadas para modelar el espacio limpio será el mismo que

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	MEMLIN 128	2.30	7.46	4.62	6.39	8.77	5.40	8.16	6.05	70.22
CLK	P-MEMLIN 128	2.30	7.80	4.90	5.89	8.39	5.71	7.48	6.02	70.47
CLK	MEMHIN 128	2.21	7.89	5.17	6.02	8.29	5.56	7.82	6.05	70.22

Cuadro 6.1: Mejores resultados con la base de datos *SpeechDat Car* en español para las técnicas de normalización de vectores de características MEMLIN, P-MEMLIN y MEMHIN en términos de WER (%). Se presentan los resultados para los diferentes entornos básicos (E1,..., E7) utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia (CLK en la columna de Entrenamiento). La columna marcada como "Reconocimiento" hace referencia a la señal empleada para reconocer, que será la ruidosa normalizada con las técnicas P-MEMLIN, MEMHIN y MEMLIN, estos últimos resultados incluidos a modo de comparación. Junto al nombre de los diferentes métodos aparece el número de Gaussianas con que se modelaron los correspondientes espacios (el limpio y los asociados a los distintos entornos básicos). Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

el utilizado para representar cada entorno básico. Asimismo se incluye en la Tabla, además del WER medio, MWER, la mejora media de WER, MIMP, en tanto por ciento, que se calcula del mismo modo que ya se explicó en el Capítulo 5.4 (ver expresión (5.29)).

Por otra parte, y a pesar de que a simple vista ya se pueden intuir los resultados, es conveniente analizar mediante la prueba de hipótesis estadística *z-test* si el comportamiento de las técnicas propuestas en este apartado, P-MEMLIN y MEMHIN, es estadísticamente diferente con respecto al del algoritmo MEMLIN para la base de datos *SpeechDat Car* en español. De este modo, comparando los métodos MEMLIN y P-MEMLIN, el valor del estadístico W, w, es $w = 0{,}0673 < 1{,}96$, por lo que la mejora que proporciona el algoritmo P-MEMLIN en este caso no se puede considerar independiente de la base de datos con un intervalo de confianza del 95 %. Asimismo, comparar los mejores resultados para las técnicas MEMLIN y MEMHIN no tiene sentido alguno ya que éstos son idénticos. De todas maneras, a la hora de valorar las conclusiones obtenidas mediante la hipótesis estadística *z-test*, hay que tener en cuenta siempre las limitaciones de la propia prueba, ya comentadas en la Sección 4.2.

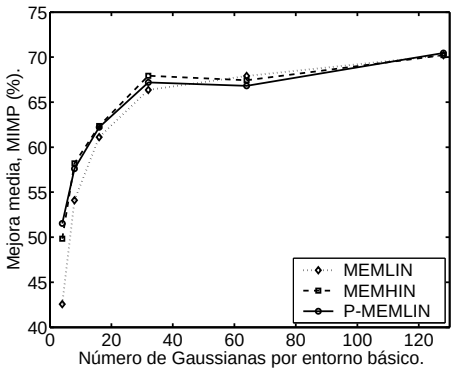


Figura 6.3: Mejora media de WER empleando las técnicas MEMLIN (línea punteada con diamantes círculos blancos), MEMHIN (línea discontinua con cuadrados blancos) y P-MEMLIN (línea continua con círculos blancos).

A la luz pues de los resultados presentados en la Tabla 6.1 se puede concluir que, teniendo en cuenta únicamente los mejores resultados medios para las distintas técnicas y para todos y cada uno de los entornos básicos, los métodos P-MEMLIN y MEMHIN no aportan ninguna mejora significativamente estadística con respecto al algoritmo MEMLIN. Sin embargo, si se representa la mejora media de WER para los métodos MEMLIN, P-MEMLIN y MEMHIN en función del número de Gaussianas con que se modela cada entorno básico (Figura 6.3) se puede apreciar que los dos últimos sí proporcionan un mejor comportamiento cuando el número de Gaussianas es reducido. Así, por ejemplo, si se aplica la técnica MEMLIN modelando cada entorno básico con 4 Gaussianas se obtiene un MIMP de 42.56 %, mientras que los algoritmos P-MEMLIN y MEMHIN alcanzan, bajo las mismas condiciones, valores sensiblemente mejores, 51.53 % y 49.84 %, respectivamente; aunque, eso sí, dicha mejora queda reducida a la mínima expresión cuando el número de Gaussianas por entorno básico se eleva por encima de 8. Este comportamiento se debe a que la compensación de la varianza de los vectores de características, que es lo que pretenden las técnicas P-MEMLIN y MEMHIN, se vuelve más importante cuando el modelado de los entornos básicos y el espacio limpio se realiza con un número reducido de Gaussianas, lo que se debe a que en ese caso el espacio modelado por cada Gaussiana es más variable y las transformaciones aprendidas en la fase de entrenamiento son más sensibles a la diferencia de varianzas entre los vectores de características asociados a las correspondientes Gaussianas. En estas situaciones, un modelo más complejo de $\mathbf{x}$ proporciona interesantes mejoras. Por otra parte, otro de los factores que puede hacer que técnicas como P-MEMLIN o MEMHIN tengan un mejor comportamiento es el tipo de ruido, ya que, por ejemplo, el ruido aditivo afecta en una mayor medida que otros a la varianza de los vectores de características, lo que se adecúa mejor a las características de métodos como P-MEMLIN o MEMHIN antes que a las del algoritmo MEMLIN. Para certificar esta afirmación se realizó una serie de experimentos haciendo uso de la base de datos *SpeechDat Car* en español comparando el comportamiento de los algoritmos MEMLIN y MEMHIN [Buera *et al.*, 2004b] atendiendo a distintas SNR. Así pues, en este caso los nuevos corpora ruidosos de entrenamiento y reconocimiento se obuvieron añadiendo artificialmente ruido aditivo de vehículo obtenido de la propia base de datos a los correspondientes corpora limpios. Para este nuevo experimento se utilizó la *parametrización UZ*, modelos acusticos de las unidades fonéticas y 8 ó 16 Gaussianas para modelar los entornos básicos y el espacio limpio para ambas técnicas. En la Figura 6.4 se muestra el WER medio en función de la SNR. Se puede apreciar que la técnica MEMHIN proporciona en todos los casos una cierta mejora, siendo ésta más importante para las situaciones más adversas (SNR reducidas).

6.5.2. Resultados para la técnica PD-MEMLIN

A continuación se comparan los resultados obtenidos con las técnicas MEMLIN y PD-MEMLIN, entrenando y empleando, en este último método, transformaciones para todos los 25 fonemas españoles más el silencio a pesar de que, para la tarea concreta de dígitos empleada en esta experimentación, no son todos necesarios. En la Tabla 6.2 se incluyen los mejores resultados para los dos métodos comparados en esta subsección, incluyendo, junto a sus respectivos nombres y en aras de establecer una comparación justa, el correspondiente número de transformaciones por entorno básico en $\log_{10}$, *Transformations per basic Environment*, TpE, que cada técnica debe calcular para normalizar un vector de características. Mediante este término se tiene una idea aproximada, para el algoritmo en cuestión, del coste computacional por vector acústico normalizado, pues se corresponde con el número de exponenciales que se han de evaluar. Así pues, el TpE se calcula, suponiendo que cada fonema de los entornos básicos ruidosos y el espacio limpio se modelan con el mismo número de Gaussianas, como

$$TpE = \log_{10}(n_{s_y^{ph}} n_{s_x^{ph}} n_{ph}), \tag{6.29}$$

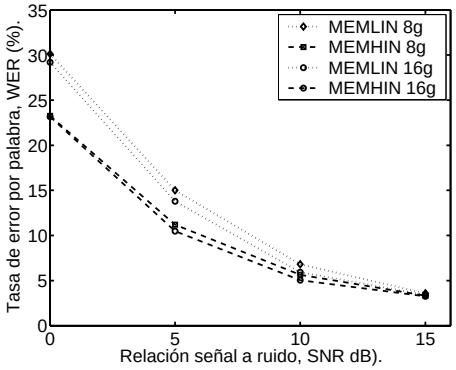


Figura 6.4: Mejora media de WER (MIMP) en % en función de la SNR para las técnicas MEMLIN y MEMHIN. En ambos casos se han modelado los entornos básicos con 8 ó 16 Gaussianas. La señal ruidosa procede de la base de datos *SpeechDat Car* en español a la que se le ha añadido artificialmente ruido aditivo de vehículo obtenido a partir de la misma base de datos.

donde $n_{s_y^{ph}}$ y $n_{s_x^{ph}}$ son el número de Gaussianas del modelo ruidoso y limpio para el fonema ph, respectivamente, y n_{ph} es el número de fonemas ($n_{ph} = 1$, para la técnica MEMLIN). En esta experimentación se utiliza el mismo número de Gaussianas para modelar cada fonema del espacio limpio y de cada entorno ruidoso básico, pudiéndo ser 2, 4, 8, 16 ó 32, y cuyos resultados completos se incluyen en el Anexo 6.9

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	MEMLIN 4.21	2.30	7.46	4.62	6.39	8.77	5.40	8.16	6.05	70.22
CLK	PD-MEMLIN 3.82	1.73	8.23	5.45	4.64	6.86	3.02	7.14	5.30	75.44

Cuadro 6.2: Mejores resultados con la base de datos *SpeechDat Car* en español para las técnicas de normalización de vectores de características MEMLIN y PD-MEMLIN en términos de WER (%). Se presentan los resultados para los diferentes entornos básicos (E1,..., E7) utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia (CLK en la columna de Entrenamiento). La columna marcada como "Reconocimiento" hace referencia a la señal empleada para reconocer, que será la ruidosa normalizada con las técnicas PD-MEMLIN o MEMLIN, estos últimos resultados incluidos a modo de comparación. Junto al nombre de los diferentes métodos aparece el número de transformaciones por entorno básico en $\log_{10}$, TpE. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

A la luz de los resultados presentados en la Tabla 6.2 se puede asegurar que la técnica PD-MEMLIN proporciona unos resultados medios, al menos para la combinación óptima de número de Gaussianas tratada, superiores a los obtenidos por el algoritmo MEMLIN.

Por otra parte, y para determinar si se puede afirmar o no que los resultados anteriores son estadísticamente significativos, se recurre a la prueba de hipótesis estadística *z-test*. En esta ocasión se comparan las técnicas MEMLIN y PD-MEMLIN bajo la base de datos, *SpeechDat Car* en español. Se puede observar que el valor del estadístico W, w, es $w = 1{,}731 < 1{,}96$, por lo que la mejora del algoritmo en este caso no se puede considerar independiente de la base de datos con un intervalo de confianza del 95 %. Sin embargo, si se compararan los mejores resultados obtenidos

por las técnicas SPLICE ME y PD-MEMLIN, se constata que $w = 3{,}25 > 1{,}96$, con lo que se puede considerar que la diferencia de comportamiento de estas dos últimas técnicas es estadísticamente significativa con un intervalo de confianza del 95 %. De todos modos, se recuerda una vez más que se han de considerar estos resultados con suma cautela dadas las limitaciones de la propia prueba, ya comentadas en la Sección 4.2.

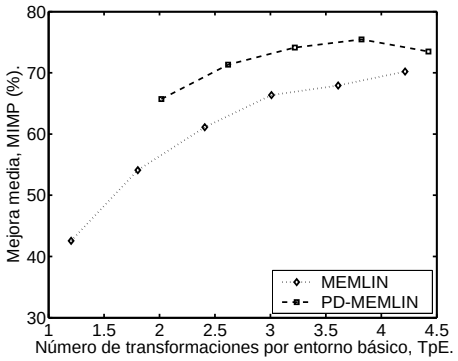
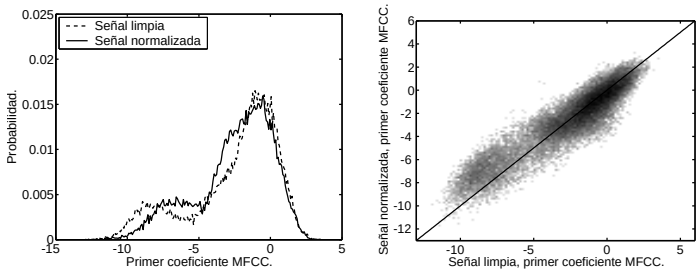


Figura 6.5: Mejora media de WER en función de TpE en $\log_{10}$ con la base de datos *SpeechDat Car* en español empleando las técnicas MEMLIN (línea punteada con diamantes blancos) y PD-MEMLIN (línea discontinua con cuadrados blancos). Se presentan los resultados utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia.

Asimismo, y para estudiar conjuntamente la tendencia del comportamiento de las técnicas MEMLIN y PD-MEMLIN en función del número de transformaciones por entorno básico, TpE en $\log_{10}$, se presenta la Figura 6.5. La tendencia observada demuestra que la técnica PD-MEMLIN proporciona una significativa mejora relativa con respecto al algoritmo MEMLIN, independientemente del número de transformaciones por entorno básico que se empleen. Por su parte, la Figura 6.6 muestra el histograma y el *log-scattegram* del primer coeficiente MFCC de los vectores de características de voz limpios del entorno básico E4 de la base de datos *SpeechDat Car* en español y los correspondientes normalizados mediante el método PD-MEMLIN empleando 16 Gaussianas por fonema. A partir de esta Figura se puede concluir que las transformaciones propuestas por el algoritmo PD-MEMLIN solucionan el problema de la proyección de gran cantidad de vectores de características ruidosos hacia el silencio del espacio limpio, como así sucedía en la técnica MEMLIN (ver Figura 5.7.b), a la vez que se reduce la incertidumbre y se elimina buena parte de los efectos que el entorno acústico introducía en los coeficientes de los vectores de características (ver Figura 5.7.a). Todo esto es consecuencia de que el algoritmo PD-MEMLIN reduce, como ya se ha indicado, el espacio de proyección de los correspondientes vectores de desplazamiento a nivel de fonema, produciendo unos vectores de características normalizados que se adaptan mejor a los modelos acústicos.

Para conocer el límite de la técnica PD-MEMLIN, se llevó a cabo un nuevo experimento en el que cada vector de características se normalizó únicamente con las transformaciones propias del correspondiente fonema "correcto", $\hat{ph}$, que se obtuvo mediante segmentación forzada en términos de fonema de la señal limpia del corpus de reconocimiento a partir del algoritmo de Viterbi. A esta pseudo-técnica, por comodidad en la nomenclatura, se le denomina KPD-MEMLIN, *Known* PD-MEMLIN, y viene definida por la expresión



(a.1) (a.2)

Figura 6.6: *Log-scattegram* e histograma realizados entre el primer coeficiente MFCC de las tramas de voz de la señal limpia (eje de abscisas) y la señal normalizada usando la técnica PD-MEMLIN con 16 Gaussianas por fonema para cada entorno básico (a) (eje de ordenadas). Las representaciones se realizaron a partir del corpus de reconocimiento del entorno básico E4 de la base de datos *SpeechDat Car*. La línea en los *log-scattegrams* representa la función $x = y$.

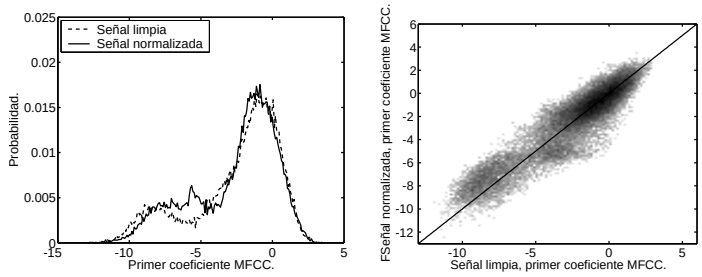
$$\hat{\mathbf{x}}_t = \mathbf{y}_t - \sum_{e} \sum_{s_y^{e,\hat{ph}}} \sum_{s_x^{\hat{ph}}} \mathbf{r}_{s_x^{\hat{ph}}, s_y^{e,\hat{ph}}} p(e|\mathbf{y}_t) p(s_y^{e,\hat{ph}}|\mathbf{y}_t, e, \hat{ph}) p(s_x^{\hat{ph}}|\mathbf{y}_t, e, \hat{ph}, s_y^e). \tag{6.30}$$

Nótese que la expresión (6.30) es, conceptualmente, la misma que (6.28), que ya fue utilizada en la fase de entrenamiento "ciega" desarrollada para la técnica PD-MEMLIN (Sección 6.4). En la Tabla 6.3 se muestran, junto con los resultados de reconocimiento obtenidos con la señal limpia (Entrenamiento CLK, Reconocimiento CLK), incluidos para comparar, los correspondientes tras aplicar la técnica KPD-MEMLIN. Dicha experimentación se realizó con la base de datos *SpeechDat Car* en español haciendo uso de la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de entrenamiento de señal limpia (Entrenamiento CLK). Por otra parte, las GMMs de los fonemas se compusieron con 16 componentes para los distintos entornos básicos y el espacio limpio. Asimismo, y por completar el estudio de la pseudo-técnica KPD-MEMLIN, también se incluyen los correspondientes *log-scattegram* e histograma. Ambos se obtuvieron a partir del primer coeficiente MFCC de los vectores de caracteristicas limpios del corpus de reconocimiento del entorno básico E4 y los correspondientes normalizados mediante la técnica KPD-MEMLIN bajo las condiciones de experimentación consideradas anteriormente (Figura 6.7).

De la Tabla 6.3 y de la Figura 6.7 se puede constatar que la mejora media en WER proporcionada por la pseudo-técnica KPD-MEMLIN se acerca al 100 % si se modela cada fonema con 16 Gaussianas; sin embargo, la incertidumbre del primer coeficiente MFCC de los vectores acústicos normalizados del entorno básico E4 no se ha reducido considerablemente con respecto a la obtenida con la técnica PD-MEMLIN bajo las mismas condiciones de experimentación (Figura 6.6). Esto último es debido a que, en este caso, las normalizaciones proyectan los vectores de características ruidosos al espacio limpio a nivel de fonema, que ya tiene de por sí una cierta incertidumbre que queda modelada mediante las varianzas de los modelos acústicos correspondientes. Este hecho puede confirmarse mediante el estudio de la tasa media de fonemas correctos, *Mean Correct Phoneme*, MCP. Para este propósito, se considera que el fonema correcto para cada vector acústico es aquel dado por la segmentación forzada en términos de fonema de la señal limpia gracias al algoritmo de Viterbi. A su vez, la tasa MCP se obtiene como la relación de fonemas correctamente reconocidos haciendo uso de las GMMs con que se modelan dichas unidades pertenecientes al espacio limpio y sin considerar modelo de lenguaje o vocabulario alguno. Los resultados de la tasa MCP obtenidos

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	CLK	1.90	2.64	1.81	1.75	1.62	0.64	0.35	1.75	–
CLK	KPD-MEMLIN 3.82	0.96	2.57	2.52	1.75	2.10	1.27	1.02	1.84	99.37

Cuadro 6.3: Resultados con la base de datos *SpeechDat Car* en español para la pseudo-técnica de normalización de vectores de características KPD-MEMLIN en términos de WER (%). Se presentan los resultados para los diferentes entornos básicos (E1,..., E7) utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia (CLK en la columna de Entrenamiento). La columna marcada como "Reconocimiento" hace referencia a la señal empleada para reconocer, que será la ruidosa normalizada con las pseudo-técnica KPD-MEMLIN. También se incluyen a modo de comparación los resultados obtenidos con señal limpia (CLK en la columna de Reconocimiento). Junto al nombre del método KPD-MEMLIN aparece el número de transformaciones por entorno básico en $\log_{10}$, TpE. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.



(a.1) (a.2)

Figura 6.7: *Log-scattegram* e histograma realizados entre el primer coeficiente MFCC de las tramas de voz de la señal limpia (eje de abscisas) y la señal normalizada usando la pseudo-técnica kPD-MEMLIN con 16 Gaussianas por fonema para cada entorno básico (a) (eje de ordenadas). Las representaciones se realizaron a partir del corpus de reconocimiento del entorno básico E4 de la base de datos *SpeechDat Car*. La línea en los *log-scattegrams* representa la función $x = y$.

con los algoritmos PD-MEMLIN y KPD-MEMLIN para los distintos entornos básicos del corpus de reconocimiento se muestran en la Tabla 6.4. En ambas técnicas cada fonema se ha modelado con 16 Gaussianas para todos los entornos básicos y el espacio limpio, pudiéndose apreciar como la pseudo-técnica KPD-MEMLIN mejora, en media, los resultados del método PD-MEMLIN más allá del 10 %, a pesar de que, como ya se ha constatado anteriormente, la incertidumbre apenas se ha visto reducida, por lo que se puede concluir que la proyección de los vectores de características a nivel de fonema correcto es sensiblemente más eficaz para el método KPD-MEMLIN que para el algoritmo PD-MEMLIN.

A partir de las Tablas 6.3 y 6.4, se puede concluir que las transformaciones dependientes de cada fonema consideradas para el método PD-MEMLIN son consistentes con respecto a los modelos acústicos, ya que los vectores de características se proyectan del espacio ruidoso al limpio a nivel de fonemas. Asimismo, y gracias a los estudios realizados, se puede definir una futura línea de trabajo basada en dotar a la técnica PD-MEMLIN de una mejor estimación de la probabilidad a posteriori del fonema ph, dado el vector de características, $\mathbf{y}_t$, y el entorno básico e, $p(ph|\mathbf{y}_t, e)$. Por otra parte, también habría que estudiar el comportamiento de la pseudo-técnica KPD-MEMLIN en sistemas de vecrificación e identificación de locutor dependiente del texto en entornos acústicos adversos, ya que, a priori, podría proporcionar interesantes resultados. En este sentido ya se han

MCP (%)	E1	E2	E3	E4	E5	E6	E7	Mean
PD-MEMLIN 16-16	32.64	31.23	30.38	32.54	32.04	34.14	31.21	32.03
KPD-MEMLIN 16-16	37.68	40.15	39.87	43.06	45.15	48.35	50.28	42.42

Cuadro 6.4: Tasa media de fonemas correctos, *Mean Correct Phoneme*, MCP, en % para los vectores de características de voz normalizadas mediante las técnicas PD-MEMLIN y KPD-MEMLIN, que han sido aplicadas al entorno básico E4 del corpus de reconocimiento de la base de datos *SpeechDat Car* en español. Para ambos métodos se modelan los fonemas con 16 Gaussianas para todos los entornos básicos y el espacio limpio.

realizado unas primeras pruebas preliminares [Buera *et al.*, 2005d], llegándose a la conclusión de que la proyección de los vectores de características ruidosos a un espacio limpio genérico puede eliminar parte de la especificidad propia de cada locutor lo que, de cara a tareas de verificación e identificación de ocutor, no es deseable. A pesar de ello se obtuvieron importantes mejoras en ambas tareas.

6.5.3. Resultados para la técnica PD-MEMLIN con fase de entrenamiento "ciega"

A continuación se comparan los resultados obtenidos con las técnicas MEMLIN, PD-MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega", definiéndose dicha fase a partir de la desarrollada en la Sección 6.4, esto es, tres iteraciones realizadas con señal pseudo-estéreo para estimar $p(s_x^{ph}|s_y^{e,ph}, e, ph)$ y $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$ obtenida mediante dos iteraciones con señal pseudo-estéreo, una vez que se ha calculado $\mathbf{r}_{1,s_x^{ph},s_y^{e,ph}}$ con el algoritmo EM. Para realizar la correspondiente normalización con la técnica PD-MEMLIN, independientemente de su fase de entrenamiento y al igual que la subsección 6.5.2, se entrenaron y emplearon transformaciones para todos los 25 fonemas españoles más el silencio. Por su parte, el número de Gaussianas con que se modelan los distintos fonemas puede ser 2, 4, 8, 16 ó 32, y los correspondientes resultados completos se incluyen en el Anexo 6.9. Nótese que las condiciones de la experimentación son las mismas que las definidas previamente para la subsecciones 6.5.2 y 6.5.1, por lo que los resultados son totalmente comparables. En la Tabla 6.5 se incluyen los mejores resultados para los tres métodos comparados en esta ocasión, incluyendo, junto a sus respectivos nombres, el correspondiente número de transformaciones por entorno básico en $\log_{10}$, TpE, que cada método tiene que calcular para normalizar un vector de características ruidoso.

Nuevamente se recurre a la prueba de hipótesis estadística *z-test* para determinar si se puede afirmar o no que los mejores resultados presentados anteriormente son estadísticamente significativos. En este caso, y dado que el comportamiento de la técnica estudiada en esta subsección es algo más pobre que el alcanzado por el algoritmo PD-MEMLIN, no tiene sentido calcular el correspondiente valor de w por cuanto se puede afirmar que será inferior a 1.731. Así, se puede aseverar que los métodos MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega" no presenten comportamientos estadísticamente diferentes independientemente de la base de datos, *SpeechDat Car* en español con un intervalo de confianza del 95 %. Por su parte, si se compara con la técnica SPLICE ME se puede observar que $w = 2{,}23 > 1{,}96$, por lo que la mejora del algoritmo en este caso sí se puede considerar independiente de la base de datos con el intervalo de confianza elegido. Nuevamente se hace hincapié en que a la hora de valorar las conclusiones obtenidas mediante la hipótesis estadística *z-test*, hay que tener presente siempre las limitaciones de la propia prueba, ya comentadas en la Sección 4.2.

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	MEMLIN 4.21	2.30	7.46	4.62	6.39	8.77	5.40	8.16	6.05	70.22
CLK	PD-MEMLIN 3.82	1.73	8.23	5.45	4.64	6.86	3.02	7.14	5.30	75.44
CLK	PD-MEMLIN "ciego" 3.82	2.59	6.43	4.34	6.14	8.39	4.44	9.86	5.74	72.40

Cuadro 6.5: Mejores resultados con la base de datos *SpeechDat Car* en español para las técnicas de normalización de vectores de características MEMLIN, PD-MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega" en términos de WER (%). Se presentan los resultados para los diferentes entornos básicos (E1,..., E7) utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia (CLK en la columna de Entrenamiento). La columna marcada como "Reconocimiento" hace referencia a la señal empleada para reconocer, que será la ruidosa normalizada con las técnicas PD-MEMLIN con fase de entrenamiento "ciega", PD-MEMLIN o MEMLIN, estos últimos resultados incluidos a modo de comparación. Junto al nombre de los diferentes métodos aparece el número de transformaciones por entorno básico en $\log_{10}$, TpE. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Asimismo, y para estudiar la tendencia del comportamiento de los distintos algoritmos considerados en esta subsección, en la Figura 6.8 se representan las mejoras medias de WER correspondientes en función de TpE. Los resultados muestran que la técnica PD-MEMLIN con fase de entrenamiento "ciega" es capaz de proporcionar unos resultados similares, e incluso en algún caso mejores, a los logrados por el método MEMLIN para los distintos valores de TpE estudiados, con la ventaja añadida de que no es necesario disponer se señal estéreo para obtener los vectores de desplazamiento y los modelos de probabilidad entre Gaussianas. Sin embargo, las mejoras obtenidas con esta técnica aún quedan lejos del mejor resultado alcanzado hasta el momento por la técnica PD-MEMLIN con fase de entrenamiento con señal estéreo.

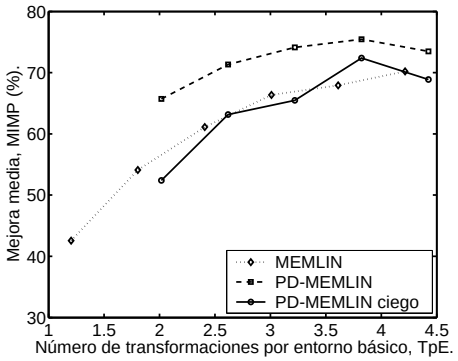


Figura 6.8: Mejora media de WER en función de TpE en $\log_{10}$ con la base de datos *SpeechDat Car* en español empleando las técnicas MEMLIN (línea punteada con diamantes blancos), PD-MEMLIN (línea discontinua con cuadrados blancos) y PD-MEMLIN con fase de entrenamiento "ciega" (línea continua con círculos blancos). Se presentan los resultados utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia.

A modo de resumen se incluyen en la Tabla 6.6 los resultados más significativos (MWER y MIMP) de las distintas técnicas presentadas en este Capítulo (P-MEMLIN, MEMHIN, PD-

MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega"), indicando a su vez el TpE requerido en cada caso. Adicionalmente, y por completar la comparación, se han incluido los resultados obtenidos con el método MEMLIN. Se puede observar que el algoritmo PD-MEMLIN obtiene los mejores resultados con el menor TpE, mientras que la versión de la misma técnica con fase de entrenamiento "ciega" alcanza, con el mismo TpE, un MWER menor que los obtenidos con técnicas como P-MEMLIN, MEMHIN y MEMLIN, que sí empleaban señal estéreo en su correspondiente fase de entrenamiento.

	TpE	MWER (%)	MIMP (%)
MEMLIN	4.21	6.05	70.22
MEMHIN	4.21	6.05	70.22
P-MEMLIN	4.21	6.02	70.47
PD-MEMLIN	3.82	5.30	75.44
PD-MEMLIN "ciego"	3.82	5.74	72.40

Cuadro 6.6: Mejores resultados en términos de WER medio (MWER) y mejora media en WER (MIMP) en % para los métodos MEMLIN, MEMHIN, P-MEMLIN, PD-MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega" (identificado como PD-MEMLIN "ciego"), indicando a su vez el TpE requerido en cada caso.

6.6. Anexo C.

En este Anexo se incluye el desarrollo teórico necesario para estimar la matriz diagonal asociada al término de pendiente y el vector que representa el término independiente del modelo de $\mathbf{x}$, $\mathbf{A}_{s_x,s_y^e}$ y $\mathbf{b}_{s_x,s_ye}$ respectivamente, correspondiente a la técnica P-MEMLIN. Para ello se hace necesario el empleo de señal de entrenamiento estéreo, $(\mathbf{X}_e, \mathbf{Y}_e) = \{(\mathbf{x}_1^e, \mathbf{y}_1^e); ...; (\mathbf{x}_{t_e}^e, \mathbf{y}_{t_e}^e); ...; (\mathbf{x}_{T_e}^e, \mathbf{y}_{T_e}^e)\}$, con $t_e \in [1, T_e]$; nótese que, por simplificar la notación, se ha eliminado el índice Tr que aparecía en la Sección 6.1 para indicar que se trata del corpus de entrenamiento. Así pues, $\mathbf{A}_{s_x,s_y^e}$ y $\mathbf{b}_{s_x,s_ye}$ se obtendrán igualando tanto la media como la desviación típica asociadas a cada par de Gaussianas, s_x y s_y^e, y correspondientes a los vectores de características limpios y los obtenidos mediante el modelo de $\mathbf{x}$ ($\mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e) = \mathbf{A}_{s_x,s_y^e}\mathbf{y}_t - \mathbf{b}_{s_x,s_y^e}$). Este criterio se toma debido a que lo que se pretende con esta técnica es acercar la pdf de los vectores de características ruidosos asociada al par de Gaussianas s_x y s_y^e a la correspondiente de los vectores limpios. De este modo, las ecuaciones que habrá que considerar son las siguientes

$$\mu_{s_x,s_y^e}^{\mathbf{x}} = \mu_{s_x,s_y^e}^{\mathbf{A}_{s_x,s_y^e}\mathbf{y}-\mathbf{b}_{s_x,s_y^e}}, \tag{C.1}$$

$$\sqrt{\mathbf{\Sigma}_{s_x,s_y^e}^{\mathbf{x}}} = \sqrt{\mathbf{\Sigma}_{s_x,s_y^e}^{\mathbf{A}_{s_x,s_y^e}\mathbf{y}-\mathbf{b}_{s_x,s_y^e}}}, \tag{C.2}$$

donde el operador $\sqrt{\bullet}$ realiza la raíz cuadrada elemento a elemento de la matriz $\bullet$. Además, hay que tener en cuenta que el vector de medias y la matriz diagonal de covarianzas asociadas al par de Gaussianas s_x y s_y^e de una determinada variable z se definen del siguiente modo

$$\mu_{s_x,s_y^e}^{\mathbf{z}} = \frac{\sum_{t_e} p(s_x|\mathbf{x}_{t_e}^e)p(s_y^e|\mathbf{y}_{t_e}^e)\mathbf{z}_{t_e}^e}{\sum_{t_e} p(s_x|\mathbf{x}_{t_e}^e)p(s_y^e|\mathbf{y}_{t_e}^e)}, \tag{C.3}$$

$$\mathbf{\Sigma}^{\mathbf{z}}_{s_x,s_y^e} = diag[\frac{\sum_{t_e} p(s_x|\mathbf{x}^e_{t_e})p(s_y^e|\mathbf{y}^e_{t_e})(\mathbf{z}^e_{t_e} - \mu^{\mathbf{z}}_{s_x,s_y^e})(\mathbf{z}^e_{t_e} - \mu^{\mathbf{z}}_{s_x,s_y^e})^T}{\sum_{t_e} p(s_x|\mathbf{x}^e_{t_e})p(s_y^e|\mathbf{y}^e_{t_e})}], \tag{C.4}$$

donde el operador $diag[\bullet]$ hace nulos todos los elementos de la matriz $\bullet$ distintos de la diagonal. Al considerar que no hay dependencia alguna entre las componentes de los vectores de características, el sistema de ecuaciones compuesto por las expresiones (C.1) y (C.2) se puede ver como tantos sistemas independiente de dos ecuaciones como componentes tengan los vectores de características. Así pues, dichas expresiones se pueden ver, sin perder generalidad y haciendo uso de (C.3) y (C.4), del siguiente modo

$$\mu^{\mathbf{x}}_{s_x,s_y^e}(i) = \mathbf{A}_{s_x,s_y^e}(i,i)\mu^{\mathbf{y}}_{s_x,s_y^e}(i) - \mathbf{b}_{s_x,s_y^e}(i), \tag{C.5}$$

$$\sqrt{\mathbf{\Sigma}^{\mathbf{x}}_{s_x,s_y^e}(i,i)} = \mathbf{A}_{s_x,s_y^e}(i,i)\sqrt{\mathbf{\Sigma}^{\mathbf{y}}_{s_x,s_y^e}(i,i)}, \tag{C.6}$$

donde i representa el índice de la componente correspondiente, ya sea para los vectores de medias como para las matrices diagonales de las covarianzas. De este modo, y despejando convenientemente, las expresiones finales para $\mathbf{A}_{s_x,s_y^e}(i,i)$ y $\mathbf{b}_{s_x,s_ye}(i)$ son

$$\mathbf{b}_{s_x,s_y^e}(i) = \frac{\sqrt{\mathbf{\Sigma}^{\mathbf{x}}_{s_x,s_y^e}(i,i)}}{\sqrt{\mathbf{\Sigma}^{\mathbf{y}}_{s_x,s_y^e}(i,i)}}\mu^{\mathbf{y}}_{s_x,s_y^e}(i) - \mu^{\mathbf{x}}_{s_x,s_y^e}(i), \tag{C.7}$$

$$\mathbf{A}_{s_x,s_y^e}(i,i) = \frac{\sqrt{\mathbf{\Sigma}^{\mathbf{x}}_{s_x,s_y^e}(i,i)}}{\sqrt{\mathbf{\Sigma}^{\mathbf{y}}_{s_x,s_y^e}(i,i)}}, \tag{C.8}$$

que coinciden elemento a elemento con las expresiones presentadas en (6.3) y (6.4), respectivamente.

6.7. Anexo D.

En este Anexo se incluye el desarrollo teórico necesario para estimar el correspondiente vector de desplazamiento $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$ para la técnica PD-MEMLIN a partir de la minimización del error cuadrático medio asociado a cada par de Gaussianas, s_x^{ph} y $s_y^{e,ph}$, ($\xi_{s_x^{ph},s_y^{e,ph}}$). Para ello es necesario un corpus de entrenamiento estéreo $(\mathbf{X}_{e,ph}, \mathbf{Y}_{e,ph}) = \{(\mathbf{x}_1^{e,ph}, \mathbf{y}_1^{e,ph}); ...; (\mathbf{x}_{t_{e,ph}}^{e,ph}, \mathbf{y}_{t_{e,ph}}^{e,ph}); ...; (\mathbf{x}_{T_{e,ph}}^{e,ph}, \mathbf{y}_{T_{e,ph}}^{e,ph})\}$, con $t_{e,ph} \in [1, T_{e,ph}]$; nótese que, por simplificar la notación, se ha eliminado el índice Tr para indicar que se trata del corpus de entrenamiento, tal y como sí está recogido en la Sección 6.3. De este modo, el error cuadrático medio asociado a cada par de Gaussianas, s_x^{ph} y $s_y^{e,ph}$, para la técnica PD-MEMLIN, $\xi_{s_x^{ph},s_y^{e,ph}}$, se define como

$$\xi_{s_x^{ph},s_y^{e,ph}} = \frac{1}{T_{e,ph}} \sum_{t_{e,ph}} p(s_x|\mathbf{x}^{e,ph}_{t_{e,ph}}, e)p(s_y^e|\mathbf{y}^{e,ph}_{t_{e,ph}}, e)$$
$$\times Tra[(\mathbf{x}^{e,ph}_{t_{e,ph}} - \Psi(\mathbf{y}^{e,ph}_{t_{e,ph}}, s_x^{ph}, s_y^{e,ph}))(\mathbf{x}^{e,ph}_{t_{e,ph}} - \Psi(\mathbf{y}^{e,ph}_{t_{e,ph}}, s_x^{ph}, s_y^{e,ph}))^T], \tag{D.1}$$

donde $\Psi(\mathbf{y}^{e,ph}_{t_{e,ph}}, s_x^{ph}, s_y^{e,ph}) = \mathbf{y}^{e,ph}_{t_{e,ph}} - \mathbf{r}_{s_x^{ph},s_y^{e,ph}}$. Teniendo en cuenta esto último, así como distintas propiedades de cálculo matricial, se puede observar, antes de llevar a cabo la minimización de $\xi_{s_x^{ph},s_y^{e,ph}}$, que

$$(\mathbf{x}_{t_e}^{e,ph} - \Psi(\mathbf{y}_{t_e}^{e,ph}, s_x^{ph}, s_y^{e,ph}))(\mathbf{x}_{t_e}^{e,ph} - \Psi(\mathbf{y}_{t_{e,ph}}^{e,ph}, s_x^{ph}, s_y^{e,ph}))^T = \mathbf{x}_{t_{e,ph}}^{e,ph}(\mathbf{r}_{s_x^{ph},s_y^{e,ph}})^T - \mathbf{x}_{t_{e,ph}}^{e,ph}(\mathbf{y}_{t_{e,ph}}^{e,ph})^T$$
$$+\mathbf{x}_{t_{e,ph}}^{e,ph}(\mathbf{y}_{t_{e,ph}}^{e,ph})^T - \mathbf{y}_{t_{e,ph}}^{e,ph}(\mathbf{r}_{s_x^{ph},s_y^{e,ph}})^T + \mathbf{y}_{t_{e,ph}}^{e,ph}(\mathbf{y}_{t_{e,ph}}^{e,ph})^T$$
$$-\mathbf{y}_{t_{e,ph}}^{e,ph}(\mathbf{x}_{t_{e,ph}}^{e,ph})^T + \mathbf{r}_{s_x^{ph},s_y^{e,ph}}^T(\mathbf{r}_{s_x^{ph},s_y^{e,ph}})^T - \mathbf{r}_{s_x^{ph},s_y^{e,ph}}(\mathbf{y}_{t_{e,ph}}^{e,ph})^T$$
$$+\mathbf{r}_{s_x^{ph},s_y^{e,ph}}(\mathbf{x}_{t_{e,ph}}^{e,ph})^T \quad \text{(D.2)}$$

A la hora de estimar el vector de desplazamiento, $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$, se procede a la minimización de la expresión (D.1) con respecto a $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$ haciendo uso de (D.2). Para ello es necesario

$$\mathbf{0} = \frac{\delta\xi_{s_x^{ph},s_y^{e,ph}}}{\delta\mathbf{r}_{s_x^{ph},s_y^{e,ph}}} = \frac{1}{T_{e,ph}}\sum_{t_{e,ph}} p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{e,ph}, e)p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, e)$$
$$\frac{\delta}{\delta\mathbf{r}_{s_x^{ph},s_y^{e,ph}}}[Tra[\mathbf{x}_{t_{e,ph}}^{e,ph}(\mathbf{r}_{s_x^{ph},s_y^{e,ph}})^T - \mathbf{x}_{t_{e,ph}}^{e,ph}(\mathbf{y}_{t_{e,ph}}^{e,ph})^T + \mathbf{x}_{t_{e,ph}}^{e,ph}(\mathbf{y}_{t_{e,ph}}^{e,ph})^T$$
$$-\mathbf{y}_{t_{e,ph}}^{e,ph}(\mathbf{r}_{s_x^{ph},s_y^{e,ph}})^T + \mathbf{y}_{t_{e,ph}}^{e,ph}(\mathbf{y}_{t_{e,ph}}^{e,ph})^T - \mathbf{y}_{t_{e,ph}}^{e,ph}(\mathbf{x}_{t_{e,ph}}^{e,ph})^T$$
$$+\mathbf{r}_{s_x^{ph},s_y^{e,ph}}^T(\mathbf{r}_{s_x^{ph},s_y^{e,ph}})^T - \mathbf{r}_{s_x^{ph},s_y^{e,ph}}(\mathbf{y}_{t_{e,ph}}^{e,ph})^T + \mathbf{r}_{s_x^{ph},s_y^{e,ph}}(\mathbf{x}_{t_{e,ph}}^{e,ph})^T]]. \quad \text{(D.3)}$$

O, lo que es lo mismo

$$\mathbf{0} = \frac{1}{T_{e,ph}}\sum_{t_{e,ph}} p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{e,ph}, e)p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, e)(\mathbf{x}_{t_{e,ph}}^{e,ph} - \mathbf{y}_{t_{e,ph}}^{e,ph} + 2\mathbf{r}_{s_x^{ph},s_y^{e,ph}} + \mathbf{x}_{t_{e,ph}}^{e,ph} - \mathbf{y}_{t_{e,ph}}^{e,ph}). \quad \text{(D.4)}$$

A partir de la expresión anterior, y tras despejar convenientemente, se obtiene finalmente la expresión óptima deseada para $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$

$$\mathbf{r}_{s_x^{ph},s_y^{e,ph}} = \frac{\sum_{t_{e,ph}} p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{e,ph}, e)p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, e)(\mathbf{y}_{t_{e,ph}}^{e,ph} - \mathbf{x}_{t_{e,ph}}^{e,ph})}{\sum_{t_{e,ph}} p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{e,ph}, e)p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, e)}. \quad \text{(D.5)}$$

6.8. Anexo E.

La distancia de Kullback Leibler, $KL(p,q)$, es una medida de similitud entre dos funciones de densidad de probabilidad, en este caso continuas y denominadas p y q, que se calcula del siguiente modo

$$KL(p,q) = \int_{\mathbf{x}} p(\mathbf{x}) log(\frac{p(\mathbf{x})}{q(\mathbf{x})}) d\mathbf{x}, \quad \text{(E.1)}$$

donde se puede apreciar que p y q son pdfs de una variable vectorial, $\mathbf{x}$, y que se suponen para este estudio Gaussianas

$$p(\mathbf{x}) = \frac{c_p}{(2\pi)^{D/2}|\mathbf{\Sigma}_p|^{1/2}} e^{-1/2(\mathbf{x}-\mu_p)^T\mathbf{\Sigma}_p^{-1}(\mathbf{x}-\mu_p)}, \quad \text{(E.2)}$$

$$q(\mathbf{x}) = \frac{c_q}{(2\pi)^{D/2}|\mathbf{\Sigma}_q|^{1/2}} e^{-1/2(\mathbf{x}-\mu_q)^T\mathbf{\Sigma}_q^{-1}(\mathbf{x}-\mu_q)}, \quad \text{(E.3)}$$

donde c_p y c_q son las probabiliades a priori de las Gaussianas correspondientes y μ_p, μ_q, $\mathbf{\Sigma}_p$ y $\mathbf{\Sigma}_q$ son los vectores de medias y las matrices diagonales de covarianzas de las pdfs p y q, respectivamente. Por último D es la dimensión delvector $\mathbf{x}$. Si se evalúa el término logarítmico de la expresión (E.1) introdciendo (E.2) y (E.3) se tiene que

$$log(\frac{p(\mathbf{x})}{q(\mathbf{x})}) = log(\frac{c_p}{c_q}) + \frac{1}{2}log(\frac{|\mathbf{\Sigma}_q|}{|\mathbf{\Sigma}_p|}) - \frac{1}{2}((\mathbf{x}-\mu_p)^T\mathbf{\Sigma}_p^{-1}(\mathbf{x}-\mu_p) - (\mathbf{x}-\mu_q)^T\mathbf{\Sigma}_q^{-1}(\mathbf{x}-\mu_q)). \quad \text{(E.4)}$$

Teniendo en cuenta que las matrices de covarianzas, $\mathbf{\Sigma}_p$ y $\mathbf{\Sigma}_q$, son diagonales se tienen las siguientes igualdades

$$|\mathbf{\Sigma}_z| = \prod_i \mathbf{\Sigma}_z(i,i), \quad \text{(E.5)}$$

$$(\mathbf{x}-\mu_z)^T\mathbf{\Sigma}_z^{-1}(\mathbf{x}-\mu_z) = \sum_i \frac{(\mathbf{x}(i)-\mu_z(i))^2}{\mathbf{\Sigma}_z(i,i)}, \quad \text{(E.6)}$$

donde z puede ser p o p, e i indica el coeficiente i-ésimo. Así pues, introduciendo (E.4), (E.5) y (E.6) en la expresión (E.1), la distancia de Kullback Leibler adopta la forma siguiente

$$KL(p,q) = c_p log(\frac{c_p}{c_q}) + \frac{c_p}{2}\sum_i log(\frac{\mathbf{\Sigma}_q(i,i)}{\mathbf{\Sigma}_p(i,i)}) - \frac{c_p}{2}$$
$$\times \int_{\mathbf{x}} \prod_i (\frac{1}{(2\pi)^{1/2}\mathbf{\Sigma}_p(i,i)^{1/2}} e^{-1/2\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\mathbf{\Sigma}_p(i,i)}}) \sum_i (\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\mathbf{\Sigma}_p(i,i)} - \frac{(\mathbf{x}(i)-\mu_q(i))^2}{\mathbf{\Sigma}_q(i,i)}) d\mathbf{x}. \quad \text{(E.7)}$$

Dado que la integral de una Gaussiana, como toda pdf, a lo largo de su dominio es igual a la unidad, la expresión (E.7), se puede simplificar del siguiente modo

$$KL(p,q) = c_p log(\frac{c_p}{c_q}) + \frac{c_p}{2}\sum_i log(\frac{\mathbf{\Sigma}_q(i,i)}{\mathbf{\Sigma}_p(i,i)}) - \frac{c_p}{2}$$
$$\times \sum_i \int_{\mathbf{x}(i)} \frac{1}{(2\pi)^{1/2}\mathbf{\Sigma}_p(i,i)^{1/2}} e^{-1/2\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\mathbf{\Sigma}_p(i,i)}} (\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\mathbf{\Sigma}_p(i,i)} - \frac{(\mathbf{x}(i)-\mu_q(i))^2}{\mathbf{\Sigma}_q(i,i)}) d\mathbf{x}(i). \quad \text{(E.8)}$$

A continuación se trata separadamente la integral de la expresión anterior, que es ya unidimensional, y a la que, por comodidad, se la denominará A. Así pues

$$A = \int_{\mathbf{x}(i)} \frac{1}{(2\pi)^{1/2}\mathbf{\Sigma}_p(i,i)^{1/2}} e^{-1/2\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\mathbf{\Sigma}_p(i,i)}} \frac{(\mathbf{x}(i)-\mu_p(i))^2}{\mathbf{\Sigma}_p(i,i)} d\mathbf{x}(i)$$
$$- \int_{\mathbf{x}(i)} \frac{1}{(2\pi)^{1/2}\mathbf{\Sigma}_p(i,i)^{1/2}} e^{-1/2\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\mathbf{\Sigma}_p(i,i)}} \frac{(\mathbf{x}(i)-\mu_q(i))^2}{\mathbf{\Sigma}_q(i,i)} d\mathbf{x}(i) = B + C. \quad \text{(E.9)}$$

Los dos términos que componen la expresión (E.9), y que, por simplificar la notación, se nombrarán a partir de este momento como B y C, respectivente, se calculan independientemente haciendo uso de cálculo integral básico. Así, se puede observar que B, tras hacer un cambio de variable y resolver por partes es

$$B = \int_{\mathbf{x}(i)} \frac{1}{(2\pi)^{1/2}\mathbf{\Sigma}_p(i,i)^{1/2}} e^{-1/2\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\mathbf{\Sigma}_p(i,i)}} \frac{(\mathbf{x}(i)-\mu_p(i))^2}{\mathbf{\Sigma}_p(i,i)} d\mathbf{x}(i) = 1. \quad \text{(E.10)}$$

A su vez, C se descompone en tres términos, D, E y F, tal y como se indica a continuación

$$C = \int_{\mathbf{x}(i)} \frac{1}{(2\pi)^{1/2}\boldsymbol{\Sigma}_p(i,i)^{1/2}} e^{-1/2\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\boldsymbol{\Sigma}_p(i,i)}} \frac{(\mathbf{x}(i)-\mu_q(i))^2}{\boldsymbol{\Sigma}_q(i,i)} d\mathbf{x}(i) =$$
$$\frac{-1}{\boldsymbol{\Sigma}_q(i,i)}\Big(\int_{\mathbf{x}(i)} \frac{1}{(2\pi)^{1/2}\boldsymbol{\Sigma}_p(i,i)^{1/2}} e^{-1/2\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\boldsymbol{\Sigma}_p(i,i)}} \mathbf{x}(i)^2 d\mathbf{x}(i)$$
$$- \int_{\mathbf{x}(i)} \frac{1}{(2\pi)^{1/2}\boldsymbol{\Sigma}_p(i,i)^{1/2}} e^{-1/2\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\boldsymbol{\Sigma}_p(i,i)}} 2\mathbf{x}(i)\mu_q(i) d\mathbf{x}(i)$$
$$+ \int_{\mathbf{x}(i)} \frac{1}{(2\pi)^{1/2}\boldsymbol{\Sigma}_p(i,i)^{1/2}} e^{-1/2\frac{(\mathbf{x}(i)-\mu_p(i))^2}{\boldsymbol{\Sigma}_p(i,i)}} \mu_q(i)^2 d\mathbf{x}(i)\Big)$$
$$= \frac{-1}{\boldsymbol{\Sigma}_q(i,i)}(D+E+F). \tag{E.11}$$

Las variables D, E y F se calculan haciendo uso de cálculo integral básico aplicando cambios de variables y resolviendo por partes. De este modo se obtienen las siguientes expresiones finales

$$D = \boldsymbol{\Sigma}_p(i,i) + \mu_p(i), \tag{E.12}$$

$$E = -2\mu_p(i) + \mu_q(i), \tag{E.13}$$

$$F = \mu_q(i)^2, \tag{E.14}$$

Así pues, y teniendo en cuenta todo lo anterior, se obtiene finalmente la expresión para la distancia de Kullback Leibler, $KL(p,q)$, entre dos pdfs Gaussianas

$$KL(p,q) = c_p log(\frac{c_p}{c_q}) + \frac{c_p}{2}\sum_i (log(\frac{\boldsymbol{\Sigma}_q(i,i)}{\boldsymbol{\Sigma}_q(i,i)}) + \frac{\boldsymbol{\Sigma}_p(i,i)}{\boldsymbol{\Sigma}_q(i,i)} + \frac{(\mu_p(i)-\mu_q(i))^2}{\boldsymbol{\Sigma}_q(i,i)} - 1), \tag{E.15}$$

que, como se puede apreciar, coincide con (6.22) cuando los vectores de medias de las dos Gaussianas que componen las pdfs que se desean comparar son iguales.

6.9. Anexo F.

En este Anexo se presenta el desarrollo teórico necesario para estimar el vector de desplazamiento $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$ para el método PD-MEMLIN con fase de entrenamiento "ciega" haciendo uso del algoritmo EM. Dicho algoritmo se aplica iterativamente en dos pasos, E y M: el paso E, *Expectation*, estima el valor esperado de los parámetros que se pretenden estimar, mientras que el M maximiza dicho valor esperado con respecto a la variable que se desea estimar.

Se considera pues, un corpus de entrenamiento constituido por vectores de características ruidosos para cada fonema ph y entorno básico e, $\mathbf{Y}_e^{ph} = \{\mathbf{y}_1^{e,ph}; ...; \mathbf{y}_{t_{e,ph}}^{e,ph}; ...; \mathbf{y}_{T_{e,ph}}^{e,ph}\}$, con $t_{e,ph} \in [1, T_{e,ph}]$, donde el fonema correspondiente a cada vector acústico se ha determinado a partir de segmentación forzada a nivel de fonema de la señal ruidosa mediante el algoritmo de Viterbi. Apréciese que se ha eliminado con respecto a la Sección 6.4 el superíndice Tr por simplificar la notación. Por otra parte, se dispone de los modelos GMMs de los vectores de características limpios y ruidosos

para cada entorno y fonema: (6.9) (6.10), (6.11) y (6.12). Con todo ello se asume que la pdf de los vectores de características ruidosos, dado el par de Gaussianas s_x^{ph} y $s_y^{e,ph}$, el entorno básico e y el fonema ph es

$$p(\mathbf{y}_{t_{e,ph}}^{e,ph}|s_x^{ph}, s_y^{e,ph}, e, ph) = \mathcal{N}(\mathbf{y}_{t_{e,ph}}^{e,ph}; \mu_{s_x^{ph}} + \mathbf{r}_{s_x^{ph}, s_y^{e,ph}}, \mathbf{\Sigma}_{s_y^{e,ph}}), \tag{F.1}$$

de modo que se puede definir la función de log-verosimilitud para todo el corpus de entrenamiento correspondiente a un determinado entorno básico e y fonema ph, $L(\mathbf{Y}_e^{ph})_e^{ph}$, como

$$L(\mathbf{Y}_e^{ph})_e^{ph} = \sum_{t_{e,ph}} log \sum_{s_y^{e,ph}} \sum_{s_x^{ph}} p(s_x^{ph}, s_y^{e,ph}|e, ph)\mathcal{N}(\mathbf{y}_{t_{e,ph}}^{e,ph}; \mu_{s_x^{ph}} + \mathbf{r}_{s_x^{ph}, s_y^{e,ph}}, \mathbf{\Sigma}_{s_y^{e,ph}}), \tag{F.2}$$

donde $p(s_x^{ph}, s_y^{e,ph}|e, ph)$ es la probabilidad conjunta del par de Gaussianas s_x^{ph} y $s_y^{e,ph}$, dado el entorno básico, e, y el fonema ph. A continuación se realiza el paso E, para lo que se define la función auxiliar $Q(\phi, \phi_{new})_e^{ph}$, donde $\phi = \{\mathbf{r}_{s_x^{ph}, s_y^{e,ph}}\}$ se corresponde con el vector de desplazamiento del que se disponga en cada iteración, esto es, el obtenido en la iteración precedente, y $\phi_{new} = \{\mathbf{r}_{new, s_x^{ph}, s_y^{e,ph}}\}$ será el nuevo vector de desplazamiento calculado

$$Q(\phi, \phi_{new})_e^{ph} = \sum_{t_{e,ph}} \sum_{s_y^{e,ph}} \sum_{s_x^{ph}} p(s_x^{ph}, s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, \phi, e, ph) log(p(\mathbf{y}_{t_{e,ph}}^{e,ph}, s_x^{ph}, s_y^{e,ph}|\phi_{new}, e, ph)), \tag{F.3}$$

Si, por comodidad, se define la variable $\mathbf{\Omega} = \mathbf{y}_{t_{e,ph}}^{e,ph} - \mu_{s_x^{ph}} - \mathbf{r}_{new, s_x^{ph}, s_y^{e,ph}}$, y teniendo en cuenta que $p(\mathbf{y}_{t_{e,ph}}^{e,ph}, s_x^{ph}, s_y^{e,ph}|\phi_{new}, e, ph) = p(s_x^{ph}, s_y^{e,ph}) p(\mathbf{y}_{t_{e,ph}}^{e,ph}|s_x^{ph}, s_y^{e,ph}\phi_{new}, e, ph)$ la expresión (F.3) se transforma en

$$Q(\phi, \phi_{new})_e^{ph} = constant + \sum_{t_{e,ph}} \sum_{s_y^{e,ph}} \sum_{s_x^{ph}} p(s_x^{ph}, s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, \phi, e, ph)(-\frac{1}{2} log|\mathbf{\Sigma}_{s_y^{e,ph}}| - \frac{1}{2}\mathbf{\Omega}^T \mathbf{\Sigma}_{s_y^{e,ph}} \mathbf{\Omega}), \tag{F.4}$$

donde *constant* no afecta a la maximización que se realizará en el paso M. Así, y una vez finalizado el paso E, en el paso M se procede a calcular el valor de $\mathbf{r}_{new, s_x^{ph}, s_y^{e,ph}}$ derivando con respecto a dicha variable la expresión (F.4), e igualando posteriormente a cero.

$$\mathbf{r}_{new, s_x^{ph}, s_y^{e,ph}} = \frac{\delta(Q(\phi, \phi_{new})_e^{ph})}{\delta(\mathbf{r}_{new, s_x^{ph}, s_y^{e,ph}})}$$

$$= \sum_{t_{e,ph}} p(s_x^{ph}, s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, \phi, e, ph) \frac{\delta(-\frac{1}{2} log|\mathbf{\Sigma}_{s_y^{e,ph}}| - \frac{1}{2}\mathbf{\Omega}^T \mathbf{\Sigma}_{s_y^{e,ph}} \mathbf{\Omega})}{\delta(\mathbf{r}_{new, s_x^{ph}, s_y^{e,ph}})} = \mathbf{0}, \tag{F.5}$$

$$\frac{\delta(-\frac{1}{2} log|\mathbf{\Sigma}_{s_y^{e,ph}}| - \frac{1}{2}\mathbf{\Omega}^T \mathbf{\Sigma}_{s_y^{e,ph}} \mathbf{\Omega})}{\delta(\mathbf{r}_{new, s_x^{ph}, s_y^{e,ph}})} = \mathbf{\Sigma}_{s_y^{e,ph}}(\mathbf{y}_{t_{e,ph}}^{e,ph} - \mu_{s_x^{ph}} - \mathbf{r}_{new, s_x^{ph}, s_y^{e,ph}}), \tag{F.6}$$

$$\mathbf{r}_{new, s_x^{ph}, s_y^{e,ph}} = \frac{\sum_{t_{e,ph}} p(s_x^{ph}, s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, \phi, e, ph)(\mathbf{y}_{t_{e,ph}}^{e,ph} - \mu_{s_x^{ph}})}{\sum_{t_{e,ph}} p(s_x^{ph}, s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, \phi, e, ph)}, \tag{F.7}$$

donde $p(s_x^{ph}, s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, \phi, e, ph)$ se puede obtener mediante la siguiente aproximación

$$p(s_x^{ph}, s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, \phi, e, ph) \simeq p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, e, ph)p(s_x^{ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, s_y^{e,ph}, \phi, e, ph)$$
$$= p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph}, e, ph)\frac{p(s_y^{e,ph})p(\mathbf{y}_{t_{e,ph}}^{e,ph}|s_x^{ph}, s_y^{e,ph}, \phi)}{\sum_{s_x^{ph}} p(s_y^{e,ph})p(\mathbf{y}_{t_{e,ph}}^{e,ph}|s_x^{ph}, s_y^{e,ph}, \phi)}, \tag{F.8}$$

que coincide con la expresión presentada en la Sección 6.4.

6.10. Anexo G.

7

Mejoras en el Modelado de la Probabilidad entre Gaussianas.

En el Capítulo 5 se plantearon distintas líneas de actuación para compensar algunas de las limitaciones observadas en la técnica MEMLIN. Una de ellas, mejorar el modelo de $\mathbf{x}$, ya se ha tratado convenientemente en el Capítulo 6; ahora le toca el turno al término correspondiente a la probabilidad entre Gaussianas, $p(s_x|\mathbf{y}_t, e, s_y^e)$, término este de gran importancia puesto que determina, a nivel de Gaussiana, el entorno de proyección del vector de características ruidoso dentro del espacio limpio y, por tanto, el nivel de incertidumbre en el que se puede mover el vector de características normalizado, que estará en función de las varianzas de las Gaussianas que modelan el espacio limpio. Hasta el momento, y en los distintos métodos presentados en este trabajo, $p(s_x|\mathbf{y}_t, e, s_y^e)$ se ha aproximado siempre por $p(s_x|\mathbf{x}_t)$, eliminando por tanto la dependencia con el correspondiente vector de características ruidoso. Esto no deja de ser una aproximación de compromiso ya que supone que el vector de características limpio asociado al correspondiente degradado ha sido generado por una Gaussiana, s_x, que únicamente depende de la Gaussiana s_y, menospreciando, en cierto modo, la capacidad degradativa de la aleatoriedad introducida por el ruido del entorno acústico concreto.

Para compensar las limitaciones mostradas por el modelado de la probabilidad entre Gaussianas apreciado en las técnicas presentadas hasta el momento, se propone modelar mediante una GMM los vectores de características ruidosos asociados a cada par de Gaussianas: s_x y s_y^e, en el caso de los métodos MEMLIN, P-MEMLIN o MEMHIN [Buera *et al.*, 2006b], o s_x^{ph} y $s_y^{e,ph}$ si se trata del algoritmo PD-MEMLIN [Buera *et al.*, 2006a]. Estas GMMs se entrenan en la fase de entrenamiento previa mediante señal estéreo y, del mismo modo que proporcionan una interesante mejora en los resultados de RAH, también son responsables de incrementar el coste computacional considerablemente. Sin embargo, tal y como se indicará a lo largo de este Capítulo, esta limitación puede ser minimizada si se reduce el número de pares de Gaussianas computadas.

En este Capítulo se presenta primeramente (Sección 7.1) un estudio sobre los efectos, tanto cualitativos como cuantitativos, que el término de probabilidad entre Gaussianas posee en las técnicas MEMLIN y PD-MEMLIN. A raíz de los resultados presentados no sólo se podrá afirmar que el término en cuestión tiene una importancia capital, sino que además el margen de mejora en términos de WER al que se puede aspirar es muy importante. Una vez demostradas las limitaciones de la aproximación del modelo de probabilidad entre Gaussianas aplicada hasta el momento en las técnicas MEMLIN y PD-MEMLIN, se procede a exponer la solución propuesta en este sentido: en la Sección 7.2 se incluye el desarrollo teórico general precisado para obtener las distintas GMMs que representan a los vectores de características ruidosos asociados a cada par de Gaussianas. La

adaptación de las expresiones calculadas en la Sección anterior para las técnicas MEMLIN y PD-MEMLIN se exponen en la Sección 7.3. Los resultados de RAH obtenidos tras la aplicación del nuevo modelado de la probabilidad entre Gaussianas basado en GMMs para las técnicas MEMLIN y PD-MEMLIN se obtuvieron, una vez más, con la base de datos *SpeechDat Car* en español, y se incluyen en la Sección 7.4. En ella queda patente el buen comportamiento de las dos extensiones propuestas, no sólo con respecto a los métodos empíricos basados en el criterio MMSE más utilizados en la actualidad (CMN, RATZ y SPLICE), sino también si se comparan con la técnicas MEMLIN y PD-MEMLIN.

7.1. El Efecto del Modelado de la Probabilidad entre Gaussianas.

Desde un primer momento ya se pensó que el modelado de la probabilidad entre Gaussianas en las técnicas de normalización de vectores de características presentadas en este trabajo hasta este momento (MEMLIN, P-MEMLIN, MEMHIN y PD-MEMLIN) podía desempeñar un papel capital. En realidad, conceptualmente hablando, este término tiene la capacidad de determinar, a nivel de Gaussiana, el entorno de proyección del vector de características ruidoso dentro del espacio limpio y, por tanto, el nivel de incertidumbre en el que se puede mover el vector de características normalizado, que estará en función de las varianzas de las Gaussianas que modelan el espacio limpio. Esta suposición, sin embargo, sólo proporcionaba una cierta idea cualitativa de la importancia del modelado de la probabilidad entre Gaussianas. Para certificarla, y además determinar cuantitativamente cual impornte es dicho término, se realizaron sendos experimentos de RAH para las técnicas MEMLIN y PD-MEMLIN.

Para ello se modificaron ambas técnicas de modo que el término de estudio, el modelo de la probabilidad entre Gaussianas, se calculara a partir de la señal de reconocimiento limpia, esto es, $p(s_x|\mathbf{y}_t, e, s_y^e) \simeq p(s_x|\mathbf{x}_t)$, para el caso del algoritmo MEMLIN y $p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph}) \simeq p(s_x^{ph}|\mathbf{x}_t)$, si se trata del método PD-MEMLIN. De esta manera, el cálculo de las nuevas variables se realizará haciendo uso de (5.7) y (5.8) o de (6.11) y (6.12), respectivamente.

$$p(s_x|\mathbf{x}_t) = \frac{p(s_x)p(\mathbf{x}_t|s_x)}{\sum_{s_x} p(s_x)p(\mathbf{x}_t|s_x)}, \tag{7.1}$$

$$p(s_x^{ph}|\mathbf{x}_t) = \frac{p(s_x^{ph})p(\mathbf{x}_t|s_x^{ph})}{\sum_{s_x^{ph}} p(s_x^{ph})p(\mathbf{x}_t|s_x^{ph})}. \tag{7.2}$$

Con este experimento, tanto más alejado de la realidad conforme mayor sea el número de Gaussianas con que se modele el espacio limpio, se pretende conocer el límite de los métodos MEMLIN y PD-MEMLIN al que se puede aspirar cuando el modelado de la probabilidad entre Gaussianas es óptimo. Así pues los resultados de RAH para las dos técnicas modificadas se pueden observar en la Tabla 7.1, donde se han incluido además, a modo de comparación, los resultados ya obtenidos cuando se reconoce la señal limpia (Entrenamiento CLK, Reconocimiento CLK) y ruidosa (Entrenamiento CLK, Reconocimiento HF). Asimismo se introducen los resultados de WER medio (MWER) y mejora media de WER (MIMP). En la experimentación se ha utilizado la base de datos *SpeechDat Car* en español, *parametrización UZ* y modelos acústicos para unidades fonéticas. Por otra parte, para el caso de la modificación sobre la técnica MEMLIN se han empleado 128 Gaussianas para modelar el espacio limpio y cada entorno básico; sin embargo, para el algoritmo modificado PD-MEMLIN se han utilizado 16 Gaussianas para componer los distintos modelos

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	CLK	1.90	2.64	1.81	1.75	1.62	0.64	0.35	1.75	
CLK	HF	5.91	14.49	14.55	20.17	21.07	16.19	35.71	16.21	
CLK	MEMLIN 128	1.63	3.52	1.82	1.50	2.29	0.79	0.35	1.99	98.36
CLK	PD-MEMLIN 16	1.25	3.78	2.66	2.13	3.53	1.27	1.36	2.50	94.83

Cuadro 7.1: Resultados en términos de WER (%) con la base de datos *SpeechDat Car* para las técnicas MEMLIN y PD-MEMLIN cuando se emplea señal limpia para determinar el modelado de la probabilidad entre Gaussianas. Se presentan los resultados para los diferentes entornos básicos (E1,..., E7) utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia (CLK en la columna de Entrenamiento). "Reconocimiento" hace referencia a la señal empleada para RAH: limpia (CLK), ruidosa (HF) y normalizada tras aplicar las técnicas modificadas MEMLIN y PD-MEMLIN. En estas dos últimas se incluye, junto al nombre identificador de los métodos, el número de Gaussianas empleadas para modelar, bien los espacios (limpio y los entornos básicos), bien los distintos fonemas.

GMM de cada fonema. Nótese que la experimentación propuesta posee los mismos parámetros que los empleados en las Secciones 5.4 y 6.5, por lo que los resultados son totalmente comparables.

Tal y como se puede apreciar en la Tabla 7.1, el margen de mejora que puede proporcionar el término de modelado de la probabilidad entre Gaussianas es muy elevado, tanto que permite acercarse a la normalización perfecta (100 % de MIMP). De este modo se certifica, ya de un modo cuantitativo, la suposición que sobre la importancia de este término ya se tenía desde un primer momento. Por otra parte, el hecho de que los resultados obtenidos con la variación de la técnica MEMLIN sean algo superiores a los alcanzados con la modificación del algoritmo PD-MEMLIN se debe a que en este último caso hay otro término que influye de un modo importante y que en este caso no se está optimizando: la probabilidad a posteriori del fonema ph, dado el vector de características ruidoso $\mathbf{y}_t$ y el entorno básico e, $p(ph|\mathbf{y}_t, e)$, tal y como ya quedó reflejado en la Sección 6.5 al presentarse la pseudo-técnica KPD-MEMLIN.

Por otra parte, también se presentan en la Figura 7.1 los histogramas y *log-scattegrams* obtenidos a partir del primer coeficiente MFCC de los vectores de características de voz de la señal limpia y la normalizada mediante las dos extensiones propuestas anteriormente para los algoritmos MEMLIN y PD-MEMLIN. Dichas representaciones se han obtenido a partir de las señales del corpus de reconocimiento del entorno básico E4 de la base de datos *SpeechDat Car* en español. Asimismo, en las Figuras 7.1.a, y a modo de comparación, se vuelven a incluir las representaciones correspondientes obtenidas a partir de la señal limpia y la ruidosa, pudiéndose apreciar en este caso el efecto, tanto en términos de pdf (Figura 7.1.a.1) como de incertidumbre (Figura 7.1.a.2), que el entorno acústico produce en los coeficientes de la señal limpia. A su vez en las Figuras 7.1.b se presentan las gráficas obtenidas tras aplicar la técnica MEMLIN modificada, incluyendo los histogramas de la señal limpia y la normalizada (Figura 7.1.b.1) y el correspondiente *log-scattegram* (Figura 7.1.b.2). Si se comparan estas representaciones con las obtenidas con la técnica MEMLIN convencional (Figuras 5.7.b), se puede constatar una mejor aproximación del histograma normalizado con respecto al de la señal limpia, así como una importante reducción de la incertidumbre. Por último, en las Figuras 7.1.c se incluyen las gráficas obtenidas con la modificación de la técnica PD-MEMLIN; se incluyen tanto los histogramas de la señal limpia y la normalizada (Figura 7.1.c.1), como el correspondiente *log-scattegram* (Figura 7.1.c.2). Si se comparan estas representaciones con las obtenidas con la técnica PD-MEMLIN convencional (Figuras 6.6.b), se puede observar una mejor aproximación del histograma normalizado con respecto al de la señal limpia, así como una

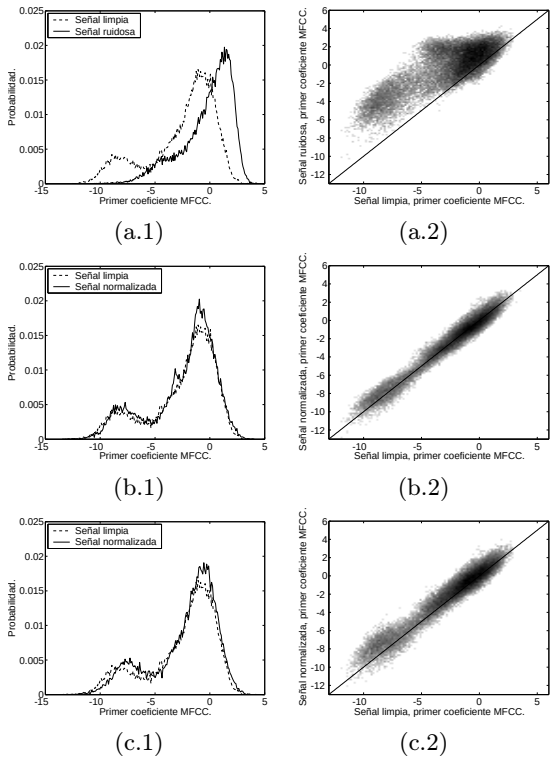


Figura 7.1: *Log-scattegrams* e histogramas realizados entre el primer coeficiente MFCC de las tramas de voz de la señal limpia (eje de abscisas) y la señal ruidosa (a) (eje de ordenadas), o normalizada usando la técnica MEMLIN con 128 Gaussianas por entorno básico (b) (eje de ordenadas) cuando se emplea señal limpia para calcular el modelado de la probabilidad entre Gaussianas. En la figura (c) se representa el *log-scattegram* y el histograma obtenidos a partir de la señal normalizada con la técnica PD-MEMLIN con 16 Gaussianas por fonema y entorno básico cuando se emplea señal limpia para calcular el modelado de la probabilidad entre Gaussianas. Todas las representaciones se realizaron a partir del corpus de reconocimiento del entorno básico E4 de la base de datos *SpeechDat Car* en español. La línea en los *log-scattegrams* representa la función $x = y$.

importante reducción de la incertidumbre.

Con todo lo anterior se puede concluir, ya definitivamente, que la potencialidad de las técnicas MEMLIN y PD-MEMLIN es tal, que podría llegar a alcanzar resultados de RAH con la señal normalizada propios de la señal limpia, hecho este que no siempre es posible decir de otras técnicas de normalización. Sin embargo, todo esto pasa por mejorar considerablemente el modelado de la probabilidad entre Gaussianas. A continuación se presenta la solución propuesta, que consiste en modelar los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y^e para la técnica MEMLIN, o s_x^{ph} y $s_y^{e,ph}$ si se trata del algoritmo PD-MEMLIN, mediante GMMs.

7.2. Modelado de la Probabilidad entre Gaussianas basado en GMMs.

Tal y como se ha comentado anteriormente, para mejorar el modelado de la probabilidad entre Gaussianas, hasta ahora aproximado siempre mediante una expresión independiente del vector de

características ruidoso, se propone hacer uso de GMMs, de modo que representen a dichos vectores de características degradados asociados a cada par de Gaussianas de los modelos de los entornos básicos y el espacio limpio (s_x y s_y^e para las técnicas MEMLIN, P-MEMLIN y MEMHIN o $s_x^{e,ph}$ y $s_y^{e,ph}$ para el caso del algoritmo PD-MEMLIN). A continuación, y de cara a simplificar la notación del desarrollo teórico que se va a exponer para estimar los parámetros que definen las correspondientes GMMs asociadas al modelado de la probabilidad entre Gaussianas, se considerará únicamente un entorno básico y un fonema, de modo que se eliminarán esas dependencias. Esto no resta generalidad alguna puesto que cada entorno básico y fonema se pueden tratar independientemente considerando que, tal y como se ha venido haciendo en capítulos precedentes (Secciones 6.3 y 5.3), cada vector de características del corpus de entrenamiento se puede etiquetar como perteneciente a un entorno básico y fonema concretos.

Sea pues la GMM que modela los vectores de características ruidosos asociada al par de Gaussianas de los espacios limpio y degradado s_x y s_y, y compuesta por C''' componentes (se considera que los vectores de características degradados asociados a los distintos pares de Gaussianas se representan con el mismo número de componentes)

$$p(\mathbf{y}_t|s_x, s_y) = \sum_{s'_y=1}^{C'''} p(\mathbf{y}_t|s_x, s_y, s'_y)p(s'_y|s_x, s_y), \tag{7.3}$$

$$p(\mathbf{y}_t|s_x, s_y, s'_y) = \mathcal{N}(\mathbf{y}_t; \mu_{s_x,s_y,s'_y}, \mathbf{\Sigma}_{s_x,s_y,s'_y}), \tag{7.4}$$

donde μ_{s_x,s_y,s'_y}, $\mathbf{\Sigma}_{s_x,s_y,s'_y}$, y $p(s'_y|s_x, s_y)$ son el vector de medias, la matriz diagonal de covarianzas y la probabilidad a priori asociados a la componente s'_y del modelo de probabilidad propuesto para el par de Gaussianas s_x y s_y. Para obtener las estimaciones de estos tres parámetros se hace uso del criterio de máxima verosimitud, ML; para lo cual se define previamente una función de verosimitud a partir de los parámetros dadas las observaciones, que en este caso se corresponden con los vectores de características ruidosos, y posteriormente se maximiza dicha función con respecto a cada uno de los tres parámetros que definen la GMM. Este proceso, dado que no tiene generalmente una solución directa sencilla, se suele llevar a cabo a partir del uso del algoritmo EM [Dempster *et al.*, 1977] tal y como se indica a continuación.

Dado un corpus de entrenamiento compuesto por señal estéreo $(\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}_1, \mathbf{y}_1), ...(\mathbf{x}_t, \mathbf{y}_t)..., (\mathbf{x}_T, \mathbf{y}_T)\}$ con $t \in [1, T]$, y donde se han eliminado las dependencias del entorno básico y los fonemas tal y como se ha explicado con anterioridad, del mismo modo que el superíndice Tr, que aparecía en otras secciones para denotar la pertenencia de los distintos vectores de características al corpus de entrenamiento, y que en esta ocasión no aparece por simplificar la notación. Por otra parte, se asume, además del modelo de probabilidad entre Gaussianas para los vectores de características ruidosos y cuyas variables se pretenden estimar, (7.3) y (7.4), que los vectores de características ruidosos se pueden modelar mediante una GMM de C' componentes, identificadas como s_y, (5.13) y (5.14), así como que los vectores de características limpios quedan representados mediante una GMM de C componentes, s_x, (5.7) y (5.8). Nótese que estas dos últimas aproximaciones ya se han tenido en cuenta en las distintas técnicas de normalización previamente presentadas. Para finalizar, y por completar académicamente el problema, también se considerará una GMM de C'' componentes, identificadas como s'_x, y que modela la probabilidad entre Gaussianas para los vectores de características limpios.

Con todo lo anterior, cada $\mathbf{y}_t$ se puede ver como un vector de características etiquetado de modo incompleto (*missing* o *incomplete data*), que, para completarlo (*complete data*), son necesarios dos vectores indicadores, a saber, $\mathbf{w}_t \in \{0, 1\}^{C'}$, que poseerá un uno en la posición correspondiente a

la Gaussiana s_y que ha generado $\mathbf{y}_t$ y ceros en el resto de las C' posiciones ($\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_T\}$); por su parte, el segundo vector indicador es $\mathbf{z}_{y,t} \in \{0,1\}^{C'''}$, que estará compuesto por un uno en la posición correspondiente a la Gaussiana s'_y del modelo de probabilidad entre Gaussianas que genera $\mathbf{y}_t$ y ceros en el resto de las C''' posiciones ($\mathbf{Z}_y = \{\mathbf{z}_{y,1}, ..., \mathbf{z}_{y,T}\}$). Asimismo, cada vector de características limpio $\mathbf{x}_t$ se puede ver igualente como un vector etiquetado de modo incompleto que precisaría de dos nuevos vectores indicadores para completarlo; el primero de ellos es en este caso $\mathbf{v}_t \in \{0,1\}^C$, que incluiría un uno en aquella posición correspondiente a la Gaussiana s_x que ha generado $\mathbf{x}_t$ y ceros en el resto de las C posiciones ($\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_T\}$), mientras que el segundo vector indicador sería $\mathbf{z}_{x,t} \in \{0,1\}^{C''}$, que estaría compuesto por un uno en la posición correspondiente a la Gaussiana s'_x del modelo de probabilidad entre Gaussianas que genera $\mathbf{x}_t$ y ceros en el resto de las C'' posiciones ($\mathbf{Z}_x = \{\mathbf{z}_{x,1}, ..., \mathbf{z}_{x,T}\}$). Así pues, la pdf de los datos completos se puede ver como

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{z}_x, \mathbf{z}_y) \quad &\simeq \quad p(\mathbf{v}, \mathbf{w}, \mathbf{z}_x) p(\mathbf{x}|\mathbf{v}, \mathbf{w}, \mathbf{z}_x) \times \\ &\quad p(\mathbf{v}, \mathbf{w}, \mathbf{z}_y) p(\mathbf{y}|\mathbf{v}, \mathbf{w}, \mathbf{z}_y), \end{aligned} \tag{7.5}$$

donde se ha supuesto que $\mathbf{x}$ e $\mathbf{y}$ son independientes, del mismo modo que $\mathbf{x}$ y $\mathbf{z}_y$, e $\mathbf{y}$ y $\mathbf{z}_x$. Dado que los cuatro vectores indicadores considerados ($\mathbf{v}$, $\mathbf{w}$, $\mathbf{z}_x$ y $\mathbf{z}_y$) se corresponden con multinomiales, la pdf de los datos completos (7.5) se puede expresar como (7.6), donde v_{s_x}, w_{s_y}, z_{y,s'_y} y z_{x,s'_x} son las componentes de los vectores $\mathbf{v}$, $\mathbf{w}$, $\mathbf{z}_x$ y $\mathbf{z}_y$ asociadas a las Gaussianas s_x, s_y, s'_y y s'_x respectivamente.

$$\begin{gathered} p(\mathbf{x}, \mathbf{y}, \mathbf{v}, \mathbf{w}, \mathbf{z}_x, \mathbf{z}_y) \simeq \\ \prod_{s_x} \prod_{s_y} \prod_{s'_x} [p(v_{s_x} = 1, w_{s_y} = 1, z_{x,s'_x} = 1) p(\mathbf{x}|v_{s_x} = 1, w_{s_y} = 1, z_{x,s'_x} = 1)]^{v_{s_x} w_{s_y} z_{x,s'_x}} \times \\ \prod_{s_x} \prod_{s_y} \prod_{s'_y} [p(v_{s_x} = 1, w_{s_y} = 1, z_{y,s'_y} = 1) p(\mathbf{y}|v_{s_x} = 1, w_{s_y} = 1, z_{y,s'_y} = 1)]^{v_{s_x} w_{s_y} z_{y,s'_y}}. \end{gathered} \tag{7.6}$$

El algoritmo EM, tal y como ya se ha indicado en alguna ocasión, es una técnica iterativa que consta de dos pasos. En el primero de ellos, llamado E, *Expectation*, se estima el valor esperado de la función de log-verosimilitud de los datos completos con respecto a los vectores indicadores y dados los datos incompletos. Por otra parte, en el segundo paso, o M, *Maximization*, se obtienen los parámetros deseados, en este caso los que definen el modelado de la probabilidad entre Gaussianas. Para esto último se maximiza con respecto a dichos parámetros la expresión del valor esperado obtenido en el paso E. Ambos procesos se repiten tantas veces como sea necesario, pudiéndose demostrar que en cada iteración se incrementa la log-verosimilitud de los datos a la vez que se garantiza la convergencia del método a un máximo local de la función de log-verosimilitud [Dempster *et al.*, 1977] [Wu, 1983]. A continuación se trata cada uno de los pasos por separado.

7.2.1. El paso E.

A la hora de evaluar el paso E, se define inicialmente la función de log-verosimilitud considerando los datos completos, esto es, y para este caso concreto, los vectores de características estéreos, limpios $\mathbf{X}$ y ruidosos $\mathbf{Y}$, y los vectores indicadores: $\mathbf{V}$, $\mathbf{W}$, $\mathbf{Z}_x$ y $\mathbf{Z}_y$, $\mathcal{L}(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y}, \mathbf{V}, \mathbf{W}, \mathbf{Z}_x, \mathbf{Z}_y)$, donde $\mathbf{\Theta}$ incluye todos los parámetros de la GMM de la probabilidad entre Gaussianas que se pretenden estimar ($p(\mathbf{y}_t|s_x, s_y, s'_y)$, μ_{s_x,s_y,s'_y} y $\mathbf{\Sigma}_{s_x,s_y,s'_y}$).

$$\begin{gathered} \mathcal{L}(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y}, \mathbf{V}, \mathbf{W}, \mathbf{Z}_x, \mathbf{Z}_y) = \textstyle\sum_t log(p(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t, \mathbf{w}_t, \mathbf{z}_{x,t}, \mathbf{z}_{y,t}|\mathbf{\Theta})) = \\ \textstyle\sum_t \sum_{s_x} \sum_{s_y} \sum_{s'_x} v_{s_x} w_{s_y} z_{x,s'_x} [log(p(v_{s_x} = 1, w_{s_y} = 1, z_{x,s'_x} = 1)) + log(p(\mathbf{x}_t|v_{s_x} = 1, w_{s_y} = 1, z_{x,s'_x} = 1))] + \\ \textstyle\sum_t \sum_{s_x} \sum_{s_y} \sum_{s'_y} v_{s_x} w_{s_y} z_{y,s'_y} [log(p(v_{s_x} = 1, w_{s_y} = 1, z_{y,s'_y} = 1)) + log(p(\mathbf{y}_t|v_{s_x} = 1, w_{s_y} = 1, z_{y,s'_y} = 1))], \end{gathered} \tag{7.7}$$

donde, si se considera que v_{s_x} y w_{s_y} son independientes, se tiene que

$$\begin{gathered} p(v_{s_x}=1, w_{s_y}=1, z_{x,s'_x}=1) \simeq \\ p(v_{s_x}=1)p(w_{s_y}=1)p(z_{x,s'_x}=1|v_{s_x}=1, w_{s_y}=1) = P_{s_x}P_{s_y}P_{s_x s_y s'_x}, \end{gathered} \tag{7.8}$$

$$\begin{gathered} p(v_{s_x}=1, w_{s_y}=1, z_{y,s'_y}=1) \simeq \\ p(v_{s_x}=1)p(w_{s_y}=1)p(z_{y,s'_y}=1|v_{s_x}=1, w_{s_y}=1) = P_{s_x}P_{s_y}P_{s_x s_y s'_y}, \end{gathered} \tag{7.9}$$

siendo P_{s_x} y P_{s_y} las probabilidades a priori de las componentes s_x y s_y, respectivamente; mientras que $P_{s_x s_y s'_x}$ y $P_{s_x s_y s'_y}$ son, por su parte, las probabilidades a priori de las Gaussianas s'_x y s'_y, de las GMMs asociadas al par s_x y s_y, respectivamente. El problema, llegado a este punto, es determinar las Gaussianas de los distintos modelos que generan los datos incompletos; para ello se consideran como constantes los parámetros del modelado de probabilidad entre Gaussianas para la iteración k-ésima, $\mathbf{\Theta}^{(k)}$, siendo k la iteración previa. Asimismo se hace uso de la función $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$, que está relacionada con la función de log-verosimilitud, y que se define del siguiente modo $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)}) = E[log(p(\mathbf{X}, \mathbf{Y}, \mathbf{V}, \mathbf{W}, \mathbf{Z}_x, \mathbf{Z}_y|\mathbf{\Theta}))|\mathbf{X}, \mathbf{Y}, \mathbf{\Theta}^{(k)}]$, donde el operador $E[\bullet]$ representa el valor esperado de $\bullet$. Considerando esto último se puede observar que

$$\begin{gathered} Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)}) = \\ \sum_t \sum_{s_x} \sum_{s_y} \sum_{s'_x} (v_{s_x} w_{s_y} z_{x s'_x})^{(k)} [log(P_{s_x} P_{s_y} P_{s_x s_y s'_x}) + log(p(\mathbf{x}_t|v_{s_x}=1, w_{s_y}=1, z_{x s'_x}=1))] + \\ \sum_t \sum_{s_x} \sum_{s_y} \sum_{s'_y} (v_{s_x} w_{s_y} z_{s'_y})^{(k)} [log(P_{s_x} P_{s_y} P_{s_x s_y s'_y}) + log(p(\mathbf{y}_t|v_{s_x}=1, w_{s_y}=1, z_{y s'_y}=1))]. \end{gathered} \tag{7.10}$$

$$\begin{gathered} (v_{s_x} w_{s_y} z_{x s'_x})^{(k)} = E[v_{s_x} w_{s_y} z_{x s'_x}|\mathbf{x}_t, \mathbf{y}_t, \mathbf{\Theta}^{(k)}] \simeq \\ E[v_{s_x}|\mathbf{x}_t] E[w_{s_y}|\mathbf{y}_t] E[z_{x s'_x}|\mathbf{x}_t, v_{s_x}, w_{s_y}, \mathbf{\Theta}^{(k)}] = AB^{(k)}, \end{gathered} \tag{7.11}$$

$$\begin{gathered} (v_{s_x} w_{s_y} z_{y s'_y})^{(k)} = E[v_{s_x} w_{s_y} z_{y s'_y}|\mathbf{x}_t, \mathbf{y}_t, \mathbf{\Theta}^{(k)}] \simeq \\ E[v_{s_x}|\mathbf{x}_t] E[w_{s_y}|\mathbf{y}_t] E[z_{y s'_y}|\mathbf{y}_t, v_{s_x}, w_{s_y}, \mathbf{\Theta}^{(k)}] = AC^{(k)}, \end{gathered} \tag{7.12}$$

donde se ha considerado que v_{s_x} y w_{s_y} son independientes, del mismo modo que v_{s_x} e $\mathbf{y}_t$, w_{s_y} y $\mathbf{x}_t$, y $z_{x s_x}$ e $\mathbf{y}_t$. Se asume igualmente que la esperanza de las Gaussianas v_{s_x} y w_{s_y}, dados los vectores de características $\mathbf{x}_t$ e $\mathbf{y}_t$ no dependen del modelo de probabilidad entre Gaussianas propuesto. Por otra parte, $E[z_{s'_y}|\mathbf{y}_t, v_{s_x}, w_{s_y}, \mathbf{\Theta}^{(k)}]$ se estima haciendo uso de las expresiones (7.3) y (7.4), dando lugar a (7.13); mientras, $E[v_{s_x}|\mathbf{x}_t]$ y $E[w_{s_y}|\mathbf{y}_t]$, se pueden obtener a partir de las GMMs que representan tanto el espacio limpio (5.7) y (5.8), como el ruidoso (5.13) y (5.14), respectivamente, de manera similar a la que se ha obtenido (7.13); sin embargo, en este trabajo a la hora de estimar estas dos variables se ha adoptado una decisión *hard*, esto es, tomarán el valor 1 si las Gaussianas s_x o s_y son respectivamente las más probables, ó 0 en cualquier otro caso. Por último, $E[z_{s'_x}|\mathbf{x}_t, v_{s_x}, w_{s_y}, \mathbf{\Theta}^{(k)}]$ se podría estimar del mismo modo que $E[z_{s'_y}|\mathbf{y}_t, v_{s_x}, w_{s_y}, \mathbf{\Theta}^{(k)}]$ para el hipotético modelado de la probabilidad entre Gaussianas para los vectores de características limpios, pero no es necesario hacerlo ya que para el desarrollo teórico que se pretende resulta intrascendente.

$$E[z_{s'_y}|\mathbf{y}_t, v_{s_x}, w_{s_y}, \mathbf{\Theta}^{(k)}] = \frac{p(s'_y|s_x, s_y)^{(k)} \mathcal{N}(\mathbf{y}_t|\mu^{(k)}_{s_x,s_y,s'_y}, \mathbf{\Sigma}^{(k)}_{s_x,s_y,s'_y})}{\sum_{s'_y} p(s'_y|s_x, s_y)^{(k)} \mathcal{N}(\mathbf{y}_t|\mu^{(k)}_{s_x,s_y,s'_y}, \mathbf{\Sigma}^{(k)}_{s_x,s_y,s'_y})}. \tag{7.13}$$

7.2.2. El paso M.

Para obtener las estimaciones de máxima verosimilitud para los distintos parámetros que definen el modelo de la probabilidad entre Gaussianas considerado, se maximiza, como ya se había

indicado con anterioridad, la función $Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$ con respecto a ellos, dando lugar de ese modo a las correspondientes expresiones para la iteración $(k+1)$.

Estimación de la probabilidad a priori de la Gaussiana s'_y del modelado de la probabilidad entre Gaussianas.

Para realizar la maximización de la probabilidad a priori de la Gaussiana s'_y del modelado de la probabilidad entre las Gaussianas s_x y s_y, se debe tener en cuenta la restricción de que las probabilidades a priori han de sumar la unidad, por lo que se hace necesario introducir el multiplicador de Lagrange $\lambda_{s_x s_y}$. Así pues, la función que se debe maximizar en este caso es

$$\mathcal{L}(\mathbf{\Theta}, \lambda_{s_x s_y}) = Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)}) - \sum_{s_x}\sum_{s_y}\lambda_{s_x s_y}\sum_{s'_y}[P_{s_x s_y s'_y} - 1], \tag{7.14}$$

donde $\lambda_{s_x s_y}$ son los correspondientes multiplicadore de Lagrange. De este modo, a la hora de obtener las probabilidades a priori óptimas, $P_{s_x s_y s'_y}$, es preciso maximizar la función $\mathcal{L}(\mathbf{\Theta}, \lambda_{s_x s_y})$ con respecto a dichas probabilidades y los multiplicadores de Lagrange. Para el primero de los casos se tiene

$$\frac{\delta\mathcal{L}(\mathbf{\Theta}, \lambda_{s_x s_y})}{\delta P_{s_x s_y s'_y}}|_{\Theta=\Theta^{k+1}} = \sum_t \frac{(v_{s_x} w_{s_y} z_{y s'_y})^{(k)}}{P^{k+1}_{s_x s_y s'_y}} - \lambda_{s_x s_y} = 0, \tag{7.15}$$

$$P^{(k+1)}_{s_x s_y s'_y} = \frac{1}{\lambda_{s_x s_y}}\sum_t (v_{s_x} w_{s_y} z_{y s'_y})^{(k)}. \tag{7.16}$$

Si ahora se maximiza la función $\mathcal{L}(\mathbf{\Theta}, \lambda_{s_x s_y})$ con respecto a los multiplicadores de Lagrange se obtienen las siguientes expresiones

$$\frac{\delta\mathcal{L}(\mathbf{\Theta}, \lambda_{s_x s_y})}{\delta\lambda_{s_x s_y}}|_{\Theta=\Theta^{k+1}} = -\sum_{s'_y} P^{(k+1)}_{s_x s_y s'_y} + 1 = 0, \tag{7.17}$$

$$\sum_{s'_y} P^{(k+1)}_{s_x s_y s'_y} = 1. \tag{7.18}$$

A partir de las expresiones anteriores (7.16) y (7.18), se puede obtener la estimación final para la probabilidad a priori de la Gaussiana s'_y del modelo de probabilidad entre Gaussianas s_x y s_y, que será

$$p(s'_y|s_x, s_y)^{(k+1)} = \frac{\sum_t (v_{s_x} w_{s_y} z_{s'_y})^{(k)}}{\sum_t \sum_{s'_y} (v_{s_x} w_{s_y} z_{s'_y})^{(k)}} = \frac{\sum_t AC^{(k)}}{\sum_t \sum_{s'_y} AC^{(k)}}, \tag{7.19}$$

donde debe recordarse que $(v_{s_x} w_{s_y} z_{y s'_y})^{(k)} = AC^{(k)}$.

Estimación del vector de medias de la Gaussiana s'_y del modelado de la probabilidad entre Gaussianas.

Para estimar los vectores de medias de la Gaussiana s'_y del modelado de la probabilidad entre las Gaussianas s_x y s_y, se deberá maximizar con respecto a dicho vector de medias, μ_{s_x, s_y, s'_y}, la función siguiente $\mathcal{L}(\mathbf{\Theta}) = Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$. Teniendo en cuenta que dicho vector de medias aparece únicamente en el segundo término aditivo de (7.10), se puede observar que la maximización requerida es

$$\begin{array}{c}\frac{\delta \mathcal{Q}(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})}{\delta \mu_{s_x s_y s'_y}}|_{\Theta=\Theta^{k+1}} = \mathbf{0} = \\ \sum_t (v_{s_x} w_{s_y} z_{y s'_y})^{(k)} \frac{\delta}{\delta \mu_{s_x s_y s'_y}}|_{\Theta=\Theta^{k+1}}[log(p(\mathbf{y}_t|v_{s_x}=1, w_{s_y}=1, z_{s'_y}=1))] = \\ \sum_t (v_{s_x} w_{s_y} z_{y s'_y})^{(k)} \frac{\delta}{\delta \mu_{s_x s_y s'_y}}|_{\Theta=\Theta^{k+1}}[-\frac{1}{2}(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}(\mathbf{y}_t - \mu_{s_x s_y s'_y})],\end{array} \tag{7.20}$$

donde se ha hecho uso de

$$p(\mathbf{y}_t|v_{s_x}=1, w_{s_y}=1, z_{y s'_y}=1) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_{s_x,s_y,s'_y}|^{1/2}} e^{-\frac{1}{2}(\mathbf{y}_t - \mu_{s_x,s_y,s'_y})^T \mathbf{\Sigma}^{-1}_{s_x,s_y,s'_y}(\mathbf{y}_t - \mu_{s_x,s_y,s'_y})}, \tag{7.21}$$

donde d es la dimensión de los vectores de características. Mediante propiedades del cálculo matricial, y teniendo en cuenta que la matriz de covarianza es diagonal, se puede observar que

$$\begin{array}{c}(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}(\mathbf{y}_t - \mu_{s_x s_y s'_y}) = Tr[(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}(\mathbf{y}_t - \mu_{s_x s_y s'_y})] = \\ Tr[\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mathbf{y}_t \mathbf{y}_t{}^T] - Tr[\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mathbf{y}_t \mu_{s_x s_y s'_y}{}^T] - \\ Tr[\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mu_{s_x s_y s'_y} \mathbf{y}_t{}^T] + Tr[\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mu_{s_x s_y s'_y} \mu_{s_x s_y s'_y}{}^T].\end{array} \tag{7.22}$$

$$\begin{array}{c}\frac{\delta}{\delta \mu_{s_x s_y s'_y}}|_{\Theta=\Theta^{k+1}}[-\frac{1}{2}(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}(\mathbf{y}_t - \mu_{s_x s_y s'_y})] = \\ -\frac{1}{2}[-\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mathbf{y}_t - \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}{}^T \mathbf{y}_t + (\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} + \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}{}^T)\mu_{s_x s_y s'_y}] = \\ -\frac{1}{2}[-2\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mathbf{y}_t + 2\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mu_{s_x s_y s'_y}].\end{array} \tag{7.23}$$

Con todo lo anterior, e introduciendo la expresión (7.23) en (7.20), se tiene finalmente que

$$\begin{array}{c}\frac{\delta \mathcal{Q}(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})}{\delta \mu_{s_x s_y s'_y}}|_{\Theta=\Theta^{k+1}} = \mathbf{0} = \\ \sum_t (v_{s_x} w_{s_y} z_{y s'_y})^{(k)}[-\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mathbf{y}_t + \mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mu_{s_x s_y s'_y}{}^{(k+1)}],\end{array} \tag{7.24}$$

$$\mu^{(k+1)}_{s_x,s_y,s'_y} = \frac{\sum_t (v_{s_x} w_{s_y} z_{s'_y})^{(k)} \mathbf{y}_t}{\sum_t (v_{s_x} w_{s_y} z_{s'_y})^{(k)}}. \tag{7.25}$$

Estimación de la matriz de covarianzas de la Gaussiana s'_y del modelado de la probabilidad entre Gaussianas.

Para estimar las matrices diagonales de covarianzas de la Gaussiana s'_y del modelado de la probabilidad entre las Gaussianas s_x y s_y, se deberá maximizar con respecto a dicha matriz de covarianzas, $\mathbf{\Sigma}_{s_x,s_y,s'_y}$, la función siguiente $\mathcal{L}(\mathbf{\Theta}) = Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$. Teniendo en cuenta que $\mathbf{\Sigma}_{s_x,s_y,s'_y}$ aparece únicamente en el segundo término aditivo de la expresión (7.10), se puede observar que la maximización requerida, aplicando (7.21), es

$$\begin{array}{c}\frac{\delta \mathcal{Q}(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})}{\delta \mathbf{\Sigma}_{s_x s_y s'_y}}|_{\Theta=\Theta^{k+1}} = \mathbf{0} = \\ \sum_t (v_{s_x} w_{s_y} z_{y s'_y})^{(k)} \frac{\delta}{\delta \mathbf{\Sigma}_{s_x s_y s'_y}}|_{\Theta=\Theta^{k+1}}[log(p(\mathbf{y}_t|v_{s_x}=1, w_{s_y}=1, z_{s'_y}=1))] = \\ \sum_t (v_{s_x} w_{s_y} z_{y s'_y})^{(k)} \frac{\delta}{\delta \mathbf{\Sigma}_{s_x s_y s'_y}}|_{\Theta=\Theta^{k+1}}[-\frac{1}{2}(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}(\mathbf{y}_t - \mu_{s_x s_y s'_y})].\end{array} \tag{7.26}$$

Mediante propiedades del cálculo matricial, y teniendo en cuenta que la matriz de covarianza es diagonal, así como la expresión (7.22), se puede observar que

$$\frac{\delta}{\delta \boldsymbol{\Sigma}_{s_x s_y s'_y}} |_{\Theta=\Theta^{k+1}} [-\frac{1}{2}(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \boldsymbol{\Sigma}^{-1}_{s_x s_y s'_y} (\mathbf{y}_t - \mu_{s_x s_y s'_y})] = \\ -\frac{1}{2}[{\boldsymbol{\Sigma}^{(k+1)}_{s_x s_y s'_y}}^{-1} - {\boldsymbol{\Sigma}^{(k+1)}_{s_x s_y s'_y}}^{-1} (\mathbf{y}_t - \mu_{s_x s_y s'_y})(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T {\boldsymbol{\Sigma}^{(k+1)}_{s_x s_y s'_y}}^{-1}]. \tag{7.27}$$

Con todo lo anterior, y llevando la expresión (7.27) a (7.26), se tiene finalmente que

$$\boldsymbol{\Sigma}^{(k+1)}_{s_x, s_y, s'_y} = \frac{1}{\sum_t (v_{s_x} w_{s_y} z_{s'_y})^{(k)}} \times \\ \sum_t (v_{s_x} w_{s_y} z_{s'_y})^{(k)} (\mathbf{y}_t - \mu^{(k)}_{s_x, s_y, s'_y})(\mathbf{y}_t - \mu^{(k)}_{s_x, s_y, s'_y})^T. \tag{7.28}$$

Obsérvese que si se mantienen todas las aproximaciones consideradas hasta el momento, y los vectores de características ruidosos se modelan en (7.3) con la misma pdf uniforme para todos los pares de Gaussianas s_x y s_y en vez de mediante Gaussianas, la expresión para el modelo de probabilidad cruzada que se debería emplear en la fase de normalización coincidiría con (5.27). Así pues, se puede decir que el modelo de probabilidad entre Gaussianas propuesto en este Capítulo no deja de ser una generalización del que ya se había considerado en Capítulos precedentes.

7.3. Aplicación del modelado de probabilidad entre Gaussianas basado en GMMs a las técnicas MEMLIN y PD-MEMLIN.

Hasta el moemnto simplemente se ha propuesto un modelo basado en GMMs para representar los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y. Sin embargo, no se ha indicado como aplicar este nuevo modelado a las técnicas de normalización de vectores de características propuestas en este trabajo. Llegado a este punto, se podrían considerar todas ellas: MEMLIN, P-MEMLIN, MEMHIN y PD-MEMLIN. Pero, y dado que hasta el momento los mejores resultados en cuanto a tasas de reconocimiento se han obtenido con las técnicas MEMLIN y PD-MEMLIN, a continuación se incluyen únicamente las extensiones de ambos métodos para incluir el nuevo modelado de la probabilidad entre Gaussianas. No obstante, la aplicación para los algoritmos P-MEMLIN y MEMHIN es directa a partir de la desarrollada para el método MEMLIN.

7.3.1. Extensión para la técnica MEMLIN

Para extender el método MEMLIN incluyendo el modelado de la probabilidad entre Gaussianas basado en GMMs es necesario, tal y como ya se adelantó, estimar los parámetros del mismo para cada entorno básico, e [Buera *et al.*, 2006b]. Esto se realiza de modo independiente para cada uno de ellos, de manera que las expresiones (7.19), (7.25) y (7.28) se evalúan con la correspondiente señal estéreo del corpus de entrenamiento de cada entorno básico, obtiéndose el consiguiente modelo

$$p(\mathbf{y}_t | s_x, s^e_y, e) = \sum_{s'_y} p(\mathbf{y}_t | s_x, s^e_y, s'_y, e) p(s'_y | s_x, s^e_y, e), \tag{7.29}$$

$$p(\mathbf{y}_t | s_x, s^e_y, s'_y, e) = \mathcal{N}(\mathbf{y}_t; \mu_{s_x, s^e_y, s'_y}, \boldsymbol{\Sigma}_{s_x, s^e_y, s'_y}), \tag{7.30}$$

donde μ_{s_x, s^e_y, s'_y}, $\boldsymbol{\Sigma}_{s_x, s^e_y, s'_y}$, y $p(s'_y | s_x, s^e_y)$ son el vector de medias, la matriz diagonal de covarianzas y la probabilidad a priori asociados a la componente s'_y del modelo de probabilidad entre Gaussianas propuesto para s_x y s^e_y. Con todo ello la nueva estimación de $p(s_x | \mathbf{y}_t, e, s^e_y)$ se calcula del siguiente modo

$$p(s_x|\mathbf{y}_t, e, s_y^e) = \frac{p(\mathbf{y}_t|s_x, s_y^e, e)}{\sum_{s_x} p(\mathbf{y}_t|s_x, s_y^e, e)}. \tag{7.31}$$

Nótese que en este caso se ha eliminado la aproximación $p(s_x|\mathbf{y}_t, e, s_y^e) \simeq p(s_x|e, s_y^e)$ considerada en el Capítulo 5. Por otra parte, si se desease realizar la correspodiente extensión para incluir el modelado de la probabilidad entre Gaussianas basado en GMMs en los métodos P-MEMLIN o MEMHIN, ésta sería exactamente la misma que la recientemente expuesta para la técnica MEMLIN.

7.3.2. Extensión para la técnica PD-MEMLIN

Para extender el método PD-MEMLIN incluyendo el modelado de la probabilidad entre Gaussianas basado en GMMs es necesario estimar los parámetros de la misma para cada entorno básico, e, y fonema, ph [Buera *et al.*, 2006a]. Esto se lleva a cabo de modo independiente, de manera que las expresiones (7.19), (7.25) y (7.28) se evalúan con la correspondiente señal estéreo del corpus de entrenamiento de cada entorno básico y fonema. Haciendo esto, se obtiene el siguiente modelo

$$p(\mathbf{y}_t|s_x^{ph}, s_y^{e,ph}, e, ph) = \sum_{s'_y} p(\mathbf{y}_t|s_x^{ph}, s_y^{e,ph}, s'_y, e, ph) p(s'_y|s_x^{ph}, s_y^{e,ph}, e, ph), \tag{7.32}$$

$$p(\mathbf{y}_t|s_x^{ph}, s_y^{e,ph}, s'_y, e, ph) = \mathcal{N}(\mathbf{y}_t; \mu_{s_x^{ph}, s_y^{e,ph}, s'_y}, \mathbf{\Sigma}_{s_x^{ph}, s_y^{e,ph}, s'_y}), \tag{7.33}$$

donde $\mu_{s_x^{ph}, s_y^{e,ph}, s'_y}$, $\mathbf{\Sigma}_{s_x^{ph}, s_y^{e,ph}, s'_y}$, y $p(s'_y|s_x^{ph}, s_y^{e,ph})$ son el vector de medias, la matriz diagonal de covarianzas y la probabilidad a priori asociados a la componente s'_y del modelo de probabilidad entre Gaussianas propuesto para s_x^{ph} y $s_y^{e,ph}$. Con todo ello la nueva estimación de $p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph})$ se calcula del siguiente modo

$$p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph}) = \frac{p(\mathbf{y}_t|s_x^{ph}, s_y^{e,ph}, e, ph)}{\sum_{s_x^{ph}} p(\mathbf{y}_t|s_x^{ph}, s_y^{e,ph}, e, ph)}. \tag{7.34}$$

Nótese que en este caso se ha eliminado nuevamente la aproximación $p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph}) \simeq p(s_x^{ph}|e, ph, s_y^{e,ph})$.

7.4. Resultados con la base de datos *SpeechDat Car* en español.

La experimentación realizada con las técnicas de normalización empíricas MEMLIN y PD-MEMLIN aplicando el nuevo modelado de la probabilidad entre Gaussianas basado en GMMs se realizó con la base de datos *SpeechDat Car* en español. A la hora de obtener los distintos componentes de las correspondientes funciones del modelado de $\mathbf{x}$, eto es, los vectores de desplazamiento, así como los parámetros de las GMMs que componen el modelo de la probabilidad entre Gaussianas, se hará uso del corpus de entrenamiento correspondiente a cada entorno básico y fonema, esto último sólo para el caso de la técnica PD-MEMLIN. Por otra parte, y una vez que se ha llevado a cabo la normalización de los vectores acústicos degradados con las correspondientes técnicas, se aplicará el método CMS. Para toda esta experimentación se utilizó la *parametrización UZ* y modelos acústicos de unidades fonéticas, de modo que los resultados de referencia se pueden consultar en la Sección 4.3. Se puede apreciar igualmente que todos los parámetros que definen los experimentos en este caso coinciden con los aplicados en las Secciones 5.4 y 6.5, de manera que los

resultados son totalmente comparables. Asimismo la Figura 5.5 sigue siendo válida para explicar los tres pasos precisados para llevar a cabo la experimentación.

7.4.1. Resultados para la técnica MEMLIN con modelado de la probabilidad entre Gaussianas basado en GMMs.

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	MEMLIN 128	2.30	7.46	4.62	6.39	8.77	5.40	8.16	6.05	70.22
CLK	MEMLIN MP 128-2	2.01	6.43	3.92	5.76	6.48	4.13	4.42	4.86	78.48

Cuadro 7.2: Mejores resultados con la base de datos *SpeechDat Car* en español para las técnicas de normalización de vectores de características MEMLIN y MEMLIN con Modelado de la Probabilidad entre Gaussianas basado en GMMs (MEMLIN MP) en términos de WER (%). Se presentan los resultados para los diferentes entornos básicos (E1,..., E7) utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia (CLK en la columna de Entrenamiento). La columna marcada como "Reconocimiento" hace referencia a la señal empleada para reconocer, que será la ruidosa normalizada con las técnicas MEMLIN y MEMLIN MP, estos primeros resultados incluidos a modo de comparación. Junto al nombre de los diferentes métodos aparece el número de Gaussianas con que se modelaron los correspondientes espacios (el limpio y los asociados a los distintos entornos básicos), incluyendo además para el caso del método MEMLIN MP el número de componentes de las GMMs que constituyen el modelado de la probabilidad entre Gaussianas. Se presenta igualmente el WER medio, MWER, así como la mejora media, MIMP.

En la Tabla 7.2 se pueden apreciar los mejores resultados de RAH para la técnica de normalización de vectores de características MEMLIN con Modelado de Probabilidad entre Gaussianas basado en GMMs, que se ha identificado como MEMLIN MP; asimismo se incluyen también, a modo de comparación, los resultados correspondientes obtenidos con el método MEMLIN y que ya fueron introducidos en la Sección 5.4. En ambos casos, junto al nombre de la técnica, MEMLIN y MEMLIN MP, se incluye el número de componentes que conforman las GMMs necesarias para obtener los correspondientes resultados en cada caso: el primer valor, 128 en este caso, se corresponde con el número de componentes empleadas para modelar los espacios limpio y el asociado a cada entorno básico (se realizó un barrido con 4, 8, 16, 32, 64 y 128 Gaussianas, cuyos resultados completos se pueden apreciar en los Anexos 5.6 y 7.5). El segundo valor para MEMLIN MP, 2, es el número de componentes con que se modela la señal ruidosa asociada a cada par de Gaussianas, s_x y s_y^e, (se realizó un barrido con 1, 2 y 4 componentes y, por cuestiones de coste computacional, se decidió emplear únicamente 2). Cabe destacar que de aquí en adelante para todas las técnicas tratadas en este Capítulo, y mientras no se indique lo contrario, el número de Gaussianas empleadas para modelar el espacio limpio será el mismo que el utilizado para representar cada entorno básico. Asimismo se incluye en la Tabla 7.2, además del WER medio, MWER, la mejora media de WER, MIMP, en tanto por ciento, que se calcula a partir de la expresión (5.29) como se indica en el Capítulo 5.4.

Por otra parte, es conveniente analizar mediante la prueba de hipótesis estadística *z-test* si el comportamiento de la técnica MEMLIN MP, es estadísticamente diferente con respecto al del algoritmo MEMLIN para la base de datos *SpeechDat Car* en español. De este modo, el valor del estadístico W, w, es $w = 2{,}8 > 1{,}96$, por lo que la mejora que proporciona el algoritmo MEMLIN MP en este caso sí se puede considerar independiente de la base de datos con un intervalo de confianza del 95 %. De todas maneras, a la hora de valorar las conclusiones obtenidas mediante la

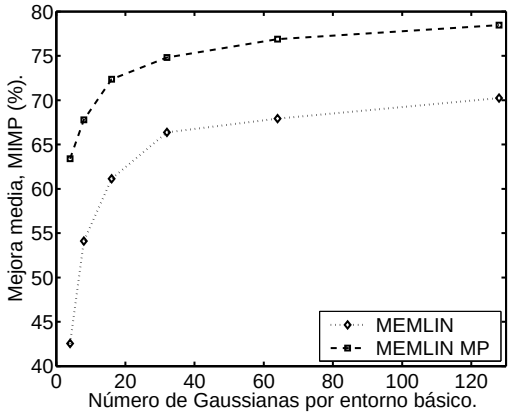


Figura 7.2: Mejora media de WER con la base de datos *SpeechDat Car* en español empleando las técnicas MEMLIN (línea punteada con diamantes blancos) y MEMLIN con modelado de la probabilidad entre Gaussianas basado en GMMs, MEMLIN PM, (línea discontinua con cuadrados blancos). Se presentan los resultados utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia

hipótesis estadística *z-test*, hay que tener en cuenta siempre las limitaciones de la propia prueba, ya comentadas en la Sección 4.2.

La Figura 7.2 muestra la mejora media en términos de WER (MIMP) en % para las técnicas MEMLIN y MEMLIN MP cuando se varía el número de Gaussianas con que se modela el espacio limpio y los distintos entornos básicos. Para el caso del método MEMLIN MP, y por cuestiones de no incrementar excesivamente el coste computacional, la probabilidad entre Gaussianas se representa mediante dos componentes para cada par de Gaussianas, s_x y s_y^e. Se puede apreciar como en todos los casos se produce una importante mejora en los resultados cuando se incluye el nuevo modelado de la probabilidad entre Gaussianas; así, cuando se representan los entornos básicos con 4 Gaussianas, la mejora media se incrementa desde 42.56 % hasta 63.39 %, mientras que si se modelan los entornos básicos con 128 componentes, el aumento se produce desde 70.22 % hasta 78.48 %.

Como resumen, y a la luz pues de los resultados presentados en las Tablas 5.1 y 7.2 se puede concluir que, teniendo en cuenta únicamente los mejores resultados para las distintas técnicas y para todos y cada uno de los entornos, la introducción del modelado de la probabilidad entre Gaussianas basado en GMMs a la técnica MEMLIN aporta una mejora estadísticamente significativa con respecto al propio algoritmo MEMLIN, mejorando asimismo claramente el comportamiento de métodos como SPLICE con selección de modelos de entorno o IRATZ. Cabe destacar del mismo modo que la mejora alcanzada es más acusada cuando el número de componentes con que se modela el espacio limpio o los entornos básicos es reducido, tal y como queda patente en la Figura 7.2.

Sin embargo, y aunque la mejora al incluir el modelado de la probabilidad entre Gaussianas basado en GMMs es clara para cualquier número de componentes con que se representen los entornos básicos, el coste computacional en este caso es mucho mayor puesto que se debe evaluar un número más elevado de *scores* de Gaussianas por entorno básico y vector de características ruidoso, n_G, lo que es, a la postre, el término más gravoso computacionalmente hablando del proceso de normalización. De hecho, en este caso n_G será

$$n_G = n_{s_y^e}(1 + n_{s_x} \times n_{{s'_y}^e}), \tag{7.35}$$

donde $n_{s_y^e}$ es el número de Gaussianas con que se modela cada entorno básico, n_{s_x} es el número de componentes con que se constituye la GMM que representa el espacio limpio y finalmente $n_{{s'_y}^e}$ es el número de Gaussianas con que se representan los vectores de características ruidosos asociados a cada par de Gaussianas s_x y s_y^e. Se puede observa como la diferencia con respecto al algoritmo MEMLIN ($n_G = n_{s_y^e}$) puede llegar a ser muy importante. Para reducir el coste computacional de la técnica MEMLIN MP se propone evaluar en el proceso de normalización únicamente aquellos pares de Gaussianas, s_x y s_y^e, más probables para cada vector de características. Para ello, primeramente se determinan aquellas Gaussianas, $n'_{s_y^e}$, de los modelos de cada entorno básico con mayor *score* haciendo uso de las expresiones (5.19) y (5.20). Posteriormente se elige para cada una de ellas aquellas del modelo limpio más probables, n'_{s_x}, mediante (5.27) o (5.28), atendiendo al tipo de solución, *hard* o *soft* respectivamente. De este modo el número final de *scores* que se han de evaluar para cada vector de características y entorno básico en la fase de normalización pasa a ser

$$n_G = n_{s_y^e} + n'_{s_y^e} \times n'_{s_x} \times n_{{s'_y}^e}, \tag{7.36}$$

En la Tabla 7.3 se muestran los resultados para la técnica MEMLIN MP para distintos valores de $n'_{s_y^e}$ y n'_{s_x}, siendo en todos los casos $n_{{s'_y}^e} = 2$. Adicionalmente, junto al nombre de la técnica se añade el número de Gaussianas con que se han modelado el espacio limpio y los entornos básicos. Se puede observar como, si bien la disminución del número de componentes evaluadas, n_G, reduce ligeramente las prestaciones del método, los resultados obtenidos siguen siendo satisfactorios, a la vez que se minimiza el coste computacional hasta en un factor 15.

	$n'_{s_y^e}$	n'_{s_x}	MWER	MIMP
MEMLIN MP 4	4	4	7.04	63.40
MEMLIN MP 8	4	4	6.87	64.40
MEMLIN MP 16	8	8	5.67	72.87
MEMLIN MP 32	8	8	5.62	73.23
MEMLIN MP 64	16	16	5.44	74.46
MEMLIN MP 128	32	32	5.11	76.77

Cuadro 7.3: Resultados medios (MWER y MIMP) con la base de datos *SpeechDat Car* en español para la técnica de normalización de vectores de características MEMLIN MP en términos de WER (%), cuando se reduce el número de Gaussianas evaluadas en el proceso de normalización ($n'_{s_y^e}$ y n'_{s_x}). Se presentan los resultados utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de entrenamiento de la señal limpia. Junto al nombre del método aparece el número de Gaussianas con que se modelaron los correspondientes espacios (el limpio y los asociados a los distintos entornos básicos). El número de componentes de las GMMs que constituyen el modelado de la probabilidad entre Gaussianas es, en todos los casos, 2.

7.4.2. Resultados para la técnica PD-MEMLIN con modelado de la probabilidad entre Gaussianas basado en GMMs.

A continuación se comparan los resultados obtenidos con las técnicas PD-MEMLIN y PD-MEMLIN con Modelado de la Probabilidad entre Gaussianas basado en GMMs, que se identificará en lo sucesivo como PD-MEMLIN MP. Para realizar la correspondiente normalización se

entrenaron y emplearon transformaciones para los 25 fonemas españoles más el silencio a pesar de que, como ya se ha indicado, para esta tarea concreta de dígitos no todos los fonemas son necesarios.

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	PD-MEMLIN 16	1.73	8.23	5.45	4.64	6.86	3.02	7.14	5.30	75.44
CLK	PD-MEMLIN MP 16-2	1.92	7.46	5.31	5.14	5.82	3.81	4.08	4.97	77.72

Cuadro 7.4: Mejores resultados con la base de datos *SpeechDat Car* en español para las técnicas de normalización de vectores de características PD-MEMLIN y PD-MEMLIN con Modelado de la Probabilidad entre Gaussianas basado en GMMs (PD-MEMLIN MP) en términos de WER (%). Se presentan los resultados para los diferentes entornos básicos (E1,..., E7) utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia (CLK en la columna de Entrenamiento). La columna marcada como "Reconocimiento" hace referencia a la señal empleada para reconocer, que será la ruidosa normalizada con las técnicas PD-MEMLIN y PD-MEMLIN MP, estos primeros resultados incluidos a modo de comparación. Junto al nombre de los diferentes métodos aparece el número de Gaussianas con que se modelaron los correspondientes espacios (el limpio y los asociados a los distintos entornos básicos), incluyendo además para el caso del método PD-MEMLIN MP el número de componentes de las GMMs que constituyen el modelado de la probabilidad entre Gaussianas. Se presenta igualmente el WER medio, MWER, así como la mejora media, MIMP.

En la Tabla 7.4 se pueden apreciar los mejores resultados para las técnicas de normalización de vectores de características PD-MEMLIN (ya incluidos en la Sección 6.5) y PD-MEMLIN MP. En ambos casos, junto al nombre de la técnica se incluye el número de componentes que conforman las GMMs necesarias para obtener los correspondientes resultados en cada caso: el primer valor, 16 para ambas técnicas, se corresponde con el número de Gaussianas empleadas para modelar cada fonema de los espacios limpio y de los entornos básicos (se realizó un barrido con 2, 4, 8, 16 y 32 componentes, cuyos resultado completos se pueden observar en los Anexos 6.10 y 7.5). Por su parte, el segundo valor que acompaña al nombre de la técnica PD-MEMLIN MP (2) es el número de componentes con que se modela la señal ruidosa asociada a cada par de Gaussianas de cada fonema, s_x^{ph} y $s_y^{e,ph}$. Asimismo se incluye en la Tabla, además del WER medio, MWER, la mejora media de WER, MIMP, en tanto por ciento, que se calcula a partir de la expresión (5.29), ya presentada en el Capítulo 5.4.

A la luz de los resultados presentados en la Tabla 7.4 se puede asegurar que la técnica PD-MEMLIN MP proporciona unos resultados, al menos para la mejor combinación tratada de número de Gaussianas, superiores que los alcanzados por el algoritmo PD-MEMLIN.

Por otra parte, es conveniente analizar mediante la prueba de hipótesis estadística *z-test* si el comportamiento de la técnica PD-MEMLIN MP, es estadísticamente diferente con respecto al del algoritmo PD-MEMLIN para la base de datos *SpeechDat Car* en español. De este modo, el valor del estadístico W, w, es $w = 0{,}8 < 1{,}96$, por lo que la mejora que proporciona el algoritmo PD-MEMLIN MP en este caso no se puede considerar independiente de la base de datos con un intervalo de confianza del 95 % con respecto a la técnica PD-MEMLIN. Sin embargo, si se realiza la comparación entre los algoritmos PD-MEMLIN MP y MEMLIN, el valor del estadístico es $w = 2{,}53 > 1{,}96$, de modo que en este caso sí hay una diferencia de comportamiento estadísticamente significativo con un intervalo de confianza del 95 %. De todas maneras, a la hora de valorar las conclusiones obtenidas mediante la hipótesis estadística *z-test*, hay que tener en cuenta siempre las limitaciones de la propia prueba, ya comentadas en la Sección 4.2.

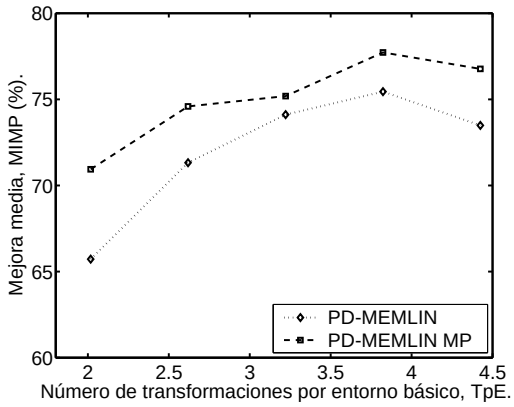


Figura 7.3: Mejora media de WER con la base de datos *SpeechDat Car* en español empleando las técnicas PD-MEMLIN (línea punteada con diamantes blancos) y PD-MEMLIN con modelado de la probabilidad entre Gaussianas basado en GMMs, MEMLIN PM, (línea discontinua con cuadrados blancos). Se presentan los resultados utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de señal limpia

Para estudiar la tendencia del comportamiento de las técnicas PD-MEMLIN y PD-MEMLIN MP en función del número de transformaciones por entorno básico, TpE en $\log_{10}$, se presenta la Figura 7.3, en la que se muestra la mejora media en términos de WER (MIMP) en % para ambas técnicas. En todos los casos, la GMM asociada a la probabilidad entre Gaussianas se modela mediante dos componentes para cada par de Gaussianas y fonema, s_x^{ph} y $s_y^{e,ph}$. Se puede apreciar como para todos los TpE estudiados se produce una cierta mejora en los resultados cuando se incluye el nuevo modelado de la probabilidad entre Gaussianas, así, cuando se modela cada fonema con 2 componentes, se incrementa la mejora media desde 65.72 % hasta 70.94 %, mientras que si los fonemas se representan con 32 Gaussianas, la mejora aumenta desde 75.44 % hasta 76.76 %.

Así pues, y a la luz de los resultados presentados en las Tablas 5.1 y 7.4, se puede concluir que, teniendo en cuenta únicamente los mejores resultados para las distintas técnicas y para todos y cada uno de los entornos, la introducción del modelado de la probabilidad entre Gaussianas basado en GMMs a la técnica PD-MEMLIN aporta una cierta mejora con respecto al algoritmo PD-MEMLIN, proporcionando igualmente mejores resultados con respecto a métodos como SPLICE con selección de modelos de entorno o IRATZ.

Sin embargo, y aunque la mejora al incluir el modelado de la probabilidad entre Gaussianas basado en GMMs es clara para cualquier número de transformaciones por entorno básico, TpE, el coste computacional es, como sucedía con el método MEMLIN MP, mucho mayor puesto que se debe evaluar un número más elevado de *scores* de Gaussianas por vector de características ruidoso en la fase de normalización, n_G, lo que, a la postre, supone el mayor coste computacional de dicha fase. Así pues, en este caso n_G será

$$n_G = n_{ph} \times n_{s_y^{e,ph}}(1 + n_{s_x^{ph}} \times n_{s'_y{}^{e,ph}}), \tag{7.37}$$

donde $n_{s'_y{}^{e,ph}}$ es el número de componentes con que se modelan los vectores de características ruidosos asociados a cada par de Gaussianas s_x^{ph} y $s_y^{e,ph}$. A su vez, se recuerda que n_{ph} es el número de fonemas, $n_{s_y^{e,ph}}$ es el número de Gaussianas con que se modela cada fonema ph del entorno básico e y, finalmente, $n_{s_x^{ph}}$ se corresponde con el número de componentes con que se

representa cada fonema *ph* en el espacio limpio. Se puede apreciar como el número de *scores* que se han de evaluar por entorno básico y vector de carcaterísticas en la fase de normalización puede llegar a ser, en este caso, mucho mayor que el necesitado para el algoritmo PD-MEMLIN ($n_G = n_{ph} \times n_{s_y^{e,ph}}$). Para reducir el coste computacional se propone evaluar en el proceso de normalización únicamente aquellos pares de Gaussianas, s_x^{ph} y $s_y^{e,ph}$, más probables para cada vector de características. Para ello, primeramente se calculan los n'_{ph} fonemas más probables para cada entorno básico mediante (6.9) y (6.10). Una vez seleccionados los fonemas más probables, y haciendo uso de las mismas expresiones, se eligen las $n'_{s_y^{e,ph}}$ Gaussianas de los modelos de cada fonema seleccionado y entorno básico con mayor *score*. Finalmente se toma para cada una de las componentes de las GMMs ruidosas seleccionadas, aquellas del modelo limpio más probables, $n'_{s_x^{ph}}$, haciendo uso de (6.20) o (6.21), atendiendo al tipo de solución, *hard* o *soft* respectivamente. De este modo el número final de *scores* por entorno básico que se han de evaluar para cada vector de características en la fase de normalización es

$$n_G = n_{ph} \times n_{s_y^{e,ph}} + n'_{ph} \times n'_{s_y^{e,ph}} \times n'_{s_x^{ph}} \times n_{s'^{e}_{y}}, \tag{7.38}$$

En la Tabla 7.5 se muestran los resultados para la técnica PD-MEMLIN PM para distintos valores de n'_{ph}, $n'_{s_y^{e,ph}}$ y $n'_{s_x^{ph}}$, siendo en todos los casos $n_{s'^{e,ph}_{y}} = 2$. Junto al nombre del algoritmo se añade el número de Gaussianas con que se ha modelado los distintos fonemas para el espacio limpio y los entornos básicos. Se puede observar como, si bien la disminución del número de componentes calculadas reduce ligeramente las prestaciones del método, los resultados obtenidos siguen siendo satisfactorios reduciendo llegando a reducir el coste computacional hasta en un factor 4.38.

	n'_{ph}	$n'_{s_y^{e,ph}}$	$n'_{s_x^{ph}}$	MWER	MIMP
PD-MEMLIN MP 2	8	2	2	6.00	70.58
PD-MEMLIN MP 4	8	4	4	5.58	73.49
PD-MEMLIN MP 8	8	6	6	5.39	74.82
PD-MEMLIN MP 16	13	12	12	5.04	77.25
PD-MEMLIN MP 32	13	25	25	5.02	77.36

Cuadro 7.5: Resultados medios (MWER y MIMP) con la base de datos *SpeechDat Car* en español para la técnica de normalización de vectores de características PD-MEMLIN MP en términos de WER (%), cuando se reduce el número de Gaussianas evaluadas en el proceso de normalización (n'_{ph}, $n'_{s_y^{e,ph}}$ y $n'_{s_x^{ph}}$). Se presentan los resultados utilizando la *parametrización UZ* y modelos acústicos para las unidades fonéticas generados a partir del corpus de entrenamiento de la señal limpia. Junto al nombre del método aparece el número de Gaussianas con que se modelaron los correspondientes espacios (el limpio y los asociados a los distintos entornos básicos). El número de componentes de las GMMs que constituyen el modelado de la probabilidad entre Gaussianas es, en todos los casos, 2.

Si se comparan los resultados obtenidos con las técnicas MEMLIN MP y PD-MEMLIN MP, se puede constatar que el segundo de los métodos proporciona una ligera mejora relativa con respecto al algoritmo MEMLIN MP. Adicionalmente, y por completar la comparación, se ha incluido la Figura 7.4, que muestra los histogramas y *log-scattegrams* del primer coeficiente MFCC de los vectores de características de voz limpios del entorno básico E4 de la base de datos *SpeechDat Car* en español y los correspondientes normalizados mediante los métodos MEMLIN MP y PD-MEMLIN MP. Para el primer caso se emplean 128 Gaussianas por entorno básico, mientras que la técnica

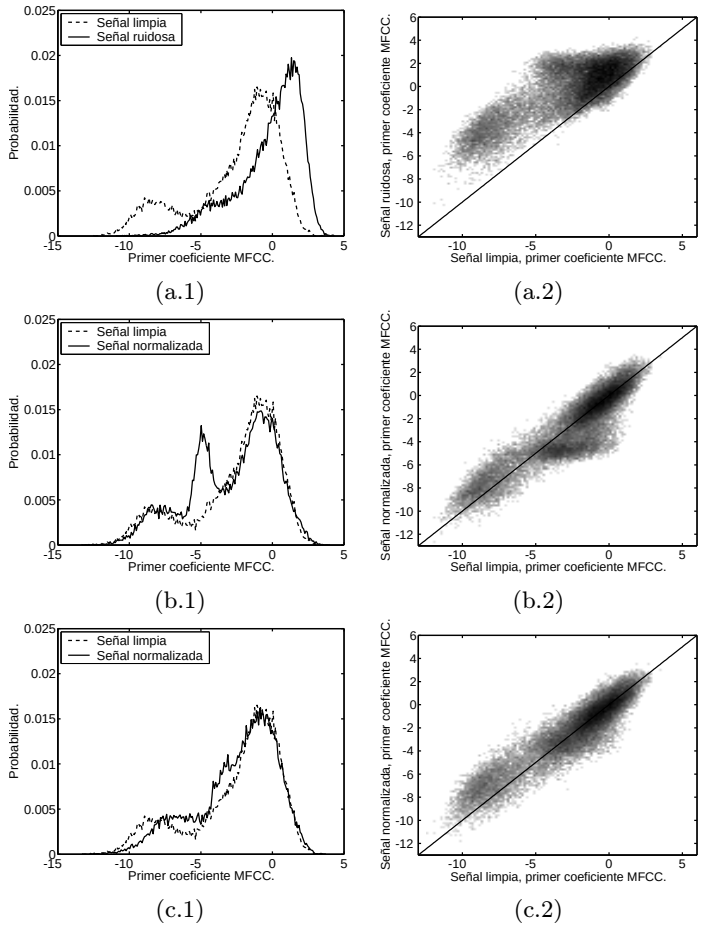


Figura 7.4: *Log-scattegrams* e histogramas realizados entre el primer coeficiente MFCC de las tramas de voz de la señal limpia (eje de abscisas) y la señal ruidosa (a) (eje de ordenadas), o normalizada usando la técnica MEMLIN MP con 128 Gaussianas por entorno básico (b) (eje de ordenadas). En la figura (c) se representa el *log-scattegram* y el histograma obtenidos a partir de la señal normalizada con la técnica PD-MEMLIN MP con 16 Gaussianas por entorno básico y fonema. Las GMMs que componen el modelo de probabilidad entre Gaussianas para ambas técnicas están compuestas por 2 componentes. Todas las representaciones se realizaron a partir del corpus de reconocimiento del entorno básico E4 de la base de datos *SpeechDat Car* en español. La línea en los *log-scattegrams* representa la función $x = y$.

PD-MEMLIN MP se llevó a cabo haciendo uso de 16 componentes por fonema y entorno básico (las combinaciones que mejores resultados en términos de WER han proporcionado en cada caso). Ciertamente, y a nivel visual, no hay grandes diferencias con las representaciones ya incluidas para los métodos MEMLIN (Figura 5.7) y PD-MEMLIN (Figura 6.6). Así pues, teniendo en cuenta todos los resultados presentados en esta Sección, se puede concluir que el nuevo modelado entre Gaussianas basado en GMMs propuesto proporciona una importante mejora de comportamiento en términos de WER, a costa, eso sí, de un mayor incremento del coste computacional. Sin embargo, este último inconveniente se ve minimizado si se reduce el número de pares de Gaussianas computadas, sin que eso suponga un elevado coste en las tasas de reconocimiento.

7.5. Anexo H.

8

Técnicas híbridas de normalización de vectores de características y adaptación de modelos acústicos.

En el Capítulo 3 se presentó una taxonomía formal de los distintos métodos empleados a la hora de dotar de robustez a los sistemas de RAH. En ella se distinguía principalmente entre dos grandes líneas de actuación: las técnicas de normalización de vectores de características y los algoritmos de adaptación de modelos acústicos. En el primero de los casos se proyectan los vectores acústicos de espacio ruidoso al de referencia, que normalmente coincide con el limpio; mientras que en la segunda línea de actuación son los modelos acústicos que representan al espacio de referencia los que se acercan estadísticamente a las condiciones de los vectores acústicos ruidosos. Igualmente se consideraron en dicho Capítulo las ventajas e inconvenientes que ambas filosofías poseen, siendo en algunos casos complementarias. Basándose en estas características, así como en la incapacidad de que las técnicas de normalización de vectores acústicos proporcionen una transformación perfecta debido a la naturaleza aleatoria del ruido, surge la idea de combinar ambos tipos de algoritmos, definiéndose así las soluciones híbridas [Sankar and Lee, 1996].

Si se hace un repaso a los métodos presentados hasta el momento en este trabajo y se analizan aquellos cuyos comportamientos han resultados más satisfactorios, se puede comprobar que en todos ellos, PD-MEMLIN, MEMLIN MP y PD-MEMLIN MP, se propone una transformación lineal con término dependiente unitario, de modo que, si bien se compensa eficazmente el desplazamiento de los vectores de características producido por el entorno acústico, tal y como se ha podido constatar a partir de los correspondientes histogramas y *scattegrams* presentados en Capítulos anteriores, no hace lo propio con otros efectos, como por ejemplo las rotaciones, deformaciones estas que pueden producirse, entre otras causas, por la variabilidad inter-locutor.

Para tratar de solucionar conjuntamente las alteraciones anteriormente consideradas: tanto el desplazamiento de los vectores de características, como la rotación de los mismos, en este Capítulo se propone el uso de técnicas híbridas, de modo que con el método de normalización correspondiente, que en general puede ser cualquiera de los presentados previamente en este trabajo, se pretende compensar el desplazamiento de los vectores de características, cosa que ya ha quedado demostrada a partir de los *log-scattegrams* e histogramas obtenidos a partir de los distintos experimentos realizados; mientras que por otra parte, con la técnica de adaptación de modelos acústicos se busca modificar el modelado del espacio limpio proyectándolo sobre el normalizado, entendiendo por espacio normalizado aquél que representa a los vectores acústicos ruidosos compensados con

la técnica de normalización correspondiente; con ello se pretende reducir de un modo estadístico aquellos desajustes de los vectores de características que la técnica de normalización no ha considerado. Esta segunda fase puede ser supervisada, si se precisan las trascripciones de los datos empleados para reestimar los modelos acústicos, o no supervisadas, si no son necesarias.

A modo esquema conceptual se incluye la Figura 8.1, en la que se presenta la filosofía que conjuntamente siguen las distintas técnicas híbridas consideradas en este Capítulo. En la parte de la izquierda de la misma se aprecia el efecto de la compensación producida por la técnica de normalización correspondiente, que desplaza los vectores de características ruidosos hacia el espacio normalizado, que, en general, no coincide con el de referencia. Por su parte, el efecto de la técnica de adaptación de modelos acústicos se muestra en la parte de la derecha de la Figura, en la que se observa como se modifica el modelado del espacio de referencia al proyectarlo sobre el normalizado.

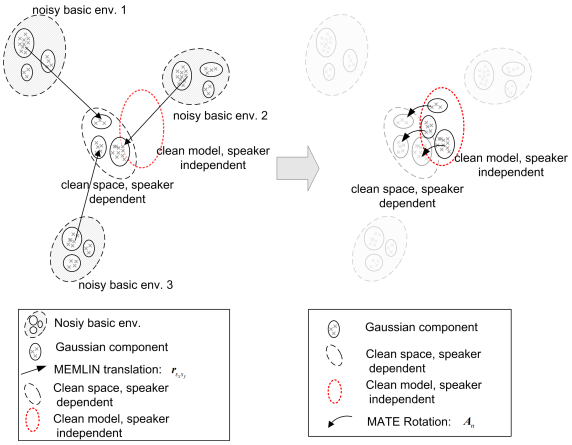


Figura 8.1: Representación gráfica de las técnicas híbridas propuestas en este Capítulo. La parte izquierda está dedicada a la normalización de los vectores de características, cuya misión es proyectar los ruidosos desde un determinado entorno básico a un espacio normalizado que, por las limitaciones del método en cuestión, no coincide con el limpio. La parte derecha está dedicada a la transformación de los modelos acústicos, que los acerca desde el espacio de referencia al normalizado.

Ya se ha indicado que cualquiera de las técnicas de normalización de vectores de características presentada en este trabajo puede formar parte, como primera fase, de los distintos algoritmos híbridos incluidos en este Capítulo. Por otra parte, para la segunda fase, compuesta por el método de adaptación de modelos acústicos, se propone emplear dos posibles líneas de trabajo, según si éstos son supervisados o no. Así, para la segunda opción se desarrollaron diversas técnicas basadas en matrices de rotación dependientes de GMMs (técnicas híbridas a partir del cálculo de matrices de rotación dependientes de GMMs), de modo que a cada vector acústico normalizado se le asignará una de dichas matrices en el proceso de decodificación mediante la optimización del criterio ML. Como segunda línea de trabajo a la hora de definir técnicas híbridas, se presenta asimismo la posibilidad, si se posee la suficiente cantidad de datos, de estimar de modo supervisado unos nuevos modelos acústicos asociados al espacio normalizado tras compensar convenientemente el corpus de entrenamiento ruidoso de que se disponga.

Este Capítulo se organiza del siguiente modo: en la Sección 8.1 se presentan las distintas técnicas híbridas no supervisadas basadas en el cálculo de matrices de rotación dependientes de

GMMs. Por su parte, las técnicas híbridas supervisadas, en las que se generan nuevos modelos acústicos mediante el corpus de entrenamiento ruidoso compensado, se introducen en la Sección 8.2. Finalmente los resultados obtenidos con los distintos algoritmos a partir de la base de datos *SpeechDat Car* en español se incluyen en la Sección 8.3.

8.1. Técnicas Híbridas no Supervisadas Basadas en Matrices de Rotación.

A la hora de compensar conjuntamente tanto el desplazamiento como la rotación de los vectores de características se puede recurrir, como una posible solución, a actuar sobre los modelos acústicos del espacio de referencia tras considerar un modelado de los vectores acústicos limpios, $\mathbf{x}$, semejante a los planteados hasta el momento para las distintas técnicas de normalización introducidas en este trabajo. De hecho, y bajo ciertas circunstancias, algunos métodos de normalización de vectores de características se pueden ver como algoritmos de adaptación de modelos acústicos. Así por ejemplo se da con la técnica MEMLIN: sea un modelado acústico compuesto por HMMs a cuyos estados les corresponde GMMs como pdfs asociadas; a su vez, la estimación de la trama limpia, según el método MEMLIN, se obtiene como $\hat{\mathbf{x}}_t = \mathbf{y}_t - \mathbf{u}_t$, donde $\mathbf{u}_t$ es el vector de desplazamiento final para el instante de tiempo t. Finalmente, y tal y como se ha venido realizando hasta el momento, la decodificación del vector compensado, $\hat{\mathbf{x}}_t$, se lleva a cabo con los modelos acústicos limpios; lo que es matemáticamente idéntico a reconocer el vector ruidoso, $\mathbf{y}_t$, con los modelos acústicos limpios adaptados, entendiendo por adaptación en este caso a modificar todos los vectores de medias de las distintas Gaussianas, sumándoles el vector $\mathbf{u}_t$. Esta equiparación entre técnicas de compensación de vectores de características y métodos de adaptación de modelos acústicos no es única, pudiéndose decir lo mismo de métodos como RATZ, SPLICE, PD-MEMLIN..., y, en general, de todos aquellos que propongan como compensación únicamente la adición de un vector de desplazamiento.

Sin embargo, el utilizar una transformación basada únicamente en un vector de desplazamiento como la anterior hace inviable compensar cualquier tipo de rotación sobre los vectores de características, por lo que es necesario plantear un nuevo modelo de degradación algo más complejo que incluya, al menos, un término multiplicativo. En ese sentido, el más sencillo posible consistiría en $\mathbf{y}_t = \mathbf{A}_t\mathbf{x}_t + \mathbf{g}_t$, donde $\mathbf{A}_t$ es la matriz de rotación para el instante de tiempo t y $\mathbf{g}_t$ es el vector de desplazamiento entre el vector de características ruidoso y el correspondiente limpio rotado. Nótese que la expresión propuesta es similar a la considerada para la técnica P-MEMLIN en la Seccion 6.1, aunque en este caso no se restringe que la matriz de rotación deba ser diagonal. El suponer un modelo como el anterior implica considerar que los vectores de medias, $\mu_{y,t}$, y las matrices de covarianza, $\mathbf{\Sigma}_{y,t}$ asociados a los vectores de características ruidosos en el instante de tiempo t son, con respecto a los consiguientes parámetros asociados a los vectores acústicos limpios, μ_x y $\mathbf{\Sigma}_x$ respectivamente, $\mu_{y,t} = \mathbf{A}_t\mu_x + \mathbf{g}_t$ y $\mathbf{\Sigma}_{y,t} = {\mathbf{A}_t}^T\mathbf{\Sigma}_x\mathbf{A}_t$. Lamentablemente, para este caso no es posible definir una equiparación como la tratada anteriormente, ya que decodificar los vectores de características normalizados con los modelos acústicos limpios y hacer lo propio con los vectores degradados y los modelos acústicos compuestos por $\mu_{y,t}$ y $\mathbf{\Sigma}_{y,t}$ no proporciona los mismos resultados. Esto es debido a que, si bien el exponente de las distintas Gaussianas computadas es idéntico, no lo es el término multiplicativo de la exponencial, que en el primer caso incluirá la expresión $|\Sigma_x|$ y en el segundo $|{\mathbf{A}_t}^T\mathbf{\Sigma}_x\mathbf{A}_t|$, produciéndose por tanto un desajuste que hace más recomendable de cara a RAH el uso de la segunda de las opciones, esto es, reconocer los vectores de características ruidosos haciendo uso de los modelos acústicos adaptados, aunque para ello haya que pagar un mayor coste computacional. Obsérvese que esta solución es idéntica a decodificar los vectores acústicos normalizados, $\hat{\mathbf{x}}_t = \mathbf{y}_t - \mathbf{g}_t$, con los modelos acústicos adaptados haciendo

uso únicamente de la matriz de rotación $\mathbf{A}_t$: $\mu_{\hat{x},t} = \mathbf{A}_t\mu_x$ y $\mathbf{\Sigma}_{\hat{x},t} = {\mathbf{A}_t}^T\mathbf{\Sigma}_x\mathbf{A}_t$. De este modo se independiza el efecto compensatorio del vector de desplazamiento con el de la matriz de rotación. A este tipo de técnicas híbridas se les va a denominar basadas en matrices de rotación.

Nótese por otra parte que, para el ejemplo anterior, en el que la normalización de los vectores de características se lleva a cabo únicamente mediante un vector de desplazamiento, la técnica híbrida se puede ver, desde un punto de vista conceptual, no como una combinación de una técnica de normalización con una de adaptación de modelos acústicos, sino como un nuevo algoritmo de adaptación de modelos acústicos *on line*. Sin embargo, y a pesar de ello, en este trabajo se les seguirá denominando técnicas híbridas y como tales se tratarán.

A modo de resumen se incluye la Figura 8.2, en la que se reproducen los distintos pasos que se han de seguir para evaluar una técnica híbrida basada en matrices de rotación. Adviértase que, en previsión de emplear técnicas de normalización de vectores de características como las planteadas en este trabajo, se ha incluido un bloque de entrenamiento para las mismas que, en principio y en un caso general, no siempre sería necesario. En el bloque denominado "Estimación de matriz" se obtiene un conjunto de matrices de rotación, $\mathbf{A}_i$, a partir del cual se elegirá posteriormente $\mathbf{A}_t$ para cada instante de tiempo t, esto último ya en la fase de decodificación mediante el vector de características normalizado. El bloque identificado como "Normalización" incluye la técnica de compensación de vectores de características empleada en el método híbrido.

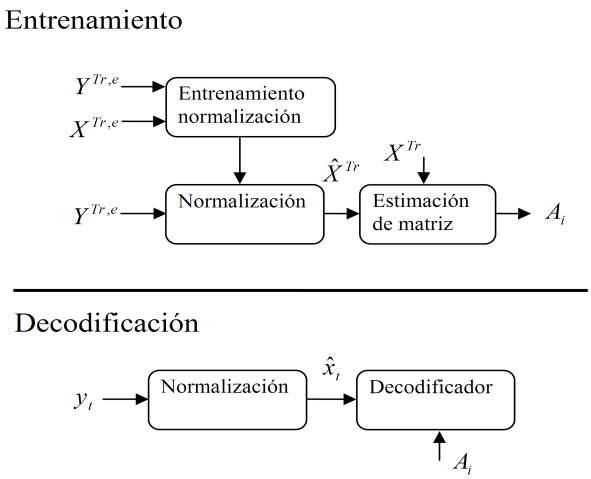


Figura 8.2: Esquema gráfico de las técnicas híbridas basadas en matrices de rotación. Se incluye tanto la fase de entrenamiento como la de decodificación. La primera de ellas está compuesta por tres bloques: "Entrenamiento normalización", que se ha incluido en previsión de utilizar técnicas de normalización de vectores de características que la precisen. Por su parte, el sistema de "Normalización" proporciona las señales limpias estimadas a partir de las correspondientes degradadas. Finalmente el bloque "Estimación de matriz" calcula un conjunto de matrices de rotación, una de las cuales será seleccionada por cada vector de características limpio estimado en la fase de decodificación en el bloque "Decodificador".

8.1.1. Técnicas Híbridas a Partir del Cálculo de Matrices de Rotación Dependientes de GMMs.

En esta Sección, dejando a un lado la primera fase de la técnica híbrida constituida por la normalización de los vectores de características, que ya ha sido tratada convenientemente en Capítulos anteriores, se propone obtener un conjunto de matrices de rotación dependientes de los distintos pares de Gaussianas con que se modelan los espacios normalizado y limpio, aplicando una filosofía similar a la empleada en las distintas técnicas de normalización de vectores de características presentadas en este trabajo. Una vez obtenido el conjunto de posibles matrices de rotación se selecciona para cada vector acústico normalizado que se pretenda decodificar aquélla, $\mathbf{A}_t$, que maximice el criterio ML.

Dado que para ello se sigue un esquema similar al utilizado para múltiples de las técnicas de normalización presentadas en este trabajo, de modo que se precisa de tres aproximaciones

- Suponiendo que la técnica de normalización seleccionada independiza los vectores compensados, $\hat{\mathbf{x}}_t$, de los entornos básicos, e, se puede considerar que el consiguiente espacio normalizado generado es lo suficientemente homogéneo como para modelarse mediante una mezcla de Gaussianas independiente de posibles nuevos entornos básicos

$$p(\hat{\mathbf{x}}_t) = \sum_{s_{\hat{x}}} p(\hat{\mathbf{x}}_t|s_{\hat{x}})p(s_{\hat{x}}), \tag{8.1}$$

$$p(\hat{\mathbf{x}}_t|s_{\hat{x}}) = \mathcal{N}(\hat{\mathbf{x}}_t; \mu_{s_{\hat{x}}}, \mathbf{\Sigma}_{s_{\hat{x}}}), \tag{8.2}$$

  donde $s_{\hat{x}}$ hace referencia a la correspondiente Gaussiana del modelo normalizado, mientras que $\mu_{s_{\hat{x}}}$, $\mathbf{\Sigma}_{s_{\hat{x}}}$, y $p(s_{\hat{x}})$ son el vector media, la matriz diagonal de covarianzas y la probabilidad a priori asociados a $s_{\hat{x}}$.

- Los vectores de características limpios se modelan mediante una GMM tal y como se indica en las expresiones (5.7) y (5.8).

- Para finalizar, dado el par de Gaussianas s_x y $s_{\hat{x}}$, el vector de características normalizado se aproxima mediante una función lineal multiplicativa del vector de características limpio dependiente del par de Gaussianas s_x y $s_{\hat{x}}$: $\hat{\mathbf{x}}_t \approx \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t$, donde $\mathbf{A}_{s_x,s_{\hat{x}}}$ es la matriz de rotación entre las tramas $\hat{\mathbf{x}}_t$ y $\mathbf{x}_t$ asociada al par de Gaussianas s_x y $s_{\hat{x}}$.

De cara a estimar la matriz de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$, se hace uso de señal estéreo en una fase de entrenamiento previa: $(\mathbf{X}^{Tr}, \hat{\mathbf{X}}^{Tr}) = \{(\mathbf{x}_1^{Tr}, \hat{\mathbf{x}}_1^{Tr}); ...; (\mathbf{x}_t^{Tr}, \hat{\mathbf{x}}_t^{Tr}); ...; (\mathbf{x}_T^{Tr}, \hat{\mathbf{x}}_T^{Tr})\}$, con $t \in [1, T]$, donde $\hat{\mathbf{X}}^{Tr}$ se obtiene tras aplicar la correspondiente técnica de normalización a los vectores de características ruidosos que componen $\mathbf{Y}^{Tr}$, que es, a su vez, la concatenación de los distintos vectores acústicos degradados de los diferentes entornos básicos, e, del corpus de entrenamiento, $\mathbf{Y}^{Tr,e}$. Así pues, a la hora de estimar las correspondientes matrices de rotación para los pares de componentes s_x y $s_{\hat{x}}$, se minimiza con respecto a $\mathbf{A}_{s_x,s_{\hat{x}}}$ el error cuadrático medio, $\xi_{s_x,s_{\hat{x}}}$, asociado a cada par de Gaussianas s_x y $s_{\hat{x}}$, y que se define como (8.3), obteniéndose finalmente la expresión (8.4)

$$\xi_{s_x,s_{\hat{x}}} = \frac{1}{T}\sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})Tra[(\hat{\mathbf{x}}_t^{Tr} - \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t^{Tr})(\hat{\mathbf{x}}_t^{Tr} - \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t^{Tr})^T], \tag{8.3}$$

$$\mathbf{A}_{s_x,s_{\hat{x}}} = \underset{\mathbf{A}_{s_x,s_{\hat{x}}}}{arg\,min}(\xi_{s_x,s_{\hat{x}}}) = \sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})(\hat{\mathbf{x}}_t^{Tr}(\mathbf{x}_t^{Tr})^T)[\sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})(\mathbf{x}_t^{Tr}(\mathbf{x}_t^{Tr})^T)]^{-1}, \tag{8.4}$$

donde $p(s_x|\mathbf{x}_t^{Tr})$ es la probabilidad a posteriori de la Gaussiana del modelo limpio, s_x, dado el vector de características limpio del corpus de entrenamiento, $\mathbf{x}_t^{Tr}$; mientras que $p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})$ es la probabilidad a posteriori de la componente del modelo normalizado $s_{\hat{x}}$, dado el vector acústico del corpus de entrenamiento normalizado $\hat{\mathbf{x}}_t^{Tr}$. Dichas probabilidades se estiman haciendo uso de las expresiones (5.7) y (5.8), para el primero de los casos (8.5), y (8.1) y (8.2) para el segundo (8.6). En el Anexo 8.4 en este mismo Capítulo se puede consultar el desarrollo teórico completo para obtener la expresión (8.4) a partir de (8.3).

$$p(s_x|\mathbf{x}_t^{Tr}) = \frac{p(\mathbf{x}_t^{Tr}|s_x)p(s_x)}{\sum_{s_x} p(\mathbf{x}_t^{Tr}|s_x)p(s_x)}, \tag{8.5}$$

$$p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr}) = \frac{p(\hat{\mathbf{x}}_t^{Tr}|s_{\hat{x}})p(s_{\hat{x}})}{\sum_{s_{\hat{x}}} p(\hat{\mathbf{x}}_t^{Tr}|s_{\hat{x}})p(s_{\hat{x}})}. \tag{8.6}$$

Tal y como ya se ha indicado previamente, a partir del conjunto de matrices de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$, se ha de seleccionar para cada vector de características normalizado, $\hat{\mathbf{x}}_t$, aquélla, $\mathbf{A}_t$, que maximice el criterio ML en la fase de decodificación. Para ello se generan unos modelos extendidos del mismo modo que se realiza para el método *augMented stAte space acousTic modEl*, MATE, [Miguel *et al.*, 2005] [Miguel *et al.*, 2006], cuya principal motivación consiste en encontrar una serie de modelos acústicos que sean capaces de representar la variabilidad inter-locutor basándose en el algoritmo VTLN, *Vocal Tract Length Normalization*, [Lee and Rose, 1998]. Así pues, a continuación se indica cómo se ha de llevar a cabo tanto la extensión de los modelos acústicos, así como la generalización que es preciso proponer para el proceso de decodificación, ya que la elección del modelo extendido o, lo que es equivalente, $\mathbf{A}_t$, para cada vector de características normalizado que se pretende reconocer se realiza en ese momento a partir del algoritmo ML.

Considerando que los modelos acústicos del espacio referencia están compuestos por HMMs, cada estado se expande a partir de las distintas matrices de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$. De este modo, un estado original q ($q \in [1,Q]$) se transformará en tantos nuevos como pares de Gaussianas s_x $s_{\hat{x}}$ haya, N, y vendrán identificados por las variables $(q,n), n \in [1,N]$. Por su parte, los parámetros que componen las pdfs asociadas a cada uno de los estados expandidos se generarán a partir de los de la pdf del estado original q haciendo uso de las correspondientes matrices de rotación, $\mathbf{A}_{s_x,s_{\hat{x}}}$, incluyendo de este modo en ellos la deformación propia de la rotación. Así, asumiendo que las distintas pdfs de los modelos acústicos del espacio de referencia están compuestas por GMMs, una componente s_q del estado original q, $\mathcal{N}(\mathbf{x}_t;\mu_{s_q},\mathbf{\Sigma}_{s_q})$, donde μ_{s_q} y $\mathbf{\Sigma}_{s_q}$ son el vector de medias y la matriz de covarianza de la correspondiente Gaussiana, se transformará, para el estado extendido (q,n), en la componente $s_{q,n}$ del siguiente modo: $\mathcal{N}(\mathbf{x}_t;\mathbf{A}_n\mu_{s_q},\mathbf{A}_n\mathbf{\Sigma}_{s_q}\mathbf{A}_n^T)$. En cuanto a las probabilidades a priori de las distintas componentes expandidas, éstas quedan inalteradas con respecto a las de las correspondientes Gaussianas del estado inicial q ($p(s_{q,n}) = p(s_q)$). Ya para finalizar, las probabilidades de transición entre estados expandidos, $\mathbf{\Pi}$, que se definen como (8.7), se pueden estimar a partir del algoritmo EM [Miguel *et al.*, 2006], aunque en los experimentos que se desarrollarán en este trabajo no se va a realizar de este modo, considerándolas equiprobables entre cualquier par de estados. Sin emabrgo, sí se pretende tratar este término en futuros trabajos ya que, en el fondo, supone aprender la evolución temporal de los pares de Gaussianas, en este caso representantes de los espacios limpio y normalizado, y esto, al menos a priori y tras unas pruebas preliminares, parece que podría proporcionar una interesante mejora.

$$\mathbf{\Pi} = \{\pi_{q',n',q,n}\}_{q'=1,n'=1,q=1,n=1}^{Q,N,Q,N}, \tag{8.7}$$

donde $\pi_{q',n',q,n}$ es la probabilidad de transición del estado (q', n') al (q, n).

Obsérvese que, desde el punto de vista de la generación de los vectores de características, los nuevos modelos acústicos expandidos pueden verse como un proceso de producción de los mismos muy flexible, por cuanto se pueden obtener secuencias de vectores acústicos generadas tras considerar distintos grados de rotación.

Una vez que se han expandido los estados de los modelos acústicos del espacio de referencia, de modo que, como ya se ha comentado, por cada uno asociado a estos últimos se obtienen N nuevos correspondientes a las distintas matrices de rotación, es necesario adaptar el algoritmo de decodificación. Esto es debido a que es en dicho proceso, tal y como ya se ha adelantado, donde se pretende estimar, de entre las N diferentes matrices $\mathbf{A}_{s_x,s_{\hat{x}}}$, la correspondiente a cada vector de características normalizado que se pretende decodificar para el instante de tiempo t, $\mathbf{A}_t$. De este modo, la variable $\phi_{q,n}(t)$, que es el *score* del estado (q, n) para el instante de tiempo t dado el vector de características $\mathbf{w}_t$, y que constituye la base para ejecutar el algoritmo de Viterbi, se calculará como

$$\phi_{q,n}(t) = \underset{n',q'}{arg\,max}(\phi_{q',n'}(t-1) \cdot \pi_{q',n',q,n} \cdot p(\mathbf{w}_t|n,q)), \tag{8.8}$$

donde $p(\mathbf{w}_t|n, q)$ es el *score* del vector de características $\mathbf{w}_t$, dado el estado extendido (q, n), $(p(\mathbf{w}_t|n,q) = \sum_{s_{q,n}} \mathcal{N}(\mathbf{w}_t; \mathbf{A}_n\mu_{s_q}, \mathbf{A}_n\mathbf{\Sigma}_{s_q}\mathbf{A}_n^T)p(s_{q,n}))$. Se puede observar como la expresión recursiva anterior es similar a la expuesta en [Miguel *et al.*, 2005], aunque, en este caso, la introducción de la deformación propia de la rotación se produce en los modelos acústicos en lugar de en los vectores de características, lo que no es exactamente lo mismo tal y como se ha indicado anteriormente, ya que al transformar las matrices de covarianzas se incluye indirectamente la normalización Jacobiana en el modelo [Pitz and Ney, 2005], hecho este que no se hubiera producido si únicamente se normalizaran los vectores de características, generando por tanto un cierto desajuste.

Así pues, y a modo de resumen, se puede decir que las técnicas híbridas a partir del cálculo de matrices de rotación dependientes de GMMs, tras estimar una serie de matrices, $\mathbf{A}_{s_x,s_{\hat{x}}}$, que modelan la variabilidad rotacional, selecciona para cada vector de características normalizado que se pretende decodificar aquélla, $\mathbf{A}_t$, que maximiza el criterio ML. Adviértase que el hecho de elegir dicho criterio no es gratuito, ya que, a fin de cuentas, los sistemas de RAH se basan en él.

Tal y como se ha presentado teóricamente la base de las técnicas híbridas basadas en el cálculo de matrices de rotación dependientes de GMMs, el método de normalización previo necesario podría ser genérico, ya que el algoritmo final es independiente de él. Sin embargo, en la experimentación propuesta para este trabajo se han empleado únicamente los algoritmos MEMLIN y MEMLIN MP, dando lugar a sendas técnicas híbridas.

Cabe destacar que este tipo de técnicas híbridas posee la importante ventaja de que modifica los modelos acústicos de un modo no supervisado e independiente de la tarea de reconocimiento; condicionantes estos difícilmente asumibles en muchos casos cuando se emplean técnicas básicas de adaptación de modelos acústicos como MAP, MLLR...

8.2. Técnicas Híbridas Supervisadas Basadas en Reentrenamiento.

A la hora de conjugar las técnicas de normalización de vectores de características con las de adaptación de modelos acústicos, además de la opción comentada anteriormente, esto es, los algoritmos híbridos basados en el cálculo de matrices de rotación dependientes de GMMs, la solución más directa, así como la que más recursos precisa, consiste en decodificar los vectores acústicos ruidosos compensados mediante el método de normalización elegido para la ocasión haciendo uso de los correspondientes modelos acústicos representantes del espacio normalizado y reentrenados de forma supervisada. El fundamento de este tipo de fusión es el mismo que el de la opción anteriormente comentada, esto es: considerar que la normalización propuesta, sea cual sea, no es perfecta, de modo que a pesar de que trate de proyectar los vectores acústicos ruidosos desde un cierto entorno básico hasta el espacio de referencia (normalmente el limpio), esto no se da a la perfección, surgiendo de este modo un nuevo espacio pseudo-limpio, al que se denomina normalizado, y que por tanto no queda perfectamente representado con los modelos acústicos de referencia. Para evitar este desajuste y obtener los consiguientes modelos acústicos que representen adecuadamente el nuevo espacio de proyección, se propone, en este caso, transformar los modelos acústicos de referencia a partir de la señal del corpus de entrenamiento ruidoso previamente normalizada; nótese que dicha compensación ha de ser la misma con que posteriormente se transformará la señal que se pretenda decodificar. La diferencia de esta nueva solución con respecto a la mostrada anteriormente consiste en que los nuevos modelos acústicos se obtienen de modo supervisado haciendo uso de las técnicas clásicas de entrenamiento. De este modo, si el corpus de entrenamiento ruidoso es suficientemente amplio, se puede emplear el algoritmo ML, pero si éste no es el caso se tendría que recurrir a técnicas clásicas de adaptación de modelos acústicos como MAP, MLLR..., según las necesidades de la aplicación y las limitaciones de los datos disponibles. En este trabajo se decidió emplear el algoritmo ML.

A modo de resumen se incluye la Figura 8.3, en la que se reflejan los distintos procesos que se han de seguir para llevar a cabo la experimentación con las técnicas híbridas supervisadas basadas en reentrenamiento. Del mimso modo que en la Sección 8.1, y en previsión de emplear técnicas de normalización de vectores de características que precisen de una fase previa de entrenamiento, se ha incluido un bloque para la misma que, en principio y en un caso general, no siempre sería necesario. En el bloque denominado "Adaptación HMM" se obtienen los nuevos modelos acústicos (HMM del espacio normalizado) a partir de los de referencia mediante el algoritmo correspondiente, y que, posteriormente y ya en la fase de decodificación, se emplearán para reconocer los vectores de características ruidosos normalizados.

Al igual que en la sección anterior, el esquema propuesto para las técnicas híbridas supervisadas basadas en reentrenamiento es independiente del método de normalización aplicado, pudiéndose ser éste cualquiera. Sin embargo, en este trabajo se han empleado únicamente los algoritmos MEMLIN y MEMLIN MP.

8.3. Resultados con la Base de Datos *SpeechDat Car* en Español.

La experimentación comparativa de las técnicas híbridas tratadas en las Secciones 8.1 y 8.2 se llevó a cabo con la base de datos *SpeechDat Car* en español, utilizándose los corpora de entrenamiento de los diversos entornos básicos previamente definidos para realizar los distintos procesos

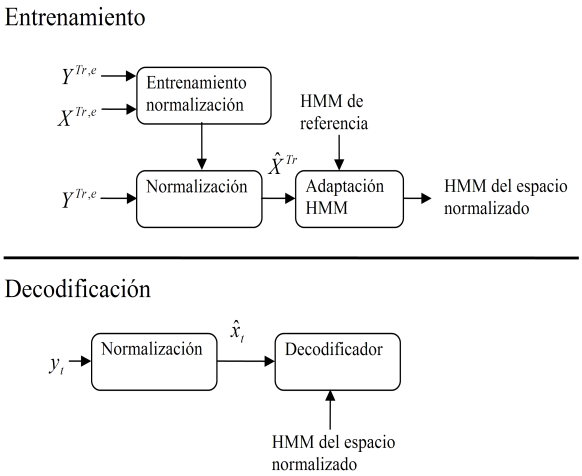


Figura 8.3: Esquema gráfico de las técnicas híbridas supervisadas basadas en reentrenamiento. Se incluye tanto la fase de entrenamiento como la de decodificación. La primera de ellas está compuesta por tres bloques: "Entrenamiento normalización", que se ha incluido en previsión de utilizar técnicas de normalización de vectores de características que la precisen. Por su parte, el sistema de "Normalización" proporciona las señales limpias estimadas a partir de las correspondientes degradadas. Finalmente el bloque "Adaptación HMM" calcula los nuevos modelos acústicos asociados al espacio normalizado a partir de los limpios y de la señal del corpus de entrenamiento degradado previa compensación. Dichos modelos se emplean finalmente para reconocer los vectores de características normalizados en el bloque de "Decodificación".

de entrenamiento, necesarios en esta ocasión tanto para las técnicas de normalización de vectores de características seleccionadas, como para generar los nuevos modelos acústicos. Por otra parte, en todas las experimentaciones propuestas en esta Sección se aplicará en última instancia el método CMS a los vectores de acústicos que se pretendan reconocer, teniéndolo también esto en cuenta, como es natural, a la hora de obtener los distintos modelos acústicos adaptados. Asimismo se empleará la *parametrización estándar ETSI* y modelos acústicos de palabras, pudiéndose, de este modo, consultar los resultados de referencia correspondientes en la Tabla 4.4. Por otra parte, y en aras de establecer comparaciones de un modo más justo, todas las técnicas de normalización de vectores de características se aplicarán únicamente sobre los coeficientes estáticos de los mismos, tal y como hasta ahora se ha venido realizando, calculando las correspondientes derivadas posteriormente.

8.3.1. Resultados de las técnicas híbridas a partir del cálculo de matrices de rotación dependientes de GMMs.

Antes de presentar los resultados obtenidos tras aplicar las técnicas híbridas a partir del cálculo de matrices de rotación dependientes de GMMs es recomendable revisar, de cara a completar los parámetros que definen la experimentación, la Figura 8.2, donde se indican los dos pasos que componen el algoritmo. Así, primeramente, y en una fase de entrenamiento previo ("Entrenamiento"), se estiman los vectores de desplazamiento, $\mathbf{r}_{s_x,s_y^e}$, y el modelado de la probabilidad entre Gaussianas, $p(s_x|\mathbf{y}_t, e, s_y^e)$, necesarios para las técnicas MEMLIN y MEMLIN MP, para lo que, tal y como se ha comentado en las Secciones 5.3 y 7.3, se hace uso de los vectores de caracetrísticas del corpus de entrenamiento estéreo de la base de datos. En caso de la técnica MEMLIN MP la correspondiente GMM con que se modelan los vectores de características ruidosos para cada par de

Gaussianas s_x s_y^e, se genera con dos componentes. Asimismo, y en la misma fase de entrenamiento, se calculan, como ya se indicó en la Seccion 8.1, las matrices de rotación, $\mathbf{A}_{s_x,s_{\hat{x}}}$, haciendo uso de los vectores de características completos, esto es, considerando los coeficientes dinámicos y tras evaluar la técnica CMS. En lo sucesivo, y salvo que se indique lo contrario, se calcularán siempre 16 matrices $\mathbf{A}_{s_x,s_{\hat{x}}}$, que provienen de modelar el espacio limpio y normalizado con 4 componentes. El segundo paso, denominado "Decodificación", consiste en reconocer las tramas compensadas con el algoritmo MEMLIN haciendo uso de los modelos expandidos construidos a partir de los modelos acústicos limpios y las distintas matrices de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$.

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	SPLICE ME 64	2.48	6.52	3.92	7.64	9.06	5.08	12.93	6.25	74.08
CLK	MEMLIN 128	2.00	6.26	3.78	7.27	8.48	5.24	11.90	5.89	75.79
CLK	MEMLIN MP 64	1.81	4.80	1.82	5.76	6.01	3.81	6.80	4.23	83.89
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	MEMLIN 64	2.19	3.95	2.10	3.26	3.24	1.90	2.38	2.86	90.54
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	MEMLIN MP 64	1.91	3.86	1.68	2.88	2.96	1.27	2.04	2.54	92.07

Cuadro 8.1: Mejores resultados con la base de datos *SpeechDat Car* en español para las técnicas SPLICE ME, MEMLIN, MEMLIN MP y las técnicas híbridas a partir del cálculo de matrices de rotación dependientes de GMMs basadas en los algoritmos de compensación MEMLIN y MEMLIN MP en términos de WER (%). Se presentan los resultados para los diferentes entornos básicos (E1,..., E7) utilizando la *parametrización estándar ETSI* y modelos acústicos para las palabras del vocabulario (dígitos) generados a partir del corpus de señal limpia (CLK en la columna de "Entrenamiento"), o extendidos a partir de los anteriores haciendo uso de 16 matrices de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$, (CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$ en la columna de "Entrenamiento"). La columna marcada como "Reconocimiento" hace referencia a la señal empleada para reconocer, que será la ruidosa normalizada con la técnica SPLICE ME, MEMLIN o MEMLIN MP. Junto al nombre de los diferentes métodos aparece el número de Gaussianas con que se modelaron los correspondientes espacios para la fase de normalización (el limpio y los asociados a los distintos entornos básicos). Por su parte, para el método MEMLIN MP se modelan los vectores de características asociados a cada par de Gaussianas s_x y s_y^e con dos componentes. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

En la Tabla 8.1 se pueden apreciar los mejores resultados para las técnicas SPLICE ME ("Entrenamiento" CLK, "Reconocimiento" SPLICE ME), MEMLIN ("Entrenamiento" CLK, "Reconocimiento" MEMLIN), MEMLIN MP ("Entrenamiento" CLK, "Reconocimiento" MEMLIN MP) y las técnicas híbridas a partir del cálculo de matrices de rotación dependientes de GMMs basadas en los algoritmos MEMLIN y MEMLIN MP ("Entrenamiento" CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$, "Reconocimiento" MEMLIN o MEMLIN MP, respectivamente). Los tres primeros experimentos se incluyen a modo de comparación, puesto que, bajo estas condiciones concretas de experimentación, no se habían presentado aún las correspondientes tasas de RAH. En todos los casos, junto al nombre de la técnica de normalización de vectores de características, se incluye el número de componentes que conforman las GMMs con que se modelan los entornos básicos ruidosos y el espacio limpio (se realizó un barrido con 4, 8, 16, 32, 64 y 128 componentes). En cuanto a los métodos híbridos, los modelos acústicos expandidos se construirán a partir de los limpios haciendo uso de las correspondientes matrices de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$. Asimismo se incluye en la Tabla 8.1, además del WER medio, MWER, la mejora media de WER, MIMP, en tanto por ciento, y calculadas del mismo modo que ya se explicó en el Capítulo 5.4 (ver expresión (5.29)).

Por otra parte, y para determinar si se puede afirmar o no que los resultados anteriores son

estadísticamente significativos, se recurre a la prueba de hipótesis estadística *z-test*. En esta ocasión se comparan las técnicas MEMLIN y MEMLIN MP con sus respectivas versiones híbridas basadas en el cálculo de matrices de rotación dependientes de GMMs bajo la base de datos *SpeechDat Car* en español. Se puede observar que el valor del estadístico W, w, para el primero de los casos es $w = 7{,}91 > 1{,}96$, por lo que la mejora del algoritmo híbrido asociado a la técnica MEMLIN en esta Subsección se puede considerar independiente de la base de datos con un intervalo de confianza del 95 %. Por otra parte, para el segundo de los casos, en el que se cotejan los resultados obtenidos con la técnica MEMLIN MP y su correspondiente algoritmo híbrido, se tiene que $w = 4{,}99 > 1{,}96$, con lo que se puede considerar igualmente que la diferencia de comportamiento de estas dos últimas técnicas es estadísticamente significativa con un intervalo de confianza del 95 %. De todos modos es conveniente recordar, igual que se ha venido haciendo hasta el momento, que estos resultados se han de considerar con cautela dadas las limitaciones de la propia prueba, ya comentadas en la Sección 4.2.

A la luz pues de los resultados presentados en la Tabla 8.1 se puede concluir que, teniendo en cuenta únicamente los mejores resultados medios para las distintas técnicas comparadas y para todos y cada uno de los entornos básicos, los métodos híbridos a partir del cálculo de matrices de rotación dependientes de GMMs aportan una importante y estadísticamente significativa mejora con respecto a los resultados obtenidos con los algoritmos SPLICE ME, MEMLIN y MEMLIN MP. Y no sólo eso, sino que además su comportamiento es más satisfactorio que el obtenido tras reconocer la señal ruidosa con modelos acústicos entrenados bajo las mismas condiciones: MWER de 4.63 % ("Entrenamiento" HF, "Reconocimiento" HF en la Tabla 4.4) y MWER de 3.42 % ("Entrenamiento" HF†, "Reconocimiento" HF en la Tabla 4.4). Todo ello teniendo en cuenta que en este caso se están adaptando los modelos acústicos de un modo no supervisado y tras estimar únicamente 16 matrices de rotación, lo que supone calcular muchos menos datos que en el caso del reentrenamiento.

Llegados a este punto se podría pensar que las técnicas híbridas a partir del cálculo de matrices de rotación dependientes de GMMs son similares, conceptualmente hablando, al algoritmo de adaptación de modelos acústicos MLLR, en el que se puede considerar que se modifican las medias y varianzas del modelado acústico mediante un vector de desplazamiento y una matriz de rotación. Para comparar las prestaciones de las dos técnicas de un modo justo, esto es, no supervisado, se adaptaron los modelos limpios mediante el algoritmo MLLR, proyectándolos sobre el espacio ruidoso haciendo uso de las trascripciones provinientes de reconocer la señal ruidosa con los modelos acústicos limpios. Bajo esas condiciones de experimentación la técnica MLLR obtiene una mejora media de 78.77 %, aún lejana del 92.07 % lograda con la técnica híbrida a partir del cálculo de matrices de rotación dependientes de GMMs haciendo uso de el método MEMLIN MP.

Asimismo, si se considera la mejora media de WER para los métodos anteriormente comentados en función del número de Gaussianas con que se modela cada entorno básico (Figura 8.4), se puede apreciar que la mejora de comportamiento observado en la Tabla 8.1 es extensible a cualquier número de componentes, de modo que, por ejemplo, si se aplica la técnica MEMLIN modelando cada entorno básico con 4 Gaussianas se obtiene un MIMP de 62.57 %, mientras que los algoritmos híbridos a partir del cálculo de matrices de rotación dependientes de GMMs haciendo uso de los métodos MEMLIN y MEMLIN MP alcanzan, bajo las mismas condiciones, unos valores sensiblemente mayores, 83.55 % y 88.49 %, respectivamente. Del mismo modo, resulta interesante hacer notar que en este tipo de técnicas híbridas, los resultados son menos dependientes del número de Gaussianas con que se modelan los entornos básicos, de modo que con un número reducido de las mismas ya se pueden alcanzar importantes mejoras.

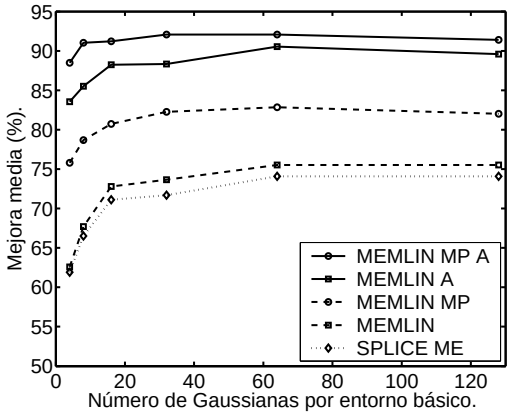


Figura 8.4: Mejora media de WER tras aplicar las técnicas SPLICE ME (línea punteada con diamantes blancos), MEMLIN (línea discontinua con cuadrados blancos), MEMLIN MP (línea discontinua con círculos blancos), la técnica híbrida a partir del cálculo de matrices de rotación dependientes de GMMs y basada en el algoritmo MEMLIN, identificada como "MEMLIN A" (línea continua con cuadrados blancos), y, ya por último, la técnica híbrida a partir del cálculo de matrices de rotación dependientes de GMMs y basada en el algoritmo MEMLIN MP y nombrada como "MEMLIN MP A" (línea continua con círculos blancos). En todos los casos se representan en función del número de Gaussianas por entorno básico empleado.

8.3.2. Resultados de las técnicas híbridas supervisadas basadas en reentrenamiento.

Paso previo a la presentación de los resultados de RAH obtenidos al decodificar los vectores de características normalizados aplicando modelos acústicos reentrenados de modo supervisado, es recomendable revisar la Figura 8.3, donde se indican los distintos pasos que se han de seguir a tal efecto. Así, primeramente es necesario, en una fase de entrenamiento previa, estimar, si es que fuera preciso, los diversos parámetros necesarios para la técnica de normalización seleccionada, para lo que, en general y para las técnicas presentadas en este trabajo, se hace uso de un corpus de entrenamiento estéreo. Asimismo, y todavía en la misma fase de entrenamiento, se estiman los modelos acústicos que representan al espacio normalizado a partir de los vectores de caracteríistias ruidosos del corpus de entrenamiento, previamente compensados. Ya por último, y en el segundo paso, que se identifica en la Figura 8.3 como "Decodificación", se normalizan los vectores de características ruidosos, reconociéndolos posteriormente con los modelos acústicos adaptados. A continuación se presentan los resultados de RAH obtenidos en función de la técnica de normalización empleada, que podrá ser MEMLIN o MEMLIN MP.

En la Tabla 8.2 se pueden apreciar los mejores resultados cuando se seleccionan como técnicas de normalización los algoritmos MEMLIN y MEMLIN MP, ("Entrenamiento" HF MEMLIN, "Reconocimiento" MEMLIN y "Entrenamiento" HF MEMLIN MP, "Reconocimiento" MEMLIN MP, respectivamente). Asimismo, y a modo de comparación, se vuelven a incluir los resultados alcanzados cuando se reconoce la señal limpia con modelos acústicos limpios ("Entrenamiento" CLK, "Reconocimiento" CLK), la señal ruidosa con modelos acústicos generados a partir de todo el corpus de entrenamiento ruidoso ("Entrenamiento" HF, "Reconocimiento" HF), y la señal ruidosa con modelos acústicos específicos para cada entorno básico ("Entrenamiento" †HF, "Reconocimiento" HF). Del mismo modo, allí donde fuera procedente, junto al nombre de la técnica de normalización empleada, se incluye el número de componentes que conforman las GMMs con que se modelan los entornos básicos ruidosos y el espacio limpio (se realizó un barrido con 4, 8, 16, 32, 64 y 128

Entrenamiento	Reconocimiento	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%
CLK	CLK	0.95	2.32	0.70	0.25	0.57	0.32	0.00	0.91	–
HF	HF	3.81	6.86	3.50	3.76	4.96	4.44	3.06	4.63	81.93
† HF	HF	1.14	4.37	1.68	2.13	2.10	2.06	23.13	3.42	87.91
HF MEMLIN 64	HF MEMLIN 64	0.57	3.09	1.26	2.51	1.43	1.27	0.34	1.67	96.33
HF MEMLIN MP 128	HF MEMLIN MP 128	0.57	2.83	0.98	2.01	1.43	1.11	0.00	1.47	97.27

Cuadro 8.2: Mejores resultados con la base de datos *SpeechDat Car* en español para las técnicas híbridas supervisadas basadas en reentrenamiento a partir de los algoritmos MEMLIN y MEMLIN MP en términos de WER (%). Se presentan los resultados para los diferentes entornos básicos (E1,..., E7) utilizando la *parametrización estándar ETSI* y modelos acústicos para las palabras del vocabulario (dígitos) adaptados al espacio normalizado a partir del corpus de entrenamiento de señal ruidosa previamente compensado con la correspondiente técnica (HF MEMLIN o HF MEMLIN MP según corresponda en la columna de "Entrenamiento"). La columna identificada como "Reconocimiento" hace referencia a la señal empleada para decodificar, que será la ruidosa normalizada con la técnica MEMLIN o MEMLIN MP. Junto al nombre de los diferentes métodos aparece el número de Gaussianas con que se modelaron los correspondientes espacios (el limpio y los asociados a los distintos entornos básicos). Adicionalmente se emplean 2 componentes para modelar los vectores de características ruidosos asociados a cada par de Gaussianas s_x y s_y^e (modelado de la probabilidad entre Gaussianas basado en GMMs para la técnica MEMLIN MP). A modo de comparación se presentan nuevamente los resultados de RAH alcanzados cuando se reconoce la señal limpia con modelos acústicos limpios ("Entrenamiento" CLK, "Reconocimiento" CLK), o la señal ruidosa con modelos acústicos generados con todo el corpus de entrenamiento degradado, o bien con modelos acústicos específicos para cada entorno básico ("Entrenamiento" HF, "Reconocimiento" HF, o "Entrenamiento" †HF, "Reconocimiento" HF, respectivamente). Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

componentes). Adicionalmente, y por completar los parámetros que definen la experimentación, se emplean 2 componentes para modelar los vectores de características ruidosos asociados a cada par de Gaussianas: s_x y s_y^e, (modelado de la probabilidad entre Gaussianas basado en GMMs para la técnica MEMLIN MP). Asimismo se incluye en la Tabla 8.2, además del WER medio, MWER, la mejora media de WER, MIMP, en tanto por ciento, que se calcula del mismo modo que ya se explicó en el Sección 5.4 (ver expresión (5.29)).

A pesar de que ya se pueden intuir los resultados, resulta conveniente, al igual que se ha venido haciendo hasta el momento, y de cara a determinar si se puede afirmar o no que los resultados anteriores son estadísticamente significativos, realizar la prueba de hipótesis estadística *z-test* para los mismos. En esta ocasión se comparan las técnicas MEMLIN y MEMLIN MP con sus respectivas versiones híbridas supervisadas basadas en reentrenamiento bajo la base de datos *SpeechDat Car* en español. Se puede observar que el valor del estadístico W, w, para el primero de los casos es $w = 11{,}81 > 1{,}96$, por lo que en esta ocasión la mejora del algoritmo híbrido supervisado asociado a la técnica MEMLIN se puede considerar independiente de la base de datos con un intervalo de confianza del 95 %. Del mismo modo para el segundo de los casos, en el que se cotejan los resultados obtenidos con la técnica MEMLIN MP y su correspondiente algoritmo híbrido supervisado basado en reentrenamiento, se tiene que $w = 8{,}56 > 1{,}96$, con lo que se puede considerar del mismo modo que para el caso anterior que la diferencia de comportamiento de estas dos últimas técnicas es estadísticamente significativa con un intervalo de confianza del 95 %. De todos modos es conveniente recordar que estos resultados se han de tratar con la debida cautela dadas las limitaciones de la propia prueba, ya comentadas en la Sección 4.2.

A partir de los resultados presentados en la Tabla 8.2 se puede concluir que, teniendo en cuenta únicamente los mejores resultados medios para los distintos casos y para todos y cada uno de los entornos básicos, decodificar los vectores de características normalizados con las técnicas MEMLIN y MEMLIN MP haciendo uso de modelos acústicos que representen los correspondientes espacios normalizados y obtenidos de modo supervisado (técnicas híbridas supervisadas basadas en reentrenamiento) aporta, para los dos casos que se han tratado, una importante y estadísticamente significativa mejora con respecto a reconocer las señales compensadas con los modelos acústicos limpios. De la misma manera la comparación es igualmente satisfactoria si se comparan los resultados alcanzados cuando las señales degradadas se decodifican con modelos acústicos generados con todo el corpus de entrenamiento degradado ("Entrenamiento" HF, "Reconocimiento" HF) o específicamente para cada entorno básico ("Entrenamiento" † HF, "Reconocimiento" HF). Esto es debido a que tras compensar la señal ruidosa, el espacio generado es mucho más compacto y homogéneo, haciendo que el entrenamiento sea más satisfactorio que si se realizara directamente sobre el entorno ruidoso, siempre mucho más heterogéneo. Asimismo, si se estudian los resultados logrados cuando se varía el número de Gaussianas con que se modela cada entorno básico para las técnicas de normalización (al igual que para los casos anteriores se realizó un barrido con 4, 8, 16, 2, 64 y 128 componentes), se puede constatar que no difieren de un modo estadísticamente significativo. Así, por ejemplo, si la técnica MEMLIN se aplica con 4 Gaussianas se obtiene un MIMP de 94.63 %, mientras que si se consideran modelos con el mismo número de componentes para el método MEMLIN MP, el MIMP alcanzado es 95.14 %. Es por esta falta de significancia entre los resultados obtenidos por lo que no se ha incluido en esta ocasión una gráfica alusiva en la que se incluyeran las mejoras medias (MIMP) para distintos números de Gaussianas con que se modelaran los entornos básicos.

8.4. Anexo I.

En este Anexo se incluye el desarrollo teórico necesario para estimar las matrices de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$ con que se representa la proyección lineal entre los datos de un espacio fuente, que en general se corresponde con el limpio, y los correspondientes al espacio objetivo, que en este caso se construye a partir de los vectores de características del espacio ruidoso normalizados mediante algún tipo de técnica de compensación de vectores de características. Sea pues un corpus de entrenamiento estéreo $(\mathbf{X}, \hat{\mathbf{X}}) = \{(\mathbf{x}_1, \hat{\mathbf{x}}_1); ...; (\mathbf{x}_t, \hat{\mathbf{x}}_t); ...; (\mathbf{x}_T, \hat{\mathbf{x}}_T)\}$, con $t \in [1, T]$; nótese que, por simplificar la notación se ha eliminado el índice Tr para indicar que se trata del corpus de entrenamiento, tal y como sí estaba recogido en la Sección 8.1.1. De este modo, el error cuadrático medio asociado a cada par de Gaussianas para la técnica propuesta, $\xi_{s_x,s_{\hat{x}}}$, se define como

$$\xi_{s_x,s_{\hat{x}}} = \frac{1}{T}\sum_t p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t)Tra[(\hat{\mathbf{x}}_t - \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t)(\hat{\mathbf{x}}_t - \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t)^T]. \tag{I.1}$$

Teniendo en cuenta ciertas propiedades del cálculo matricial, se puede observar, antes de llevar a cabo la minimización de $\xi_{s_x,s_{\hat{x}}}$, que

$$\begin{aligned}(\hat{\mathbf{x}}_t - \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t)(\hat{\mathbf{x}}_t - \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t)^T = \hat{\mathbf{x}}_t(\hat{\mathbf{x}}_t)^T - \hat{\mathbf{x}}_t(\mathbf{x}_t)^T(\mathbf{A}_{s_x,s_{\hat{x}}})^T \\ -\mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t(\hat{\mathbf{x}}_t)^T + \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t(\mathbf{x}_t)^T(\mathbf{A}_{s_x,s_{\hat{x}}})^T.\end{aligned} \tag{I.2}$$

A la hora de estimar la matriz de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$ se procede a la minimización de la expresión (I.1) con respecto a $\mathbf{A}_{s_x,s_{\hat{x}}}$ haciendo uso de (I.2)

$$\mathbf{0} = \frac{\delta \xi_{s_x, s_{\hat{x}}}}{\delta \mathbf{A}_{s_x, s_{\hat{x}}}} = \frac{1}{T} \sum_t p(s_x|\mathbf{x}_t) p(s_{\hat{x}}|\hat{\mathbf{x}}_t)$$
$$\frac{\delta}{\delta \mathbf{A}_{s_x, s_{\hat{x}}}} [Tra[\hat{\mathbf{x}}_t(\hat{\mathbf{x}}_t)^T - \hat{\mathbf{x}}_t(\mathbf{x}_t)^T (\mathbf{A}_{s_x, s_{\hat{x}}})^T$$
$$-\mathbf{A}_{s_x, s_{\hat{x}}} \mathbf{x}_t(\hat{\mathbf{x}}_t)^T + \mathbf{A}_{s_x, s_{\hat{x}}} \mathbf{x}_t(\mathbf{x}_t)^T (\mathbf{A}_{s_x, s_{\hat{x}}})^T]]. \tag{I.3}$$

O, lo que es lo mismo

$$\mathbf{0} = \frac{1}{T} \sum_t p(s_x|\mathbf{x}_t) p(s_{\hat{x}}|\hat{\mathbf{x}}_t)(-\hat{\mathbf{x}}_t(\mathbf{x}_t)^T - \hat{\mathbf{x}}_t(\mathbf{x}_t)^T + \mathbf{A}_{s_x, s_{\hat{x}}} \mathbf{x}_t(\mathbf{x}_t)^T + \mathbf{A}_{s_x, s_{\hat{x}}} \mathbf{x}_t(\mathbf{x}_t)^T). \tag{I.4}$$

Finalmente, se obtiene la expresión óptima para $\mathbf{A}_{s_x, s_{\hat{x}}}$ despejando convenientemente

$$\mathbf{A}_{s_x, s_{\hat{x}}} = \sum_t p(s_x|\mathbf{x}_t) p(s_{\hat{x}}|\hat{\mathbf{x}}_t) \hat{\mathbf{x}}_t(\mathbf{x}_t)^T [\sum_t p(s_x|\mathbf{x}_t) p(s_{\hat{x}}|\hat{\mathbf{x}}_t) \mathbf{x}_t(\mathbf{x}_t)^T]^{-1}, \tag{I.5}$$

que coincide con la expresión (8.4) presentada en la Sección 8.1.1

8.5. Anexo K.

9

Resultados con la base de datos *Aurora2*

Tal y como ya se ha adelantado en el Capítulo 4, si bien la base de datos *SpeechDat Car* proporciona, a nivel de RAH, unos resultados más interesantes por cuanto se ha grabado en ambientes hostiles reales, el corpus *Aurora2* posee la intangible ventaja de ser prácticamente considerado en la actualidad como un estándar de facto a la hora de comparar distintas técnicas de robustez. De esta manera resulta, a pesar de sus limitaciones y de cara a presentar resultados a la comunidad científica, muy útil incluir siempre la experimentación con esta base de datos.

De cara a realizar posteriores análisis de los resultados obtenidos a lo largo de la experimentación, resulta conveniente recordar brevemente ciertos aspectos del corpus *Aurora2*. De este modo, conviene recordar que se generó a partir de la base de datos de dígitos aislados y conectados en inglés *TIDigits*, a la que se le añadió artificialmente tanto distintos tipos de ruidos aditivos con diferentes SNRs (20dB, 15dB, 10dB, 5dB, 0dB y -5dB), como, en algunos casos, distorsión convolucional, de modo que con ello se busca simular algunos de los escenarios más característicos propios del área de las telecomunicaciones, a saber: metro *subway*, muchedumbre *babble*, coche *car*, salón de exhibiciones *exhibition hall*, restaurante *restaurant*, calle *street*, aeropuerto *airport* y estación de tren *train station*.

A pesar de que la experimentación completa típica comprende, como se puede apreciar en la Sección 4.3, de dos apartados, según si el corpus de entrenamiento consta únicamente de señal limpia (*clean training*) o de una combinación de limpia y degradada (*multicondition training*), en los experimentos que se van a llevar a cabo en este Capítulo únicamente se tendrá en cuenta la primera de las opciones puesto que se ha desestimado la posibilidad de emplear el corpus de entrenamiento multi-condición para generar supervisadamente nuevos modelos acústicos, aunque sí se usará como la parte degradada de la base de datos estéreo que se precisa para estimar los distintos parámetros de las técnicas correspondientes.

Por su parte, cabe recordar asimismo que los tres *sets* en que se haya dividido el corpus de reconocimiento responden, en cierto modo, a otras tantas situaciones que, en conjunto, pueden dar una idea aproximada del comportamiento general de las técnicas que se pretendan comparar. Así, el *set* A comprende la distorsión convolucional y los mismos tipos de ruidos, no así las SNRs, que aparecen en el corpus de entrenamiento multi-condición, con lo que puede dar una idea aproximada de hasta que punto los algoritmos comparados responden bien ante degradaciones observadas previamente. El *set* B, por el contrario, está compuesto por los tipos de ruido aditivo que no se

encuentran representados en el corpus de entrenamiento multi-condición, no así la distorsión convolucional, que es la misma; de esta manera se pueden extraer conclusiones acerca de la eficiencia de las distintas técnicas comparadas ante ruidos aditivos no observados en la fase de entrenamiento. Por último, el *set* C consta de un escenario presente en el corpus de entrenamiento multi-codición y otro que no lo está, aunque en ambos casos se ha aplicado una distorsión convolucional distinta de la considerada en el corpus de entrenamiento multi-condición, de modo que con este *set* se puede analizar el comportamiento de las distintas técnicas que se pretendan comparar ante una distorsión convolucional no vista previamente y a la que, en uno de los casos, se le ha añadido un ruido aditivo tampoco considerado con anterioridad.

Este Capítulo se articula en tres Secciones, tantas como técnicas seleccionadas de cara a completar la experimentación con la base de datos *Aurora2*. Para tal fin se se eligieron los métodos quizás más representativos de entre los presentados en este trabajo, a saber, MEMLIN, MEMLIN CPGMM y la técnica híbrida basada en el algoritmo MEMLIN CPGMM haciendo uso de matrices de rotación dependientes de GMMs.

9.1. Resultados con la técnica MEMLIN sobre base de datos *Aurora2*.

Para llevar a cabo la experimentación sobre la base de datos *Aurora2*, no sólo ya para la técnica MEMLIN, sino para todas aquellas que precisan de una fase de entrenamiento previa con señal estéreo, se suele recurrir para tal efecto al corpus de entrenamiento multi-condición, de modo que se consideran 24 entornos básicos, que se corresponden con los cuatro tipos de ruido del *set* A: *subway*, *babble*, *car* y *exhibition hall* y 6 SNRs: *clean*, 20dB, 15dB, 10dB, 5dB y 0dB. Del mismo modo que para los experimentos realizados con el corpus *SpeechDat Car* en español, una vez normalizados los 13 coeficientes estáticos de los correspondientes vectores de características se calcularán las correspondientes derivadas y se aplicará el algoritmo CMS. En esta ocasión se utiliza la *parametrización estándar ETSI*, el modelado de lenguaje está compuesto por cualquier secuencia de dígitos y los modelos acústicos se construyen para cada palabras de vocabulario a partir de la estructura que se puede consultar en la Sección 4.3. Así pues, los resultados de referencia se pueden consultar en la Tabla 4.3.

En la Tabla 9.1 se presentan, del modo típico en que se suele hacer, los correspondientes resultados, tanto en términos de exactitud por palabra como majora relativa, obtenidos tras aplicar la técnica de compensación MEMLIN cuando, aplicando las condiciones de experientación anteriormente comentadas, se modelan tanto los 24 entornos básicos como el espacio limpio con 128 Gaussianas (se realizó un barrido para distintos números de componentes: 8, 16, 32, 64 y 128, cuyos resultados completos se pueden observar en el Apéndice 9.4 de este mismo Capítulo). Cabe destacar que, de aquí en adelante, para las distintas técnicas que se van a comparar en este Capítulo, y mientras no se indique lo contrario, el número de Gaussianas empleadas para modelar el espacio limpio será el mismo que el utilizado para representar cada entorno básico.

A la luz pues de los valores presentados en las Tablas 9.1 y 4.3 se puede concluir que, teniendo en cuenta únicamente los mejores resultados para la técnica MEMLIN, el método se comporta en general de un modo bastante satisfactorio, obteniendo una mejora media final del 62.23 %. Ahora bien, si se hace un estudio algo más específico de cada *set* de reconocimiento, se puede observar que, el *set* A alcanza, en media, las mejores tasas de acierto en palabra con respecto al resto de *set*. Esto es algo que no debe sorprender puesto que los tipos de ruido tratados habían sido previamente

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restaurar	Street	Airport	Station	Average	Subway N	Street M	Average	Average
Absolute word accuracy. If an HTK output is WORD: %Corr=99.14, Acc=98.68 [H=……], the value to enter is 98.68.	Clean	99,20	98,91	99,25	99,41	99,20	99,20	98,91	99,25	99,41	99,20	98,99	99,00	98,99	99,16
	20 dB	98,19	98,07	98,45	98,21	98,23	98,34	97,83	98,30	98,61	98,27	97,42	97,46	97,44	98,09
	15 dB	96,87	97,04	97,44	96,37	96,93	97,00	96,44	96,25	97,02	96,68	94,02	94,27	94,15	96,27
	10 dB	93,10	92,84	92,74	92,56	92,81	92,13	90,21	92,37	92,60	91,83	84,83	86,19	85,51	90,96
	5 dB	83,99	78,09	79,91	81,90	80,97	78,82	74,59	79,23	79,68	78,08	60,41	69,42	64,92	76,60
	0 dB	61,62	50,11	52,03	60,85	56,15	53,05	50,58	57,68	52,88	53,55	30,67	43,70	37,18	51,32
	-5dB	33,11	26,55	26,00	32,54	29,55	28,08	26,63	29,56	27,92	28,05	15,73	23,75	19,74	26,99
	Average	86,76	83,23	84,11	85,98	**85,02**	83,87	81,93	84,77	84,16	**83,68**	73,47	78,21	**75,84**	82,65

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restaurar	Street	Airport	Station	Average	Subway N	Street M	Average	Average
Detailed relative results in terms of error reduction. Halving the error rate = +50%	Clean	28,08%	-12,20%	21,54%	20,80%	14,55%	28,08%	-12,20%	21,54%	20,80%	14,55%	-22,07%	-9,63%	-15,85%	8,47%
	20 dB	44,26%	79,55%	46,90%	52,93%	55,91%	83,20%	47,52%	83,12%	73,39%	71,81%	61,11%	47,87%	54,49%	61,99%
	15 dB	63,09%	89,37%	77,61%	63,58%	73,41%	88,24%	69,98%	85,68%	84,12%	82,01%	57,20%	47,51%	52,35%	72,64%
	10 dB	71,82%	86,34%	80,09%	73,14%	77,85%	83,64%	71,17%	84,96%	83,49%	80,82%	45,94%	44,68%	45,31%	72,53%
	5 dB	69,60%	71,58%	70,98%	70,28%	70,61%	71,07%	60,12%	72,45%	72,92%	69,14%	19,82%	38,14%	28,98%	61,70%
	0 dB	50,51%	47,19%	46,28%	54,35%	49,58%	49,45%	40,20%	52,72%	47,93%	47,58%	8,13%	26,27%	17,20%	42,30%
	-5dB	25,14%	26,46%	20,57%	27,58%	24,94%	27,39%	19,71%	25,64%	23,21%	23,99%	3,24%	14,15%	8,69%	21,31%
	Average	59,86%	74,80%	64,37%	62,86%	65,47%	75,12%	57,80%	75,79%	72,37%	70,27%	38,44%	40,90%	39,67%	62,23%

Figura 9.1: Exatitud por palabra, *word accuracy*, y mejoras relativas obtenidas para los distintos *sets* (A, B y C) de la base de datos *Aurora2* utilizando la técnica de normalización de vectores de características MEMLIN, modelando cada uno de los entornos básicos con 128 Gaussianas. Se ha empleado la *parametrización estándar ETSI* y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*.

considerados en el corpus de entrenamiento multi-condición. Por su parte, y no lejos de los resultados alcanzados con el *set* A, se encuentran los logrados con el *set* B, lo que permite concluir que, el algoritmo MEMLIN es capaz de proporcionar un interesante comportamiento aun ante señales de entornos ruidosos no observados con anterioridad, siempre y cuando, eso sí, los que formen parte del corpus de entrenamiento sean próximos. Esto es debido al gran número de Gaussianas con que se modela el espacio ruidoso global y que hace que, a la postre, cada vector de características que se pretenda normalizar, siempre y cuando no esté fuera del ámbito de las mismas, pueda ser representada por alguna de las correspondientes a otro entorno básico. Sin embargo, no se puede decir lo mismo de los resultados presentados para el *set* C, que son sensiblemente menos satisfactorios (mejora media del 39.67 %), lo que viene a indicar que la técnica MEMLIN es mucho más sensible ante distorsiones convolucionales que ante ruidos aditivos no observados en el proceso de entrenamiento.

A la hora de observar como se comporta el algoritmo MEMLIN cuando se incrementa el número de Gaussianas con que se modelan los distintos entornos básicos, en la Figura 9.2 se muestra la mejora media cuando se realiza el siguiente barrido para el correspondiente número de componentes: 4, 8, 16, 32, 64 y 128. Se puede apreciar como, a pesar de la mejora de las prestaciones conforme se incrementa el número de componentes, ésta no llega en ningún momento a los niveles que se habían alcanzado con la base de datos *SpeechDat Car* en español, algo lógico por otra parte si se tiene en cuenta que la primera es algo más compleja.

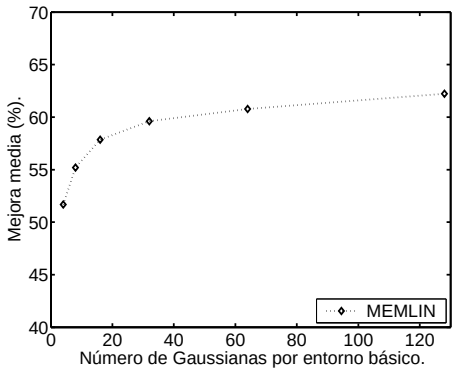


Figura 9.2: Mejoras medias de la exatitud por palabra, *word accuracy*, obtenidas para los distintos *sets* (A, B y C) de la base de datos *Aurora2* utilizando la técnica de normalización de vectores de características MEMLIN empleando distinto número de componentes para modelar los entornos básicos. Se ha empleado la *parametrización estándar ETSI* y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*.

9.2. Resultados con la técnica MEMLIN CPGMM sobre base de datos *Aurora2*.

9.3. Resultados con la técnica híbrida MEMLIN CPGMM a partir del cálculo de matrices de rotación dependientes de GMMs sobre base de datos *Aurora2*.

9.4. Anexo L.

10

Conclusiones y Líneas Futuras de Trabajo.

Durante los algo más de cuatro años que se han necesitado para completar este trabajo, a la vez que se iban completando los distintos objetivos marcados desde un principio, finalizando así las diversas tareas en que se dividió la tesis, se ponían las bases, tanto conceptuales como teóricas, para los siguientes pasos. De esta manera, utilizando como apoyo las conclusiones, estudios y resultados de todo el trabajo anterior se fueron desarrollando las distintas técnicas propuestas en esta tesis, así como las líneas de actuación futuras sobre las que seguir desarrollando soluciones para hacer más robusto el RAH ante cualquier tipo de efecto producido por elentorno acústico.

Así pues, el presente Capítulo se halla dividido en dos grandes unidades. De esta manera, en la Sección 10.1 se presentan las distintas conclusiones que, a lo largo del desarrollo del trabajo, se fueron observando y que, como ya se ha indicado, sirvieron posteriormente como punto de partida para estudios posteriores. Por su parte, en la Sección 10.2 se hace hincapié en aquellas debilidades de las técnicas propuestas que, aun observadas y constatadas durante el desarrollo de las mismas, no se trataron por cualquier causa, dejándo abiertas las puertas para futuras investigaciones.

10.1. Conclusiones.

SER MÁS EXPLÍCITO CON LO QUE SE HIZO EN CADA TÉCNICA
MIRAR LAS CONCLUSIONES INCLUIDAS EN CADA CAPÍTULO.
RELLENAR CADA PÁRRAFO.

El que las prestaciones de los sistemas de RAH decaen ante entornos acústicos adversos es algo que, no por menos sabido, debe dejarse de constatar; no tanto por la afirmación en sí sino por llegar a entender hasta que punto y de que manera afecta el entorno acústico a los vectores de características con los que posteriormente se reconocerá. En este sentido en el Capítulo 4 y sucesivos se presentan los resultados de RAH, así como distintos histogramas y *log-scattegrams* obtenidos a partir de las señales ruidosas de las bases de datos *SpeechDat Car* en español y *Aurora2*. De todo lo anterior se ha podido comprobar que el ruido propio del entorno acústico produce serias alteraciones tanto en la media como en la varianza de los componentes de los vectores acústicos, introduciendo además una gran incertidumbre, consecuencia de la aleatoriedad del ruido y que supone el mayor reto para las técnicas de normalización de vectores de características.

Una vez analizado el problema que se pretende compensar, en el Capítulo 5 se estudiaron las técnicas de normalización empíricas más empleadas en la actualidad: CMS, SPLICE y RATZ, presentando un desarrollo teórico conjunto con el que pudo comprobar que la única diferencia entre ellas reside en la aplicación de ciertas aproximaciones. De la misma manera se concluyó, de modo experimental, que las transformaciones propuestas en los tres casos no son todo lo específicas que deberían si se pretende compensar el efecto de la aleatoriedad del ruido. Teniendo en cuenta este hecho se desarrollo el algoritmo MEMLIN, que proporcionó un mejor comportamiento que los métodos anteriores debido a la utilización de vectores de transformación asociados a pares de Gaussianas, cosa que permite reducir el rango de proyección de los vectores acústicos degradados a nivel de componente y no de espacio, como sucedía en las técnicas consideradas hasta la fecha. Así, se obtuvo una mejora media para la base de datos *SpeechDat Car* en español de 70.22 %, ligeramente superior que las alcanzadas con las técnicas IRATZ (61.84 %) y SPLICE ME (65.39 %).

Tras un análisis en profundidad de la técnica MEMLIN, se detectaron dos grandes líneas de actuación que acabaron por proporcionar importantes mejoras a nivel de tasas de RAH, a saber: el modelado del vector de características limpio, que en la técnica MEMLIN se suponía lineal con término dependiente unitario, y el modelado de la probabilidad entre Gaussianas, que, en un principio, se consideraba independiente del vector acústico degradado.

El considerar que el modelo de degradación del vector acústico limpio es lineal con término dependiente unitario presupone que el entorno acústico únicamente afecta a las medias de los vectores de características, lo que, en general, es aproximadamente cierto para distorsión convolucional, pero no para ruido aditivo. Es por ello por lo que se propusieron una serie de soluciones en las que las transformaciones asociadas a cada par de Gaussianas fueran, bien lineales con término dependiente no unitario, lo que da lugar a la técnica P-MEMLIN, bien no lineales, generando el algoritmo MEMHIN. Los mejores resultados de RAH para ambas aproximaciones fueron, para la base de datos *SpeechDat Car*, muy similares a los alcanzados con el método MEMLIN: 70.47 % y 70.22 %, respectivamente. Sin emabargo sí que aportan interesantes mejoras cuando se modelan con un número reducido de componentes los entornos básicos y el espacio limpio. En esas condiciones, la modificación de las varianzas toma un importante papel que, por el incremento del número de Gaussianas se ve sensiblemente reducido. Por otra parte, esta mejora de comportamiento también se hace palpable cuando las señales ruidosas se hallan fuertemente afectadas por ruido aditivo. Así pues, aunque para la base de datos con que se realizó la mayor parte de la experimentación no aportan importantes majoras, no hay que desechar estas dos técnicas, puesto que se ha demostrado que, bajo otras condiciones de experimentación, sí pueden ser muy útiles. Ya para concluir las modificaciones sobre el modelado del vector de características limpio se propuso aprender transformaciones asociadas a cada par de Gaussianas de un mismo fonema, de manera que se definió una versión dependiente de fonemas para la técnica MEMLIN, que se denominó PD-MEMLIN. Se pudo comprobar que debido a utilizar transformaciones aún más específicas se lograba una importante mejora en el comportamiento, alcanzando un 75.44 % cuando la experimentación se realiza sobre el corpus *SpeechDat Car*. Además, con esta nueva solución se reduce el desajuste entre los vectores de características normalizados y los modelos acústicos.

Llegados a este punto, se trató uno de los problemas más conflictivos de las técnicas de normalización de vectores de características empíricas. Normalmente en todas ellas es necesario una fase de entrenamiento con señal estéreo y esto, en algunas ocasiones, no es posible. Por ello se desarrollo una fase de entrenamiento "ciega" para la técnica PD-MEMLIN en la que no sólo es necesario la señal degradada (de forma directa se obtiene la solución para el método MEMLIN). Así pues, y haciendo uso de la nueva fase de entrenamiento, se alcanzó una mejora con la base

de datos *SpeechDat Car* en español de 72.40 %, ligeramente inferior de la que se obtiene con el método PD-MEMLIN y algo superior a la alcanzada con el algoritmo MEMLIN, cerrándose de esta manera una de las grandes polémicas que las técnicas de normalización de vectores acústicos empíricas posee.

SEGUNDO PUNTO CONFLICTIVO: RUIDO NO VISTO (MEMLIN CON AURORA2).

Una vez tratado el problema del modelado del vector acústico limpio, y tras proporcionar, como se ha visto, tres nuevas soluciones cuyas mejoras quedaron patentes, se estudió la segunda gran línea de actuación sobre la técnica MEMLIN: el modelado de la probabilidad entre Gaussianas. El método MEMLIN se definió de manera que dicho modelado fuera independiente del vector acústico ruidoso, lo que, en realidad no deja de ser una aproximación. Por todo ello se decidió modelar los vectores de características ruidosos asociados a cada par de Gaussianas mediante una GMM. De esta manera se elimina la independencia temporal, creando una solución mucho más dinámica que proporciona un mejor comportamiento tanto si se aplica al método MEMLIN, como al PD-MEMLIN: 78.48 % y 77.72 % de mejora, respectivamente. Sin embargo estas dos soluciones, si bien alcanzan unas tasas de RAH satisfactorias, requieren de un mayor coste computacional. Para reducirlo se decidió no evaluar el nuevo modelado de probabilidad cruzada para todos los pares de Gaussianas sino sólo para aquellos más probables. Esto permite una elevada reducción en el número de componentes evaluadas a la vez que las tasas de RAH no se ven especialmente comprometidas. FALTA MEMLIN GMM CON AURORA2 RUIDO NO VISTO.

Dado que desde un primer momento se tuvo consciencia de las limitaciones que las técnicas de normalización de vectores de características poseen debido a la aleatoriedad del ruido, se propuso en el Capítulo 8, la combinación de las técnicas más características que se habían propuesto hasta el momento en el trabajo con métodos de adaptación de modelos acústicos. Así pues, para cada vector de características normalizado mediante los métodos MEMLIN y MEMLIN con modelado de la probabilidad cruzada basada en GMMs se le asoció una matriz de rotación en el proceso de decodificación de entre un conjunto de posibles opciones estimado previamente en una fase de entrenamiento. De esta manera, las correspondientes soluciones híbridas, que se pueden ver igualmente como un método de adpatación de modelos acústicos, alcanzan unas importantes mejoras: 90.54 % y 92.07 %, respectivamente cuando se emplea la base de datos *SpeechDat Car* en español con modelos acústicos de palabras (la técnica MEMLIN bajo esas mismas condiciones obtiene un 75.79 %). La gran ventaja de esta técnica híbrida reside en que no es necesario conocer la trascripción de la señal de entrenamiento, del mismo modo que tampoco es imprescindible emplear señal de la tarea propia de reconocimiento, lo que le proporciona una importante mejora con respecto a técnicas básicas de adaptación de modelos acústicos como MAP, MLLR... Por otra parte, si se dispone de la trascripción de la señal de entrenamiento es posible entrenar directamente modelos del espacio normalizado si se hace uso del algoritmo ML sobre la señal degradada del corpus de entrenamiento previamente normalizada. Si se hace esto, se consigue una mejora media de 96.33 % si el espacio normalizado se obtiene a partir de la técnica MEMLIN y 97.27 % si este último se genera mediante la señal normalizada con el algoritmo MEMLIN con modelado de la probabilida cruzada basada en GMMs. MEMLIN DECODER CON AURORA2.

10.2. Líneas Futuras de Trabajo.

Bibliografía

[Acero and Huang, 1995] A. Acero and X. Huang. Augmented cepstral normalization for robust speech recognition. In *Proc. IEEE Workshop on Automatic Speech Recognition*, Snowbird, UT, Dec 1995.

[Acero and Stern, 1990] A. Acero and R. M. Stern. Environmental robustness in automatic speech recognition. In *Proc. of ICASSP.*, pages 849–852, 1990.

[Acero *et al.*, 2000] A. Acero, L. Deng, T. Kristjansson, and J. Zhang. Hmm adaptation using vector taylor series for noisy speech recognition. In *Proc. ICSLP*, Beiling, China, 2000.

[Acero, 1990] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* PhD thesis, ECE Department, Carnegie-Mellon University, Pittsburgh, USA, Sep 1990.

[Andreou *et al.*, 1994] A. Andreou, T. Kamm, and J. Cohen. A parametric approach to vocal tract length normalization. In *In Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[Atal, 1983] B.-S. Atal. Efficient coding of lpc parameters by temporal decomposition. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 81–84, 1983.

[Bahl *et al.*, 1983] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:179–190, March 1983.

[Baker, 1975] J. K. Baker. *Stochastic Modelling for Automatic Speech Understanding*, pages 512–542. Academic Press, New York, NJ, USA, 1975.

[Baum, 1972] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In Oved Shisha, editor, *Inequalities III: Proceedings of the Third Symposium on Inequalities*, pages 1–8, University of California, Los Angeles, 1972. Academic Press.

[Beattie, 1992] V. L. Beattie. *Hidden Markov Model State-Based Noise Compensation.* PhD thesis, Churchill College, Cambridge University, Cambridge, UK, 1992.

[Beh and Ko, 2003] J. Beh and H. Ko. Spectral subtraction using spectral harmonics for robust speech recognition in car environments. In *International Conference on Computational Science2003*, pages 1109–1116, 2003.

[Bellegarda, 1997] J. R. Bellegarda. Statistical techniques for robust asr: Review and perspectives. In *in Proc. Eurospeech*, pages 33–36, 1997.

[Bellman and Kabala, 1965] R. Bellman and R. Kabala. Dynamic programming and modern control theory. *Academic Press Inc.*, 1965.

[Bellman, 1957] R. E. Bellman. *Dynamic Programming.* Princeton University Press., Princeton, NJ, 1957.

[Bimbot *et al.*, 1988] F. Bimbot, G. Chollet, P. Deleglise, and C. Montacie. Temporal decomposition and acoustic-phonetic decoding of speech. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 445–448, New York, USA, 1988.

[Bippus *et al.*, 1999] R. Bippus, A. Fischer, and V. Stahl. Domain adaptation for robust automatic speech recognition in car environments. In *in Proc. Eurospeech*, pages 1943–1946, 1999.

[Bocchieri, 1993] E. Bocchieri. Vector quantization for the efficient computation of continuous density likelihoods. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume II, pages 692–695, Minneapolis, USA, April 1993.

[Boll, 1979] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on ASSP*, 27:113–120, April 1979.

[Bou-Ghazale and Hansen, 2000] S. Bou-Ghazale and J.H.L. Hansen. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 8(4):429–442, 2000.

[Bourlard *et al.*, 1972] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan. Towards subband-based speech recognition. In Oved Shisha, editor, *proc. of European Signal Processing Conference*, pages 1579–1582, Trieste, Italy, 1972. Academic Press.

[Bourlard *et al.*, 1996] H. Bourlard, H. Hermansky, and N. Morgan. Towards increasing speech recognition error rates. *Speech Communication*, 18(3):205–231, 1996.

[Brown *et al.*, 1992] P. Brown, V. Della Pietra, P. Souza, J. Lai, and R. Mercer. Class-based n-gram models of natural language. *Computation Linguistics*, 18(4):467–479, 1992.

[Buera *et al.*, 2004a] L. Buera, E. Lleida, A. Miguel, and A. Ortega. Multi-environment models based linear normalization for speech recognition in car conditions. In *Acoustics, Speech, and Signal Processing, ICASSP*, Motreal, Canada, May 2004.

[Buera *et al.*, 2004b] L. Buera, E. Lleida, A. Miguel, and A. Ortega. Multi-environment models based linear normalization for robust speech recognition. In *Proceedings of the International Conference 'Speech and Computer', SPECOM*, St. Petersburg, Russia, 2004.

[Buera *et al.*, 2004c] L. Buera, E. Lleida, A. Ortega, A. Miguel, and O. Saz. Avances en la normalización cepstral con señal estéreo para el reconocimiento robusto de voz en el entorno del vehículo. In *III Jornadas en Tecnología del Habla*, Valencia, Spain, Nov 2004.

[Buera *et al.*, 2005a] L. Buera, E. Lleida, A. Miguel, and A. Ortega. Multi-environment linear normalization for robust speech analysis in cars. In *Biennial on DSP for in-Vehicle and Mobile Systems*, Sesimbra, Portugal, Sept. 2005.

[Buera *et al.*, 2005b] L. Buera, E. Lleida, A. Miguel, and A. Ortega. Recent advances in pd-memlin for speech recognition in car conditions. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, San Juan, Puerto Rico, November 2005.

[Buera *et al.*, 2005c] L. Buera, E. Lleida, A. Miguel, and A. Ortega. Robust speech recognition in cars using phoneme dependent multi-environment linear normalization. In *Interspeech - Eurospeech,9th European Conference on Speech Communication and Technology*, Sept. 2005.

[Buera *et al.*, 2005d] L. Buera, E. Lleida, J. D. Rosas, J. Villalba, A. Miguel, A. Ortega, and O. Saz. Speaker verification and identification using phoneme dependent multi-environment models based linear normalization in adverse and dynamic environments. In *Summer school for advanced studies on biometrics for secure authentication and system integration*, Alghero, Italy, June 2005.

[Buera *et al.*, 2006a] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz. Time-dependent cross-probability model for feature vector normalization. In *IV Jornadas en Tecnología del Habla*, Nov. 2006.

[Buera *et al.*, 2006b] L. Buera, E. Lleida, J.A. Nolazco, A. Miguel, and A. Ortega. Time-dependent cross-probability model for multi-environment model based linear normalization. In *ICSLP*, Sept. 2006.

[Buera *et al.*, 2006c] L. Buera, E. Lleida, J. D. Rosas, J. Villalba, A. Miguel, A. Ortega, and O. Saz. Verificación e identificación de locutor con normalización de vectores de características en entornos acústicos adversos. In *Terceras Jornadas de Reconocimiento Biométrico de Personas*, Sevilla, Spain, Nov 2006.

[Buera *et al.*, 2007] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz. Cepstral vector normalization based on stereo data for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2007.

[Carlson and Clements, 1991] B. A. Carlson and M. A. Clements. Application of a weighted projection measurement for robust hidden markov model based speech recognition. In *Proc. ICASSP*, 1991.

[Cerf-Danon and El-Béze, 1991] H. Cerf-Danon and M. El-Béze. Three different probabilistic language models: Comparison and combination. In *Proc. ICASSP*, pages 297–300, 1991.

[Cerisana *et al.*, 2000] C. Cerisana, L. Rigazio, R. Boman, and J.-C. Junqua. Transformation of jacobian matrices for noisy speech recognition. In *Proc. ICSLP*, pages 179–182, 2000.

[Chesta *et al.*, 1999] C. Chesta, O. Siohan, and C-H Lee. Maximum a posteriori linear regression for hidden markov model adaptation. In *in Proc. Eurospeech*, volume 1, pages 211–214, Budapest, Hungary, 1999.

[Cole *et al.*, 1995] R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Biermann, M. Bush, M. Clements, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D. G. Novick, M. Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Splitz, A. Waibel, C. Weinstein, S. Zahorian, and V. Zue. The challenge of spoken language systems: Research directions for the nineties. *IEEE Transactions on Speech and Audio Processing*, 1(3):1–21, Jan. 1995.

[Cook *et al.*, 2001] M. Cook, Ph. Green, L. Josifovski, , and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34(34):267–285, 2001.

[Davis and Mermelstein, 1980] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[de la Torre *et al.*, 2001] A. de la Torre, D. Fohr, and J. P. Haton. On the comparison of front-ends for robust speech recognition in car environments. In *ISCA ITR-Workshop on Adaptation Methods for Speech Recognition*, pages 109–112, Aug 2001.

[de la Torre *et al.*, 2005] A. de la Torre, A.M. Peinado, J.C. Segura, J.L. Pérez-Córdoba, C. Benítez, and A.J. Rubio. Histogram equalization of the speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(3):355–366, may 2005.

[Deleglise, 1990] P. Deleglise. Décomposition temporelle : une technique cinématique de segmentation et de décodage acoustico-phonétique. In *évaluations. XVIIIème JEP*, pages 347–352, Montréal, Canada, 1990.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society*, 39(1):1–21, 1977.

[Deng *et al.*, 2000] L. Deng, A. Acero, M. Plumpe, and X. D. Huang. Large-vocabulary speech recognition under adverse acoustic environments. In *Proc. ICSLP*, pages 806–809, Beijing, China, 2000.

[Deng *et al.*, 2003] L. Deng, J. Droppo, and A. Acero. Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 11(6):568–580, Nov. 2003.

[Digalakis, 1992] V. V. Digalakis. *Segment-based stochastic models of spectral dynamics for continuous speech recognition.* PhD thesis, Boston University, Boston, MA, USA, 1992.

[Droppo *et al.*, 2001] J. Droppo, L. Deng, and A. Acero. Evaluation of the splice algorithm on the aurora2 database. In *in Proc. Eurospeech*, volume 1, Sept. 2001.

[Droppo *et al.*, 2002] J. Droppo, L. Deng, and A. Acero. Uncertainty decoding with splice for noise robust speech recognition. In *Proc. ICASSP*, Florida, USA, May 2002.

[Droppo *et al.*, 2005] J. Droppo, M. Mahajan, A. Gunawardana, and A. Acero. How to train a discriminative front end with stochastic gradient descent and maximum mutual information. In *Proc. ASRU*, Puerto Rico, Dec 2005.

[Duda and Hart, 1973] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* J. Wiley and sons, 1973.

[Duda *et al.*, 2000] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* Wiley-Interscience, New York, 2000.

[Ephraim and Malah, 1985] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. ASSP*, 33(2):443–445, Apr 1985.

[Erzin *et al.*, 1995] E. Erzin, A.E. Cetin, and Y. Yardimci. Subband analysis for speech recognition in the presence of car noise. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 417–420, Detroit, USA, 1995.

[ETSI, 2000] ETSI. Speech processing transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms. Technical report, ETSI ES 201 108 version 1.1.2, April 2000.

[Fred and Leitao, 1994] A. L.Ñ. Fred and J. M.Ñ. Leitao. Improving sentence recognition in stochastic context-free grammars. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 9–12, 1994.

[Fritsch, 1997] J. Fritsch. ACID/HNN: A framework for hierarchial connectionist acoustic modelling. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, pages 164–171, Santa Barbara, USA, Dec. 1997.

[Fukunaga, 1990] K. Fukunaga. *Statistical pattern Recognition.* Academic Press, 1990.

[Furui, 1986] S. Furui. Speaker independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Speech and Audio Processing*, 34:52–59, Feb 1986.

[Gales and Young, 1993] M. J. F. Gales and S. J. Young. Hmm recognition in noise using parallel model combination. In *Proc. of EUROSPEECH*, pages 837–840, 1993.

[Gales, 1995] M. J. F. Gales. *Model-Based Techniques for Noise Robust Speech Recognition.* PhD thesis, Cambridge University, Cambridge, UK, 1995.

[Gales, 1997a] M. J. F. Gales. Transformation smoothing for speaker and environmental adaptation. In *Proc. of EUROSPEECH*, pages 2067–2070, 1997.

[Gales, 1997b] M.J.F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. Cued/finfeng/tr291, Cambridge University, 1997.

[Gauvain and Lee, 1994] J.-L. Gauvain and C.-H. Lee. Maximum a posteriory estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 2:291–298, Apr 1994.

[Gelin and Junqua, 1999] P. Gelin and J-C. Junqua. Techniques for robust speech recognition in the car environment. In *Proc. of EUROSPEECH*, pages 2483–2486, 1999.

[Ghahramani and Beal, 2000] Z. Ghahramani and M.J. Beal. *Variational Inference for Bayesian Mixtures of Factor Analysers*, pages 449–455. MIT Press, 2000.

[Ghahramani and Jordan, 1997] Z. Ghahramani and M. I. Jordan. Factorial hidden markov models. *Machine Learning*, 29(2-3):245–273, 1997.

[Ghahramani, 2002] Z. Ghahramani. *An introduction to hidden Markov models and Bayesian networks*, pages 9–42. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2002.

[Ghitza, 1992] O. Ghitza. *Auditory nerve representation as a basis for speech processing*, pages 453–486. Dekker, 1992.

[Ghitza, 1994] O. Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 2:115–132, Jan 1994.

[Gillick and Cox, 1989] L. Gillick and S. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 532–535, 1989.

[Gish and Ng, 1996] H. Gish and K.Ñg. Parametric trajectory models for speech recognition. In *Proc. ICSLP*, pages 466–469, 1996.

[Glass, 2003] J. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17(2-3):137–152, 2003.

[Goldenthal, 1994] W. D. Goldenthal. *Statistical Trajectory Models for Phonetic Recognition.* PhD thesis, Massachusetts Institute of Technology, MA, USA, 1994.

[Gong, 1995] Y. Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 3(16):261–291, 1995.

[González and Wintz, 1987] R. C. González and P. Wintz. *Digital image processing.* Addison Wesley, 1987.

[Gorin *et al.*, 1997] A. L. Gorin, G. Riccardi, and J. H. Wright. How may i help you? *Speech Communication*, 23(1-2):113–127, 1997.

[Gravier *et al.*, 1999] G. Gravier, M. Sigelle, and G. Chollet. Markov random field modeling for speech recognition. *Australian Journal of Intelligent Information Processing Systems*, 5(4):245–252, 1999.

[Gravier, 2000] G. Gravier. *Analyse statistique á deux dimensions pour la modélisation segmentale du signal de parole - application á la reconnaissance.* PhD thesis, Ecole Nationale Supérieure des Télécommunications (ENST), Paris, January 2000.

[Greenberg and Kingsbury, 2000] S. Greenberg and B. E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 1647–1650, Munich, Germany, 2000.

[Hagen and Bourlard, 2000] A. Hagen and H. Bourlard. Using multiple time scales in the framework of multi-stream speech recognition. In *ICSLP*, 2000.

[Hagen, 2000] A. Hagen. *Robust speech recognition based on multi-stream processing.* PhD thesis, Département dínformatique,EPFL, Lausanne, Switzerland, January 2000.

[Hanai and Stern, 1994] N. Hanai and R. M. Stern. Robust speech recognition in the automobile. In *in Proc. ICSLP*, pages 1339–1342, 1994.

[Hermansky and Morgan, 1994] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[Hermansky *et al.*, 1991] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Compensation for the effect of communication channel in auditory-like analysis of speech (rasta-plp). In *Proc. of EUROSPEECH*, pages 1367–1370, 1991.

[Hermansky *et al.*, 1993] H. Hermansky, N. Morgan, and H. G. Hirsch. Recognition of speech in additive and convolutional noise based on rasta spectral processing. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages 83–96, 1993.

[Hermansky *et al.*, 1996] H. Hermansky, S. Tibrewala, and M. Pavel. Towards asr on partially corrupted speech. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 544–547, October 1996.

[Hermansky, 1990] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of Acoustic Society of America*, 87(4):1738–1792, 1990.

[Hermansky, 1998] H. Hermansky. Speech beyond 10 milliseconds (temporal filtering in feature domain). Technical report, Center for Spoken Language Understanding, Department of Electrical Engineering and Applied Physics, Oregon Graduate Institute of Science and Technology, OR, USA, 1998.

[Hernández *et al.*, 2007] I. Hernández, P. García, J.Ñolazco, L. Buera, and E. Lleida. Robust automatic speech recognition using pd-meemlin. In *3rd Iberian Conference on Pattern Recognition and Image Analysis*, Gerona, Spain, June. 2007.

[Herrando and Nadeu, 1994] J. Herrando and C.Ñadeu. Speech recognition in noisy car environment based on osalpc representation and robust similarity measuring techniques. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 69–72, 1994.

[Hirsch and Pearce, 2000] H. G. Hirsch and D. Pearce. The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *Proc. in ISCA ITRW ASR2000*, Paris, France, September 2000.

[Holmes and Huckvale, 1994] W.J. Holmes and M. Huckvale. Why have hmms been so successful for automatic speech recognition and how might they be improved? *Speech, Hearing and Language, UCL Work in Progress*, 8:207–219, 1994.

[Holmes *et al.*, 1997] J.Ñ. Holmes, W. J. Holmes, and P.Ñ. Garner. Using formant frequencies in speech recognition. In *Proc. of EUROSPEECH*, pages 2083–2086, 1997.

[Hoshino, 2001] H. Hoshino. Noise-robust speech recognition in a car environment based on the acoustic features of car interior noise. Technical report, 2001.

[Huang *et al.*, 2001] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[Huo and Lee, 1997] Q. Huo and C. H. Lee. On-line adaptive learning of the continuous density hidden markov model based on approximate recursive bayes estimate. *IEEE Transactions on Speech and Audio Processing*, 5(2):161–172, March. 1997.

[ITU, 1996] ITU. Transmission performance characteristics of pulse code modulation channels. Technical report, Nov. 1996.

[Jacobs *et al.*, 2002] R. A. Jacobs, W. Jiang, and M. A. Tanner. Factorial hidden markov models and the generalized backfitting algorithm. *Neural Computation*, 14(10):2415–2437, 2002.

[Jankowski *et al.*, 1995] C. R. Jankowski, Jr. Hoang-Doan, and R. P. Lippmann. A comparison of signal processing front ends for automatic word recognition. *IEEE Transactions on Speech and Audio Processing*, 3:286–293, Jul. 1995.

[Jelinek *et al.*, 1975] F. Jelinek, L. R. Bahl, and R. L. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, IT-21(3):250–256, May. 1975.

[Jelinek, 1969] F. Jelinek. A fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development*, 13:675–685, Nov. 1969.

[Jelinek, 1976] F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, 1976.

[Jelinek, 1991] F. Jelinek. *Self-Organized Language Modelling for Speech Recognition*, chapter 6.1, pages 450–506. Morgan Kaufmann, San Mateo, CA, 1991.

[Jordan and Jacobs, 1994] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Comput.*, 6(2):181–214, 1994.

[Josifovski, 2002] L. Josifovski. *Robust Automatic Speech Recognition Missing and Unreliable Data.* PhD thesis, Department of Computer Science, University of Sheffield, UK, 2002.

[Juang *et al.*, 1987] B. H. Juang, L. R. Rabiner, and J. G. Wilpon. On the use of bandpass liftering in speech recognition. *IEEE Transactions on ASSP*, 7(35):947–954, July 1987.

[Junqua and Haton, 1996] J.-C. Junqua and J. P. Haton. *Robustness in Automatic Speech Recognition*. Kluwer Academic Publishers, 1996.

[Junqua and Wakita, 1989] J.-C. Junqua and H. Wakita. A comparative study of cepstral lifters and distance measures for all pole models of speech in noise. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 476–479, 1989.

[Kaiser, 1990] J.F. Kaiser. On a simple algorithm to calculate the énergyóf a signal. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 381–384, Albuquerque, USA, 1990.

[Kamm *et al.*, 1997] C. Kamm, M. Walker, and L. Rabiner. The role of speech processing in human-computer intelligent communication. *Speech Communication*, 4(23):263–278, 1997.

[Kanthak *et al.*, 2000] S. Kanthak, K. Schütz, and H.Ñey. Using SIMID instructions for fast likelihood calculation in LVCSR. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume III, pages 1531–1534, Istanbul, Turkey, June 2000.

[Katz, 1987] S. M. Katz. Estimation of probabilities from sparce data for the language model component of a speech recognizer. *IEEE Transactions on Speech and Audio Processing*, 35(3):400–401, March. 1987.

[Kim *et al.*, 1998] D. Y. Kim, C. K. Un, and N. S. Kim. Speech recognition in noisy environments using first-order vector taylor series. *IEEE Transactions on Signal Processing*, 5(3):57–59, March 1998.

[Kingbury *et al.*, 1998] B. E. D. Kingbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1-3):117–132, 1998.

[Kleinschmidt, 2002] M. Kleinschmidt. *Robust Speech Recognition Based on Spectro-temporal Processing*. PhD thesis, University of Oldenburg, Germany, 2002.

[Korkmazskiy *et al.*, 2000] F. Korkmazskiy, F. K. Soong, and O. Siohan. Constrained spectrum normalization for robust speech recognition in noise. In *Proc. of ASR*, pages 58–63, 2000.

[Kuhn and de Mori, 1990] R. Kuhn and R. de Mori. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583, June 1990.

[Kuhn *et al.*, 1994] T. Kuhn, H.Ñiermann, and E. G. Schukat-Talamazzini. Ergodic hidden markov models and polygrams for language modeling. In *Acoustics, Speech, and Signal Processing, ICASSP*, 1994.

[Kuhn *et al.*, 2000] R. Kuhn, E. Perronnin, P.Ñguyen, and N.Ñiedzielski. Rapid speaker adaptation in eigenvoice space robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, 2000.

[Kullback and Leibler, 1951] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–87, 1951.

[Kybic and Unser, 2003] J. Kybic and M. Unser. Fast parametric elastic image registration. *IEEE Transactions on Image Processing*, 11(12):1427–1442, 2003.

[Lafferty *et al.*, 2001] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, 2001.

[Lea, 1979] W. A. Lea. *Trends in Speech Recognition*. Ed. Lawrence Erlbaum, 1979.

[Lee and Rose, 1998] L. Lee and R. Rose. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing*, 1(6):49–60, 1998.

[Lee *et al.*, 1989] K. F. Lee, H. W. Hon, S. Hwang, S. Mahajan, and R. Reddy. The sphinx speech recognition system. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 445–448, Glasgow, Scotland, UK, 1989.

[Legetter and Woodland, 1995] C. J. Legetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continous-density hidden markov models. *Computer Speech and Language*, 9:171–185, 1995.

[Lehmann, 1975] E. L. Lehmann. *Nonparametrics*. Holden Day, 1975.

[Leonard and Doddington, 1993] R. G. Leonard and G. Doddington. Tidigits speech corpus. Technical report, Texas Instruments, Inc,, 1993.

[Lin and Chen, 1998] J. S. Lin and R. Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.

[Liu *et al.*, 1994] F. H. Liu, R. M. Stern, and A. Acero. Environment normalization for robust speech recognition using direct cepstral comparison. In *Proc. ICASSP*, 1994.

[Lleida *et al.*, 2002] E. Lleida, E. J. Magrau, A. Ortega, A. Miguel, and L. Buera. Reconocimiento automático del habla en vehículos, resultados con speech-dat car. In *Jornadas en Tecnología del Habla, JTH*, Granada, Spain, 2002.

[Lleida, 1990] E. Lleida. *Compresión y Selección de Información en Reconocimiento Automático del Habla*. PhD thesis, Universidad Politécnica de Cataluña, UPC, 1990.

[Lockwood and Boudy, 1992] P. Lockwood and J. Boudy. Experiments with a non-linear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars. *Speech Communication*, 11(2-3):215–228, 1992.

[Lombard, 1911] E. Lombard. Le signe de lélevation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37:101–119, 1911.

[Martin *et al.*, 1999] S. Martin, C. Hamacher, Liermann J, F. Wesssel, and H.Ñey. Assessment of smoothing methods and complex stochastic language modelling. In *Proc. of European Conf. on Speech Communication and Technology*, volume V, pages 1939–1942, Budapest, Hungary, Sept. 1999.

[Matassoni *et al.*, 2001] M. Matassoni, M. Omologo, and P. Svaizer. Use of real and contaminated speech for training of a hands-free in-car speech recognizer. In *Proc. of EUROSPEECH*, pages 3009–3012, Aalborg, Denmark, 2001.

[Matassoni *et al.*, 2002] M. Matassoni, M. Omologo, A. Santarelli, and P. Svaizer. On the joint use of noise reduction and mllr adaptation for in-car hands-free speech recognition. In *Acoustics, Speech, and Signal Processing, ICASSP*, Orlando, USA, 2002.

[McNemar, 1947] I. McNemar. Note on the sampling error of the diference between correlated proportions or percentages. In *Psychometrika*, volume 12, pages 153–157, 1947.

[Meyer and Simmer, 1997] J. Meyer and K. U. Simmer. Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 1167–1170, Munich, Germany, April 1997.

[Miguel *et al.*, 2005] A. Miguel, E. Lleida, R. Rose, L. Buera, and A. Ortega. Augmented state space acoustic decoding for modeling local variability in speech. In *Eurospeech,9th European Conference on Speech Communication and Technology, Interspeech*, pages 3009–3012, Lisbon, Portugal, 2005.

[Miguel *et al.*, 2006] A. Miguel, E. Lleida, A. Juan, L. Buera, A. Ortega, and O. Saz. Local transformation models for speech recognition. In *ICSLP*, Pittsburgh, USA, 2006.

[Mokbel and Cholet, 1995] C. E. Mokbel and G. F. A. Cholet. Automatic word recognition in cars. *IEEE Transactions on Speech and Audio Processing*, 3(5):346–356, Sep 1995.

[Mokbel *et al.*, 1993] C. E. Mokbel, J. Monn, and D. Jouvet. On-line adaptation of a speech recognizer to variantions in telephone line conditions. In *Proc. of EUROSPEECH*, volume 2, pages 1247–1250, 1993.

[Molau, 2003] S. Molau. *Normalization in the Acoustic Feature Space for Improved Speech Recognition*. PhD thesis, University of Aachen, Germany, Feb 2003.

[Moore, 1990] R. Moore. *Speech Processing*. McGraw Hill, 1990.

[Moreno *et al.*, 2000] Asuncion Moreno, Borge Lindberg, Christoph Draxler, Gael Richard, Khalid Choukri, Stephan Euler, and Jeffrey Allen. Speechdat-car. a large speech database for automotive environments. In *Proceedings of LREC*, volume 2, pages 895–900. Athens, Greece, June 2000.

[Moreno, 1996] P. Moreno. *Speech recognition in noisy environments*. PhD thesis, ECE Department, Carnegie-Mellon University, Apr. 1996.

[Morgan *et al.*, 1997] N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proc. of EUROSPEECH*, pages 2079–2082, 1997.

[Morgenthaler and Hansen, 1982] M. Morgenthaler and C. Hansen. Use of attributed grammars in speech signal processing. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 1311–1313, 1982.

[Mori, 1997] R. De Mori. Recent advances in feature extraction and acoustic modeling for automatic speech recognition. Technical report, Laboratoire Informatique dÁvignon, 1997.

[Morris *et al.*, 1999] A. Morris, A. Hagen, and H. Bourlard. The full combination sub-bands approach to noise robust HMM/ANN based ASR. In *Proc. of EUROSPEECH*, pages 599–602, 1999.

[Morris *et al.*, 2001] A. Morris, J. Barker, and H. Bourlard. From missing data to maybe useful data: soft data modelling for noise robust asr. IDIAP-RR 06, IDIAP, 2001.

[Nene and Nayar, 1996] S. A. Nene and S. K. Nayar. Closest point search in high dimensions. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 859–865, San Francisco, USA, June 1996.

[Neumeyer and Weintraub, 1994] L.Ñeumeyer and M. Weintraub. Probabilistic optimal filtering for robust speech recognition. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 417–420, 1994.

[Neumeyer and Weintraub, 1995] L.Ñeumeyer and M. Weintraub. Robust speech recognition in noise using adaptation and mapping techniques. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 141–144, Detroit, USA, May 1995.

[Ney and Essen, 1991] H.Ñey and U. Essen. On smoothing techniques for bigram-based natural language modelling. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 825–828, Toronto, Canada, May 1991.

[Ney *et al.*, 1987] H.Ñey, D. Mergel, A.Ñoll, and A. Paeseler. A data-driven organization of the dynamic programming beam search for continuous speech recognition. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 833–836, Dallas, USA, April 1987.

[Ney *et al.*, 1992] H.Ñey, R. Haeb-Umbach, B. H. Tran, and M. Oerder. Improvements in beam search for 10000-word continuous speech recognition. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume I, pages 9–12, San Francisco, USA, March 1992.

[Ney *et al.*, 1994] H.Ñey, U. Essen, and R. Kneser. On structuring probabilistic dependencies in language modelling. *Computer Speech and Language*, 2(8):1–38, 1994.

[Ney, 1990] H.Ñey. Stochastic grammars and pattern recognition. In *Proc. of NATO ASI*, pages 319–344, 1990.

[Ney, 1993] H.Ñey. Architecture and search strategies for large-vocabulary continuous-speech recognition. In *Proc. of NATO ASI*, pages 59–84, 1993.

[Nolazco and Young, 1994] J. A. Nolazco and S. J. Young. Continuous speech recognition in noise using spectral subtraction and hmm adaptation. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 409–412, 1994.

[Oppenheim and Schafer, 1975] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing.* Prentice-Hall, Inc., 1975.

[Ortmanns and Ney, 1995] S. Ortmanns and H.Ñey. An experimental study of the search space for 20000-word speech recognition. In *Proc. of the EUROSPEECH*, volume II, pages 901–904, Madrid, Spain, Sept. 1995.

[Ostendorf *et al.*, 1996] M. Ostendorf, V. Digilakis, and O. Kimball. From hmm's to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):360–378, 1996.

[Pallett *et al.*, 1990] D. Pallett, W. Fisher, and J. Fiscus. Tools for the analysis of benchmark speech recognition tests. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 97–100, 1990.

[Patil and Taillie, 2000] G. P. Patil and C. Taillie. A multiscale hierarchical markov transition matrix model for generating and analyzing thematic raster maps. Technical report, The Pennsylvania State University, Department of Statistics, 2000.

[Pereira and Warren, 1980] F. Pereira and D. Warren. Definite clause grammar for language analysis -survey of the formalism with augmented transition networks. *Artificial Intelligence*, 13:231–278, 1980.

[Pieraccini *et al.*, 1992] R. Pieraccini, E. Tzoukermann, Z. Gorelov, J-L. Gauvain, E. Levin, C-H Lee, and J. G. Wilpon. A speech understanding system based on statistical representation of semantics. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 193–196, 1992.

[Pitz and Ney, 2005] M. Pitz and H.Ñey. Vocal tract normalization equals linear transformation in cepstral space. *IEEE Transactions on Speech and Audio Processing*, 13(5):930–944, 2005.

[Pitz, 2005] M. Pitz. *Investigations on Linear Transformations for Speaker Adaptation and Normalization*. PhD thesis, University of Aachen, 2005.

[Potamianos *et al.*, 2003] G. Potamianos, C.Ñet, G. Gravier, A. Garg, and A.W.Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9):1306–1326, Sept. 2003.

[Pujol *et al.*, 2003] P. Pujol, S. Pol, C.Ñadeu, A. Hagen, and H. Bourlard. Comparison and combination of features in a hybrid hmm/mlp and a hmm/gmm speech recognition system. *IEEE Transactions on Speech and Audio Processing*, 12(1):14–22, 2003.

[Rabiner and Juang, 1993] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Englewood Cliffs NJ: PTR Prentice Hall (Signal Processing Series), 1993. General Intro : ISBN 0-13-015157-2.

[Rabiner, 1988] L. R. Rabiner. *A Tutorial on HMM and selected Applications in Speech Recognition*, chapter 6.1, pages 267–295. Morgan Kaufmann, 1988.

[Raj *et al.*, 1996] B. Raj, E. Gouvea, P. J. Moreno, and R. M. Stern. Cepstral compensation by polynomial approximation for environment-independent speech recognition. In *International Conference on Spoken Language Processing, ICSLP*, 1996.

[Rioul and Vetterli, 1991] O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, 8:11–38, 1991.

[Rose *et al.*, 2006] R. Rose, A. Keyvani, and A. Miguel. On the interaction between speaker normalization, environment compensation, and discriminant feature space transformations. In *ICASSP*, Toulouse, France, 2006.

[Rueckert *et al.*, 2001] D. Rueckert, A. F. Frangi, and J. A. Schnabel. Automatic construction of 3d statistical deformation models using non-rigid registration. In *MICCAI '01: Proceedings of the 4th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 77–84, London, UK, 2001. Springer-Verlag.

[Sankar and Lee, 1996] A. Sankar and C. Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4:190–202, May 1996.

[Sarikaya and Hansen, 2000] R. Sarikaya and J. H. L. Hansen. Improved jacobian adaptation for fast acoustic adaptation in noisy speech recognition. In *International Conference on Spoken Language Processing (ICSLP 2000)*, pages 702–705, Beijing, China, October 2000.

[Schwartz and Austin, 1991] R. Schwartz and S. Austin. A comparison of several approximate algorithms for finding multiple (n-best) sentence hypotheses. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 701–704, Toronto, Canada, May 1991.

[Schwartz and Chow, 1990] R. Schwartz and Y. L. Chow. The n-best algorithm: An efficient and exact procedure for finding the n most likely sentence hypotheses. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 1, pages 81–84, Albuquerque, USA, April 1990.

[Segarra and García, 1991] E. Segarra and P. García. Automatic learning of acoustic and syntactic-semantic levels in continuous speech understanding. In *Proc. EuroSpeech*, pages 861–864, Genova, Italy, Sept. 1991.

[Segura *et al.*, 2001] J. C. Segura, M. C. Benítez, A. de la Torre, and A. Rubio. Feature extraction from time-frequency matrices for robust speech recognition. In *Proc. EuroSpeech*, pages 1625–1628, Aalborg, Denmark, Sept. 2001.

[Shieber, 1985] S. M. Shieber. An introduction to unification-based approaches to grammar. *CSLI Lecture Notes: Center for the Study of Language and Information*, 1985.

[Shimodaira *et al.*, 2000] H. Shimodaira, Y. Kato, T. Akae, M.Ñakai, and S. Sagayama. Jacobian adaptation of hmm with initial model selection for noisy speech recognition. In *International Conference on Spoken Language Processing (ICSLP 2000)*, volume 2, pages 1003–1006, Beijing, China, October 2000.

[Shimodaira *et al.*, 2002] H. Shimodaira, N. Sakai, M.Ñakai, and S. Sagayama. Jacobian joint adaptation to noise, channel and vocal tract length. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 197–200, Orlando, USA, 2002.

[Steinbiss *et al.*, 1994] V. Steinbiss, B. H. Tran, and H.Ñey. Improvements in beam search. In *International Conference on Spoken Language Processing, ICSLP*, volume IV, pages 2143–2146, Yokohama, Japan, Sept. 1994.

[Stern and Larsy, 1987] R. M. Stern and M. J. Larsy. Dynamic speaker adaptation for feature-based isolated word recognition. *IEEE Transactions on Speech and Audio Processing*, 35(6):751–763, 1987.

[Stern *et al.*, 1997] R. M. Stern, B. Raj, and P.J. Moreno. Compensation for environmental degradation in automatic speech recognition. In *in Proc. of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 33–42, Pont-au-Mousson, France, April 1997.

[Sumby and Pollack, 1954] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *J. Acoustical Society America*, 26:212–215, 1954. W. H. Sumby and I. Pollack, "Visual contribution to speech intelligi- bility in noise," J. Acoustical Society America, vol. 26, pp. 212–215, 1954.

[Sun, 1995] D. X. Sun. Robust estimation of spectral center-of-gravity trajectories using mixture spline models. In *Proc. EUROSPEECH'95*, pages 749–752, Madrid, Spain, 1995.

[Uchida and Sakoe, 1998] S. Uchida and H. Sakoe. A monotonic and continuous two-dimensional warping based on dynamic programming. In *ICPR '98: Proceedings of the 14th International Conference on Pattern Recognition-Volume 1*, page 521, Washington, DC, USA, 1998. IEEE Computer Society.

[Usagawa *et al.*, 1994] T. Usagawa, M. Iwata, and M. Ebata. Speech parameter extraction in noisy environment using a masking model. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 81–84, 1994.

[van den Heuvel *et al.*, 1999] Henk van den Heuvel, Jerôme Boudy, Robrecht Comeyne, Stephan Euler, Asuncion Moreno, and G. Richard. The speechdat-car multilingual speech databases for in-car applications: some first validation results. In *Proceedings of Eurospeech*, volume 5, pages 2279–2282. Budapest, Hungary, Sept. 1999.

[van Kampen, 1992] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier Science Publishers, 1992.

[Viikki and Laurila, 1998] A. Viikki and K. Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25, 1998.

[Vintsyuk, 1971] T. K. Vintsyuk. Elementwise recognition of continuous speech composed of words from a specified dictionary. *Cybernetics*, 7:133–143, March 1971.

[Viterbi, 1967] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, IT-13(2):260–269, April 1967.

[Vizinho *et al.*, 1999] A. Vizinho, P. Green, M. Cooke, and L. Josifovski. Missing data theory, spectral subtraction and snr estimation for robust asr : An integrated study. In *Proc. EUROSPEECH'99*, pages 2407–2410, Budapest, 1999.

[Wakita, 1977] H. Wakita. Normalization of vowels by vocal tract length and its application to vowel identification. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 25:183–192, April 1977.

[Ward and Young, 1993] W. Ward and S. Young. Flexible use of semantic constraints in speech recognition. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages 49–50, 1993.

[Weber *et al.*, 2000] K. Weber, S. Bengio, and H. Bourlard. Hmm2- a novel approach to hmm emission probability estimation. In *International Conference on Spoken Language Processing (ICSLP 2000)*, pages III.147–150, Beijing, China, October 2000. IDIAP-rr 00-30.

[Weber, 2003] K. Weber. *HMM Mixtures (HMM2) for Robust Speech Recognition*. PhD thesis, Swiss Federal Institute of Technology Lausanne (EPFL), 2003.

[Welch, 1967] P. D. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.*, AU-15:70–73, June 1967.

[Wittkop, 2001] T. Wittkop. *Two-channel noise reduction algorithms motivated by models of binaural interaction*. PhD thesis, University of Oldenburg, Germany, 2001.

[Wrigley and Wright, 1991] E.Ñ. Wrigley and J. H. Wright. Computational requirements of probabilistic lr parsing for speech recognition using a natural language grammar. In *in Proc. Eurospeech*, pages 761–764, Bristol, UK, 1991.

[Wu, 1983] C. F. G. Wu. On the convergence properties of the *EM* algorithm for gaussian mixtures. *The Annals of Statistics*, 11(1):95–103, 1983.

[Yang *et al.*, 2000] H. H. Yang, S. Sharma, S. van Vuuren, and H. Hermansky. Relevance of timefrequency features for phonetic and speakerchannel classification. *Speech Communication*, 31(1):35–50, Aug 2000.

[Yapanel *et al.*, 2002] U. Yapanel, X. Zhang, and J. H. L. Hansen. High performance digit recognition in real car environments. In *Proc. ICSLP*, pages 793–796, Denver, USA, 2002.

[Young *et al.*, 2005] S. Young, G. Evermann, M. Gales, T. Hain, D. Tershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woolland. *The HTK book (for HTK version 3.3)*. Cambridge University Engineering Department, April 2005.

[Zavaliagkos *et al.*, 1995] G. Zavaliagkos, R. Schwartz, and J. McDonough. Maximum a posteriori adaptation for large scale hmm recognizers. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 725–728, Detroit, USA, 1995.

[Zhu and Ghahramani, 2002] X. Zhu and Z. Ghahramani. Towards semi-supervised classification with markov random fields. Technical report, (Technical Report CMU-CALD-02-106). Carnegie Mellon University., 2002.

[Zue, 1997] V. Zue. Conversational interfaces: Advances and challenges. In *Proc. of EUROSPEECH*, pages 9–18, 1997.