UNIVERSIDAD DE ZARAGOZA

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

TESIS DOCTORAL

AUTOR: Luis Buera Rodríguez

Grupo de Tecnologías de las Comunicaciones, GTC
Instituto de Investigación en Ingeniería de Aragón, I3A

DIRECTOR: Eduardo Lleida Solano

ZARAGOZA, 2007

Agradecimientos

A partir de mi escasa experiencia he podido comprobar que pocos son los hechos que no tienen detrás un precursor meridianamente claro: personas o acciones que desencadenan, de forma directa o indirecta, nuevos proyectos. En este sentido, sería injusto no considerar a Eduardo como el primer responsable de este trabajo que, aunque formalmente finalizado, no acaba más que de empezar. Por ello mi primer agradecimiento es para él, por ese tiempo robado a sus otras actividades, ayuda y apoyo.

También quisiera agradecer a todo el grupo de Tecnologías de la Voz de la Universidad de Zaragoza por todas las discusiones y trabajos conjuntos que, a lo largo de estos algo más de cuatro años, me mostraron en muchas ocasiones puertas donde sólo veía muros. Ciertamente resulta muy sencillo trabajar al lado de Antonio, Alfonso u Óscar.

Durante las distintas estancias breves que he realizado a lo largo de mi doctorado he podido aprender no sólo aspectos científico-técnicos, sino también nuevos métodos de trabajo e incluso diferentes culturas que me han ayudado a enfrentar los problemas de un modo más abierto y racional. De este modo, Juan Arturo y Paola en el TEC de Monterrey, o Álex y Jasha en Microsoft Research también son responsables en buena parte de este trabajo.

La tranquilidad es, opino, fundamental a la hora de realizar un trabajo como una tesis doctoral. Por ello, es justo reconocer que la financiación proporcionada por el Ministerio de Educación y Ciencia del Gobierno español y la Universidad de Zaragoza me ha proporcionado la calma necesaria para investigar sin ningún tipo de presión.

No quisiera dejar pasar esta oportunidad para agradecer a mi familia su apoyo y comprensión en todas y cada una de las decisiones que he tenido que ir tomando durante el desarrollo de la presente tesis. Sé que algunas, especialmente los primeros años, no fueron sencillas, pero nunca recibí otra respuesta que no fuera la comprensión. Pedro, Luisa, Borja y Elena, trabajar respaldado lo hace todo más fácil.

Todo aquel que ha trabajado en un proyecto largo, y una tesis doctoral creo que es uno de ellos, sabe que los buenos y malos momentos se alternan inexorablemente. Estar cerca en los primeros es sencillo, e incluso agradable, pero en los segundos no lo es tanto, principalmente porque perdemos la noción de lo que nos rodea y no vemos más allá de unos problemas que, para otros, no dejan de ser nimios. Por ello quisiera agradecer a mis amigos el apoyo prestado tantos jueves durante estos años. Gracias Carol, Diego, Silvia, Óscar, Miriam y Myriam.

Por último, no quisiera dar satisfacción alguna a ese modo de la memoria llamado olvido; por ello, las siguientes líneas en blanco representan mi agradecimiento a todos aquellos que, de un modo u otro, han convivido conmigo durante mis veintinueve años porque, guste o no, somos lo que los demás nos han hecho ser.

Resumen

El Reconocimiento Automático del Habla, RAH, pretende, dada una señal acústica, extraer la correspondiente secuencia de palabras pronunciadas. Bajo ciertas condiciones controladas, que comprenden todos los ámbitos que rodean al sistema de RAH, éste llega a proporcionar satisfactorias tasas de reconocimiento. Sin embargo, esto no se suele dar en situaciones reales, por lo que la robustez pasa a tener un papel primordial. En este sentido, uno de los campos en los que más se ha trabajado en los últimos años es la compensación de los efectos negativos que el entorno acústico puede introducir.

La tesis doctoral "Normalización y Adaptación a Entornos Acústicos para la Robustez en Sistemas de Reconocimiento Automático del Habla" versa sobre el uso de diversas técnicas de robustez ante el entorno acústico. Dichas técnicas de compensación comprenden tanto la proyección de los vectores de características ruidosos sobre el espacio representado por los modelos acústicos de referencia, lo que se denomina adaptación de la señal a los modelos acústicos, como la transformación de los propios modelos acústicos de referencia acercándolos al espacio asociado a los vectores de características, también conocida como adaptación de los modelos acústicos a la señal. En ambos casos se ha trabajado principalmente con técnicas empíricas no supervisadas, esto es, que no precisan del conocimiento de la trascripción de la señal empleada en la fase de entrenamiento.

En cuanto a los métodos de adaptación de la señal a los modelos acústicos, cabe reseñar que se ha desarrollado el algoritmo empírico *Multi-Environment Model-based LInear Normalization*, MEMLIN, que se sustenta en tres aproximaciones, a saber: modelar el espacio limpio y ruidoso con sendas *Gaussian Mixture Model*, GMM, y asumir que los vectores de características limpio y degradado se relacionan entre sí a partir de una transformación lineal de orden uno y pendiente unidad para cada par de Gaussianas, entendiendo por par de Gaussianas la combinación de una correspondiente al espacio limpio y otra al espacio degradado. Diversas experimentaciones con la bases de datos *SpeechDat Car* en español y *Aurora* 2 demostraron el satisfactorio comportamiento del algoritmo, reduciendo las tasas de error obtenidas previamente con técnicas como *multivariate Gaussian-based cepstral normalization*, RATZ, o *Stereo based Piecewise LInear Compensation for Environments*, SPLICE.

Si se estudia detenidamente la técnica MEMLIN, se puede observar que hay dos estimaciones que afectan en gran medida al comportamiento final del algoritmo. Éstas no son otras que el modelado del espacio de señal, que matemáticamente viene dado por la transformación asociada a cada par de componentes, y el modelado de la probabilidad condicionada entre espacios de señal, cuyo reflejo matemático se materializa en la probabilidad a posteriori de la Gaussiana del modelo limpio dada la del modelo degradado. En ambas líneas se ha trabajado a lo largo de esta tesis doctoral.

Buscando una transformación asociada a cada par de Gaussianas más realista, se definieron los algoritmos *Polynomial Multi-Environment Model-based LInear Normalization*, P-MEMLIN, que emplea un polinomio de orden uno cuya pendiente puede ser diferente de la unidad, *Multi-Environment Model-based HIstogram Normalization*, MEMHIN, basada en una función no lineal obtenida a partir de ecualización de histograma y *Phone*

Dependent Multi-Environment Model-based LInear Normalization, PD-MEMLIN, que es la versión dependiente del fonema para la técnica MEMLIN. Mediante estas nuevas transformaciones se buscaba transformar no sólo las medias de los vectores acústicos, sino también las varianzas. Las diferentes experimentaciones mostraron una importante mejora por parte del algoritmo PD-MEMLIN, así como un interesante comportamiento de las técnicas P-MEMLIN y MEMHIN ante ruidos aditivo.

Inicialmente, la probabilidad a posteriori de la Gaussiana del modelo limpio dada la del modelo degradado se estimaba mediante un modelo estático independiente del vector acústico ruidoso. Así se hacía por ejemplo con las técnicas MEMLIN, P-MEMLIN, MEMHIN y PD-MEMLIN. Sin embargo, y apoyado en estudios que desvelaban la fragilidad de la aproximación considerada, se definió una solución más realista consistente en modelar los vectores de características ruidosos asociados a cada par de Gaussianas mediante una nueva GMM. De este modo, las diferentes experimentaciones mostraron que las correspondientes extensiones de los algoritmos MEMLIN y PD-MEMLIN proporcionan unas muy importantes mejoras en términos de tasa de error.

Por otra parte, y en cuanto a adaptación de los modelos acústicos a la señal, se propuso entrenar una serie de matrices de rotación para modificar los modelos acústicos de referencia. Dichas matrices representan la relación entre los vectores acústicos limpios y los normalizados, siendo éstos últimos los obtenidos a partir de cualquiera de las técnicas de compensación anteriormente mencionadas. Del mismo modo que las transformaciones de las técnicas de adaptación de vectores de características ya introducidas, las matrices de rotación están asociadas igualmente a un par de Gaussianas (una del modelo del espacio limpio y otra del modelo del espacio normalizado, que también ha sido previamente representado mediante una GMM). Obsérvese que, en el fondo, la solución propuesta es híbrida en tanto que combina un algoritmo de adaptación de vectores de características con otro de adaptación de modelos acústicos. La experimentación muestra en este caso una muy significativa mejora para las distintas bases de datos consideradas, aunque el mejor comportamiento se logra con el corpus *SpeechDat Car* en español.

En general, todas las técnicas empíricas, como las presentadas en este trabajo, poseen una limitación inherente a ellas mismas. Ésta no es otra que la necesidad de disponer de señal estéreo de entrenamiento para estimar los distintos parámetros que, posteriormente, se precisan a la hora de compensar los vectores acústicos. Para eliminar dicha limitación, se ha propuesto en este trabajo un nuevo proceso de entrenamiento para el algoritmo PD-MEMLIN basado únicamente en la señal degradada. Además, los correspondientes resultados experimentales con el corpus *SpeechDat Car* en español demostraron que la pérdida derivada de emplear sólo la señal degradada en la fase de entrenamiento no es crítica.

Abstract

Given an utterance, Automatic Speech Recognition, ASR, provides usually the most likelihood sequence of words. Under certain conditions, the performance of these systems can be satisfactory. However, the accuracy degrades in real conditions due to the acoustic conditions, being the robustness very important in those conditions.

The thesis "Normalización y Adaptación a Entornos Acústicos para la Robustez en Sistemas de Reconocimiento Automático del Habla" is based on the use of several robustness techniques. These techniques can be feature vector normalization methods, which map recognition space feature vectors to the training space, and acoustic model adaptation methods, which map acoustic models from training space to recognition space. In both cases, the main proposed algorithms are empirical and unsupervised, so that the transcription is not needed in the corresponding training phase.

To improve the results obtained using state-of-the-art empirical feature normalization methods, we propose several solutions based on the joint modelling of clean and noisy space. We present Multi-Environment Model-based LInear Normalization (MEMLIN), which splits noisy space into several basic environments and models each basic noisy and clean feature spaces using GMMs. Furthermore, noisy and clean feature vectors are related by a linear function based just on a bias vector transformation for each pair of Gaussians (clean and noisy model ones). Different experiments were carried out with *SpeechDat Car* in Spanish and *Aurora* 2 corpora, obtaining an interesting improvement concerning techniques like multivariate Gaussian-based cepstral normalization, RATZ, or Stereo based Piecewise LInear Compensation for Environments, SPLICE.

Thus, most empirical feature vector normalization methods compute a bias vector transformation for each clean model Gaussian, e.g., RATZ, each noisy model Gaussian, e.g., SPLICE, or each pair of clean and noisy model Gaussians, e.g., MEMLIN. In this thesis, we propose several approximations to modify the simple bias correction term used in MEMLIN. A first-order polynomial transformation, Polynomial Multi-Environment Model-based LInear Normalization (P-MEMLIN) addresses the use of a different slope and bias term for each pair of clean and noisy model Gaussians. A non-linear transformation, Multi-Environment Model-based HIstogram Normalization (MEMHIN) addresses the use of a histogram equalization for each pair of clean and noisy model Gaussians. Those two new methods can compensate the effects of the noise over the means and the variance of the feature vectors. Furthermore, a phone dependent version of MEMLIN is also included, Phone Dependent Multi-Environment Model-based LInear Normalization (PD-MEMLIN) in which the clean and noisy spaces are split into phonemes that are modelled using GMMs. All these algorithms are included in order to improve the transformations associated to each pair of Gaussians. An important improvement was reached with PD-MEMLIN when the experimentation is made with *SpeechDat Car* in Spanish. On the other hand, P-MEMLIN and MEMHIN also improved the results reached by MEMLIN when additive noise is included.

The second critical point in MEMLIN is the estimation of the probability of the clean model Gaussian, given the noisy model one and the noisy feature vector (cross-probability

model). In MEMLIN, P-MEMLIN, MEMHIN and PD-MEMLIN a time-independent solution is considered. In order to overcome this limitation, a time-dependent solution is proposed in this theses. The time-dependent solution consists on modelling the noisy feature vectors associated to each pair of Gaussians from the clean and the noisy spaces with a GMM. The experimentation showed that the corresponding extensions for MEMLIN and PD-MEMLIN provide very important improvements concerning the corresponding algorithms with time-independent cross-probability model.

In this theses we propose also an acoustic model adaptation technique based on rotation transformations over an expanded HMM-state space. Hence, clean and normalized spaces are modelled with GMMs and a set of rotation matrices is obtained, estimating one matrix for each pair of Gaussians (clean-normalized) with stereo normalized and clean data in a previous unsupervised training process using linear regression. In recognition, each normalized feature vector is decoded with the expanded acoustic models, which are generated from the reference ones and the set of rotation matrices; so that one of the rotation matrices is selected during the search algorithm for each normalized feature vector by using the ML criterion. Thus, shift and rotation degradations are compensated jointly in an unsupervised way. Observe that the final technique can be considered as an on-line unsupervised hybrid solution which combines a feature vector normalization technique (MEMLIN or PD-MEMLIN in this theses) with the novel acoustic model adaptation. The results with *SpeechDat Car* database in Spanish showed very important improvements concerning MEMLIN and PD-MEMLIN.

In many acoustic environments and training databases, stereo data are unavailable. Since most of the proposed techniques needs stereo data and to overcome this limitation, a non-stereo data training algorithm that uses only noisy feature vectors is proposed in this theses. That "blind" technique is applied over the PD-MEMLIN method, obtaining even better recognition results than MEMLIN with *SpeechDat Car* database in Spanish.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción.

1.1 Introducción.

La tesis doctoral "Normalización y Adaptación a Entornos Acústicos para la Robustez en Sistemas de Reconocimiento Automático del Habla" versa sobre el uso de diversas técnicas de compensación de los efectos del ruido propio de los entornos acústicos en los sistemas de Reconocimiento Automático del Habla, RAH. Dichas técnicas de compensación comprenden tanto la proyección de los vectores de características ruidosos sobre el espacio representado por los modelos acústicos (adaptación de la señal a los modelos acústicos), como la transformación de los modelos acústicos de referencia, acercándolos de este modo al espacio asociado a los vectores de características correspondientes (adaptación de los modelos acústicos a la señal).

El RAH es una disciplina científica multidisciplinar que posee, como principal objetivo, 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 son precisamente, como se puede suponer, las más interesantes a la hora de aplicar las tecnologías del 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 encuentren; de modo que, sea cual sea este último, las tasas de reconocimiento se acerquen idealmente a las obtenidas al decodificar la señal limpia. Conseguir este ambicioso objetivo supondría 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 la adaptación

tanto de los vectores de características como de los modelos acústicos, desarrollando e implementando distintas técnicas que, en su conjunto, se han mostrado efectivas ante diversos y variables entornos, obteniendo unos resultados sensiblemente superiores a los logrados con los métodos más habitualmente empleados en la actualidad.

Este Capítulo se articula, dejando al margen esta primera sección, en 4 apartados. En la Sección 1.2 se trata el contexto y motivación que han llevado a la realización de la presente tesis doctoral, mientras que los objetivos que se han pretendido alcanzar se introducen en la Sección 1.3. Por su parte, en la Sección 1.4 se resume brevemente la estructura de la memoria por capítulos. Finalmente, y ya en la Sección 1.5, se incluyen las principales contribuciones logradas a partir del trabajo realizado.

1.2 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 interactúa 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, las interfaces hombre-máquina, si bien 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 las interfaces orales. Tantas, que aplicaciones que hace sólo unos pocos años eran tildadas de inverosímiles ya se han hecho realidad, llegando a formar parte, en algunos casos, del escenario cotidiano de muchas personas. Así, centralitas telefónicas, sistemas de dictado o de ayuda a minusválidos... ya integran en innumerables ocasiones interfaces hombre-máquina basadas 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 sean 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.

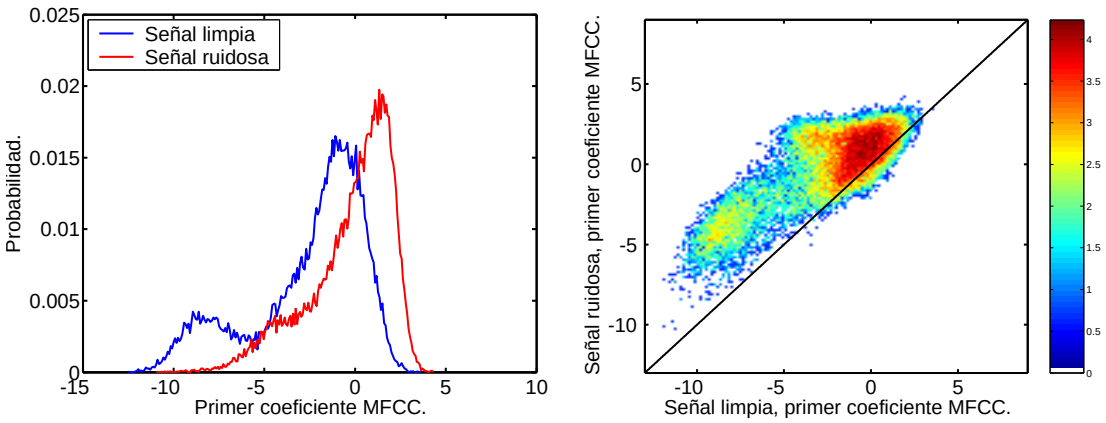
Cuando las condiciones que rodean a la interfaz hombre-máquina no están controladas, la tarea de proporcionar una funcionalidad satisfactoria se complica sobremanera debido principal, aunque no únicamente, a dos causas, a saber: la incapacidad de las interfaces para abordar de forma conveniente situaciones imprevistas y el efecto pernicioso de

los entornos acústicos adversos, que produce generalmente una severa degradación en el comportamiento del sistema de RAH. Para compensar ambas situaciones, que típicamente se dan en condiciones reales, se pueden plantear distintas soluciones que, en general, se agrupan en dos grandes líneas de actuación [Zue97]: flexibilizar el módulo de diálogo, de modo que el usuario alcance el objetivo con el mayor grado de satisfacción global [KWR97] [GRW97], y diseñar sistemas de RAH robustos ante cualquier entorno acústico adverso [CHA$^+$95] [Mor96].

Anteriormente se han nombrado dos de los sistemas de que se compone una interfaz hombre-máquina basada en el habla: el sistema de RAH y el módulo de diálogo. Para completarla, 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 la interfaz por cuanto, al encontrarse a más bajo nivel, los módulos de comprensión y diálogo dependen en gran medida 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 el sistema de RAH, 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, 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 condiciones silenciosas. 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 debido a que el usuario articula los vocablos de distinta manera a como lo haría si se encontrara en un entorno silencioso 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 ella se muestran el *log-scattergram* y el histograma del primer coeficiente *Mel Frequency Cepstral Coefficient*, MFCC, de los vectores de características de voz limpia y degradada, ambas grabadas



(1.a) (1.b)

Figura 1.1: *Log-scattergram* 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 (8.05 dB de SNR media). La línea en el *log-scattergram* representa la función identidad $x = y$.

en un vehículo durante el proceso de conducción (*Signal to Noise Rate*, SNR, media de 8.05 dB). Se puede apreciar como las condiciones que rodean a dicho entorno han 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.

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

- Parametrización o extracción robusta de características, esto es, obtener de la señal de voz aquella información que, siendo útil para el RAH, se vea lo menos afectada posible 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 transformara 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.

Igualmente existe la posibilidad de desarrollar líneas híbridas que combinen algunas de las soluciones anteriores. Asimismo cabe destacar que, en general, no se puede asegurar a priori 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 introduce 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.

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 de técnicas de adaptación tanto del sistema de RAH, como de la señal que se pretende reconocer, haciendo especial hincapié en esta última línea de actuación. 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.

1.3 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 las 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 hasta el momento de cara a proporcionar robustez a los sistemas de RAH, se decidió, desde un primer momento, enfocar la tesis doctoral principalmente hacia el desarrollo de técnicas de adaptación de vectores de características, puesto que poseen 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. Dicho esto, tampoco se dejó de lado la línea de investigación basada en la adaptación de modelos acústicos, dedicando a ella también un importante esfuerzo y buscando en todo momento la manera óptima de conjugarla con la línea de actuación anterior. Tanto en uno como en otro caso se primó en todo momento el desarrollo de algoritmos no supervisados, por ser éstos mucho menos restrictivos a la hora de su aplicación.

Así pues, y una vez centrado el dominio del trabajo, el objetivo final es único: proporcionar la menor tasa de error, mejorando en la medida de lo posible los trabajos que, en este campo, se hayan presentado 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. De este modo en este trabajo, y salvo en escasas ocasiones, se proponen métodos no supervisados, esto es, que no es necesaria la trascripción de los corpora correspondientes empleados en la fase previa de entrenamiento que dichos métodos propuestos puedan precisar. Asimismo, se trabajará principalmente en el ámbito de la normalización empírica de vectores de características, para lo que se suele asumir que los corpora de entrenamiento representan de un modo fiel a los entornos acústicos

que posteriormente se darán en la fase de reconocimiento; aunque en este trabajo se presentan igualmente 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 cualquier tipo de distorsión. Por esto último se consideró *SpeechDat Car* en español como base de datos de referencia para realizar la mayor parte de la experimentación, ya que se grabó directamente en un vehículo mientras el locutor conducía. Dicho esto también se creyó oportuno trabajar con el corpus *Aurora* 2 que, si bien está compuesto por señales degradadas de un modo artificial, actualmente está considerado como un banco de pruebas estándar sobre el que comparar las distintas técnicas de robustez para sistemas de RAH. Adicionalmente se presentan también resultados con el corpus *Hiwire*.

Así pues, con la vista puesta en obtener la mínima tasa de error posible, y teniendo en cuenta las premisas anteriores, se puede definir una serie de subobjetivos que, a lo largo del desarrollo de la presente tesis doctoral, 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 desarrolladas, 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, *Aurora* 2 y *Hiwire*, que fueron analizadas convenientemente para 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.

- Estudio e implementación de técnicas de referencia.

  A partir de la revisión bibliográfica es posible determinar aquellos algoritmos que se adecuan de un modo más conveniente a los parámetros que van a regir la investigación y que, por tanto, constituyen las técnicas de referencia con las que se deberán comparar posteriormente los métodos que se vayan desarrollando a lo largo del presente trabajo. En la primera parte del Capítulo 5 se estudian dichas técnicas de referencia, planteando a su vez algunas de las posibles limitaciones que poseen.

- Desarrollo de nuevas técnicas y depurado de las mismas.

  Como consecuencia de un estudio concienzudo de las técnicas de referencia 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 igualmente 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 ciertamente precisa de los anteriores, es el motor que ha movido todo el trabajo desarrollado, de modo 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 limitaciones detectadas tras la experimentación y posterior estudio de los distintos métodos.

1.4 Estructura de la Memoria.

La memoria se divide en once 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 de robustez propuestas e implementadas durante el desarrollo de la presente tesis doctoral, incluyendo convenientemente los resultados obtenidos tras las correspondientes experimentaciones; finalmente en el Capítulo 11 se presentan 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, y en la Sección 2.2, se estudia, desde el punto de vista matemático-estadístico, un sistema de RAH clásico, tratando por separado, y en Secciones diferenciadas, 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 (2.3) 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 la Sección 2.4. 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 (2.5). En ella se realiza un breve repaso de 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 especialmente en la opción más utilizada en estos momentos: las N-gramas.

Los procedimientos de búsqueda constituyen, en cierto modo, el motor de todo 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 (2.6) 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 un sistema clásico de RAH, en el tercer Capítulo se aborda, desde un punto de vista taxonómico, las más importantes 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, y como ya se ha adelantado, tres grandes líneas de actuación, a saber: extracción robusta de características, adaptación de modelos acústicos al espacio definido por los vectores de características que se pretenden decodificar y adaptación de vectores de características hacia el espacio representado por los modelos acústicos. Cada una de las líneas se trata de modo independiente en una Sección propia dentro del Capítulo.

  En la Sección dedicada a la extracción robusta de características (3.2), una vez explicada la filosofía de actuación que subyace bajo dicha línea de actuación, se realiza un breve repaso de aquellos algoritmos que, enmarcados en ella, son los más habituales, haciendo especial hincapié tanto en los beneficios como en las limitaciones que cada uno de ellos posee.

  La adaptación de modelos acústicos al espacio definido por los vectores de características que se pretenden decodificar, como segunda opción a la hora de dotar de robustez a los sistemas de RAH, se trata en la Sección 3.3. 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 (3.4) versa sobre la adaptación de los vectores de características hacia el espacio representado por los modelos acústicos. 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 cualitativa de las limitaciones y ventajas que los diferentes métodos poseen.

- Cuarto Capítulo.

  De igual modo que los Capítulos segundo y tercero proporcionan las bases conceptuales 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 que permitirán comparar de un modo cuantitativo los distintos métodos que, a lo largo del trabajo, se van a ir presentando. Por ello, en la Sección 4.2 se estudian brevemente las tres bases de datos que se van a emplear a tal efecto: *SpeechDat Car* en español, *Aurora* 2 y *Hiwire*, si bien es cierto que el grueso de los resultados se obtendrán con la primera de ellas por ser mucho más realista que el resto, que se genera

tras incluir artificialmente ruido aditivo y/o distorsión convolucional a alocuciones limpias procedentes del corpus *TIDigits*, en el caso de el corpus *Aurora* 2, o bien a frases grabadas por locutores no nativos, si se trata de la base de datos *Hiwire*.

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 (4.3) está dedicada a las pruebas de hipótesis estadísticas. Así, en dicha Sección se presentan los tres algoritmos más ampliamente utilizados a tal efecto en el ámbito del RAH, para, finalmente, centrarse en el denominadado *z-test*, que será el que, a pesar de sus limitaciones, se empleará en este trabajo.

Finalmente, y como último apartado de este Capítulo (4.4), 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 para las bases de datos *SpeechDat Car* en español, *Aurora* 2 y *Hiwire* y que servirán como base para comparar posteriormente los comportamientos de las distintas técnicas presentadas en este trabajo.

- Quinto Capítulo.

  El Capítulo quinto supone la puerta de entrada del segundo gran grupo temático en que ha quedado dividida 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 los vectores de características. Así, la Sección 5.2 está dedicada al análisis, tanto desde un punto de vista teórico como cualitativo, de los efectos que distintos tipos de entornos acústicos producen, pudiéndose apreciar las correspondientes alteraciones que se generan en los vectores acústicos limpios.

  Una vez estudiado el problema que se pretende tratar, y tras constatar las dificultades que solucionarlo conlleva, se propone un desarrollo teórico conjunto para las técnicas más empleadas de adaptación empírica de vectores de características basadas en el criterio bayesiano *Minimum Mean Square Error*, MMSE, (Sección 5.3). De este modo se puede apreciar que algoritmos ampliamente utilizados como *Cepstral Mean Normalization*, CMN, *multivariate Gaussian-based cepstral normalization*, RATZ, o *Stereo based Piecewise LInear Compensation for Environments*, SPLICE, 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 en la Sección 5.4 la técnica de adaptación empírica de vectores de características *Multi-Environment Model-based LInear Normalization*, MEMLIN, 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 (5.5), 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, analizando nuevas posibilidades, mejorar el modelo del espacio de señal, que para el método MEMLIN se supone lineal con término dependiente unidad. Así, en la Sección 6.2 se asume un modelo 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 Sección 6.3 se presenta el algoritmo *Multi-Environment Model-based HIstogram Normalization*, MEMHIN, en el que se asume como modelo del espacio de señal una función no lineal estimada a partir de ecualización de histograma. Obsérvese que en el fondo esta solución no deja de ser una generalización de las dos anteriores, cuyos modelos propuestos 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 Sección 6.4 de este Capítulo, la técnica *Phoneme Dependent Multi-Environment Model-based LInear Normalization*, PD-MEMLIN, que es una extensión del algoritmo MEMLIN dependiente de unidades fonéticas.

  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 6.5 una fase de entrenamiento "ciega" para la técnica PD-MEMLIN, esto es, que no precisa de señal estéreo de entrenamiento.

  Como punto final del Capítulo, la Sección 6.6 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 respecto al algoritmo MEMLIN para el corpus de experimentación seleccionado, 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 degradadas de un modo drástico si se hace uso de la fase de entrenamiento "ciega".

- Séptimo Capítulo.

  Tras desarrollar en el Capítulo anterior nuevos modelos del espacio de señal 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 de probabilidad condicionada entre espacios de señal, término este de gran importancia a la hora de estimar el 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 Sección 7.2 del Capítulo se realiza un estudio, tanto cualitativo como cuantitativo, del término de modelado de probabilidad condicionada entre espacios de señal para la técnica MEMLIN, pudiéndose 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 a tal efecto modelar convenientemente los vectores de características degradados mediante *Gaussian Mixture Models*, GMMs. Esta solución se introduce y se desarrolla matemáticamente en la Sección 7.3 del Capítulo.

En la siguiente Sección (7.4) se plantea como incluir, tanto desde un punto de vista conceptual como matemático, el nuevo Modelado de Probabilidad condicionada entre espacios de señal propuesto en las técnicas MEMLIN y PD-MEMLIN, dando lugar a los métodos MEMLIN-MP y PD-MEMLIN-MP.

Al igual que en Capítulos precedentes, éste se finaliza con la Sección dedicada a la experimentación con la base de datos *SpeechDat Car* en español (7.5), 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 espacio de la señal como estimando de un modo más eficiente el modelo de probabilidad condicionada entre espacios de señal, en el Capítulo octavo se propone combinar algunas de las técnicas de adaptación de vectores de características presentadas hasta el momento con algoritmos que, con el objetivo de compensar la rotación entre los vectores de características transformados y los limpios, proponen modificar los modelos acústicos. Estas soluciones híbridas se pueden englobar en dos líneas de actuación, según si la adaptación de los modelos acústicos es supervisada o no. Así, dentro de la segunda línea, se presenta en la Sección 8.2 un algoritmo híbrido basado en el cálculo de matrices de rotación dependientes de sendas GMMs, representando una de ellas el espacio limpio y la otra el definido por la señal normalizada.

  En la Sección 8.3 se presentan aquellas soluciones híbridas basadas en reentrenamiento supervisado de modelos acústicos en el espacio normalizado, de modo que los vectores de características compensados se decodifican directamente con dichos modelos acústicos.

  Para comparar las diversas técnicas propuestas en este Capítulo, la última Sección (8.4) 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 con respecto a los resultados obtenidos con las distintas técnicas de adaptación de vectores de características que les sirven de base.

- Noveno Capítulo.

  El Capítulo noveno se articula en cuatro Secciones, en las que se presentan los resultados de RAH obtenidos con la base de datos *Aurora* 2 tras aplicar las técnicas más representativas propuestas en este trabajo. De este modo, la Sección 9.2 está dedicada al algoritmo MEMLIN, pudiéndose observar, 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 Sección 9.3 está destinada a presentar los correspondientes resultados extraídos con la base de datos *Aurora* 2 tras aplicar la técnica MEMLIN con el modelado de

probabilidad condicionada entre espacios de señal propuesto en el Capítulo séptimo, MEMLIN MP. Se puede 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 séptimo.

En la cuarta y última Sección (9.4) se muestran los resultados de RAH sobre la base de datos *Aurora* 2 obtenidos tras aplicar soluciones híbridas basadas en la estimación no supervisada de de matrices de rotación propuestas en el Capítulo octavo. En este caso la técnica de adaptación de los vectores de características elegida es el método MEMLIN MP. Del mismo modo que sucedía con la base de datos *SpeechDat Car* en español, la transformación de los modelos acústicos a partir de matrices de rotación proporciona en algunos casos una interesante mejora.

- Décimo Capítulo.

  El décimo Capítulo comprende únicamente dos Secciones. En ellas se incluyen los resultados alcanzados con la base de datos *Hiwire* tras aplicar una técnica de normalización híbrida. De este modo, en la Sección 10.2 se expone el comportamiento que, ante dicho corpus, posee la técnica MEMLIN con el modelado de probabilidad condicionada entre espacios de señal propuesto en el Capítulo séptimo, MEMLIN MP, conjuntamente con un algoritmo de adaptación supervisada de los modelos acústicos. Tal y como se podrá apreciar, y a partir de las tasas de error alcanzadas, queda reflejada nuevamente la bondad de las técnicas propuestas en el presente trabajo.

- Undécimo Capítulo.

  La memoria finaliza con un Capítulo, el undécimo, dedicado a las conclusiones (Sección 11.2) y futuras líneas de trabajo (Sección 11.3) que, apoyándose en las técnicas, resultados e ideas propuestas a lo largo de la memoria, podrían hipotéticamente llegar a dar lugar a nuevos algoritmos que cubran alguno de los vacíos que los ya propuestos poseen. Asimismo, el Capítulo finaliza con los índices de calidad (Sección 11.4), en los que se incluyen las publicaciones, proyectos y méritos acumulados a lo largo de la realización de la presente tesis doctoral.

1.5 Principales Contribuciones.

Durante los algo más de cuatro años que han sido necesarios para completar esta tesis doctoral 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 [BLMO04a] [BLO$^+$04] se presentó la técnica MEMLIN, comparando su comportamiento ante diversos entornos acústicos adversos con respecto a otros métodos ya clásicos de adaptación de vectores de características basados en similares principios. 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 supeditada, eso sí, a ciertas aproximaciones.

Posteriormente, y tras analizar la técnica MEMLIN, se llegó a la conclusión de que el modelo de espacio de señal 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 [BLMO04b], dando lugar así a la técnica MEMHIN, y por el otro, considerar transformaciones dependientes de las unidades fonéticas [BLMO05c] [BLMO05a], naciendo de este modo el algoritmo PD-MEMLIN. Para el primero de los casos se pudo observar una importante mejora en los resultados ante entornos acústicos caracterizados por ruido aditivo, ya que mediante la ecualización se pretende compensar las degradaciones producidas sobre la varianza de los vectores de características. 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 se aplica a estas nuevas tareas bajo entornos acústicos hostiles [BLMO05a] [BLR$^+$05] [BLR$^+$06]. 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 poseen en muchas ocasiones las técnicas de adaptación empírica de vectores de características, 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 tenido en cuenta para algunas de las técnicas empíricas más utilizadas por la comunidad científica, sí se consideró en [BLMO05b], donde se presenta una fase de entrenamiento no estéreo para la técnica PD-MEMLIN, pudiéndose observar además, 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 final del método.

La técnica MEMLIN, así como todas aquellas variantes de la misma basadas en nuevos modelos de espacio de la señal, 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 [BLM$^+$07]. De este modo, y por establecer una relación entre las distintas contribuciones con la estructura de la memoria, se puede concluir que 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 espacio de la señal 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 probabilidad condicionada entre espacios de señal. 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 [BLN+06] [BLM+06] [BML+07a] hacer uso de un modelo basado en GMMs para los vectores de características ruidosos, obteniéndose así importantes mejoras, tanto si se aplica dicha solución a la técnica MEMLIN [BLN+06] [BML+07a], o al método PD-MEMLIN [BLM+06]. Tanto las ideas y conceptos como los consiguientes desarrollos teóricos necesarios para estimar los nuevos modelos de probabilidad condicionada se contemplan, junto con la experimentación realizada, en el Capítulo 7.

Una vez desarrolladas diferentes soluciones para mejorar los modelados de espacio de la señal y de la probabilidad condicionada entre espacios de señal, se consideró la opción de compensar el efecto de rotación que los entornos acústicos introducen en los vectores de características, dando lugar de este modo a las soluciones híbridas expuestas en el Capítulo 8. Éstas, compuestas por sendos métodos de adaptación de vectores de características y de modelos acústicos, se clasifican atendiendo a la naturaleza supervisada o no supervisada de estas últimas técnicas. Los métodos incluidos en el primer tipo se trataron en [BML+07a] y [BMS+07], mientras que en [MBL+07] se introdujo una primera aproximación a las soluciones propias del segundo tipo basadas en el cálculo de matrices de rotación, aunque la versión definitiva de las mismas se expuso en [BML+07b] [BMS+07] [BML+07c], trabajo este que está sirviendo actualmente como base para una futura publicación en *IEEE Transactions on Speech Language and Audio Processing*.

Adicionalmente, y como colaboración con el Departamento de Ciencias Computacionales del Tecnológico de Monterrey, campus Monterrey, se presentó el trabajo [HGN+07], en el que se emplea la técnica PD-MEMLIN tras aplicar, como paso previo, el método *Spectral Subtraction*, SS, dando lugar al algoritmo *Phoneme Dependent Multi-Environment Enhanced Model based LInear Normalization*, PD-MEEMLIN. De este modo se trata de compensar en primera instancia los efectos producidos por el ruido aditivo para, posteriormente y en un segundo paso, actuar sobre el resto de distorsiones introducidas por el entorno acústico. La experimentación presentada en el trabajo se realizó sobre la base de datos *Aurora* 2, dando lugar a unos resultados especialmente competitivos ante entornos muy hostiles. Sin embargo, en esta memoria no se han incluido dichos resultados por tratarse aún de una línea de trabajo emergente. Actualmente se sigue colaborando para formalizar la técnica PD-MEEMLIN, buscando sus debilidades e incorporando modificaciones que proporcionen un mejor comportamiento también ante entornos acústicos moderadamente adversos.

Capítulo 2

Sistemas de Reconocimiento Automático del Habla.

2.1 Introducción.

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 previamente captada mediante un micrófono o sensor. A pesar de que las primeras aproximaciones serias datan ya de los años 70 [Bau72] [Jel76], 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.

En la Tabla 2.1 [HAH01], 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 seres humanos, "WER humanos (%)", como por un sistema convencional de RAH situado a la vanguardia del estado del arte, "WER máquinas (%)". Se ha incluido asimismo junto a cada tarea el número de palabras de que se compone el vocabulario de la aplicación correspondiente, donde , WSJ hace referencia a la base de datos *Wall Street Journal* [PB92]. A partir de 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 es posible reconocer cualquier secuencia de trigramas, el sistema de RAH proporciona mejores resultados que los seres humanos, lo que es debido a que éstos dan una mayor importancia al contexto léxico.

Cuando las condiciones que rodean a los sistemas de RAH no están controladas, el proporcionar tasas de reconocimiento aceptables 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 (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 indirectamente el propio proceso de producción de la misma mediante el efecto

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

Tabla 2.1: Tasas de error por palabras, *Word Error Rate*, WER, obtenidas para distintas tareas tanto por seres humanos (columna "WER Humanos (%)") como por un sistema convencional de RAH situado a la vanguardia del estado del arte (columna "WER Máquinas (%)"). La columna "# Vocabulario" indica el número de vocablos de que se compone cada una de las tareas concretas, siendo WSJ la base de datos *Wall Street Journal*.

Lombard [Lom11]. 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 de la misma, 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 suficientemente controlado. Así pues, y atendiendo al mayor o menor grado de acotación de los problemas anteriormente comentados, los sistemas de RAH se clasifican atendiendo a distintos criterios [Moo90]

- 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 características se invierten para el caso de sistemas independientes del locutor. Asimismo, también existen soluciones intermedias, como las multilocutor y las adaptadas al locutor. En la primera de ellas, el sistema de RAH está pensado para ser empleado por un grupo reducido de usuarios, mientras que los sistemas adaptados al locutor parten de uno independiente del locutor y, tras modificar los parámetros necesarios, lo acercan a las prestaciones y condiciones de uno dependiente del locutor. Por otra parte, y dentro de este mismo criterio, los sistemas de RAH se pueden clasificar igualmente atendiendo a los grados de cooperación y experiencia de los locutores.

- 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. Sin embargo, y como contraprestación, tanto el coste computacional como las tasas de error se resienten; por todo ello resulta primordial ajustar al máximo el vocabulario a la tarea de reconocimiento en cuestión. 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 distingue desde reconocer palabras aisladas hasta habla continua, pasando por cualquier estado intermedio. Cuando se pronuncian palabras aisladas, esto es, con importantes pausas entre ellas, las tasas de reconocimiento son mucho mayores que si, por el contrario, 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 también en cuenta el nivel de rechazo ante habla extraña, lo que no deja de ser otra clasificación dependiente del tipo de discurso.

- Dependencia con respecto a la estructura del diálogo. En este caso los sistemas de RAH se diferencian atendiendo a su capacidad de procesamiento del lenguaje, clasificando de forma gradual los sistemas a partir de la perplejidad o de la tarea. Así, se distingue desde el reconocimiento de comandos aislados, que se corresponde con el nivel más bajo, hasta el 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 da una idea de las posibles fuentes de variabilidad que se pueden presentar a la hora de plantear una cierta aplicación, elementos estos de capital importancia ya que la robustez del sistema ante los mismos determina en muchos casos el rendimiento final del sistema.

Este Capítulo, centrado en el estudio de los sistemas de RAH, se estructura del siguiente modo: en la Sección 2.2 se explicarán los fundamentos matemático-estadísticos de los sistemas de RAH basados en un enfoque probabilístico bayesiano, que son los más comúnmente empleados hasta la fecha. Posteriormente, y en las siguientes Secciones, se irá desgranando cada uno de los componentes que constituyen el sistema de RAH completo. Así, en la Sección 2.3 se tratará el proceso de la extracción de las características a partir de la señal de voz, que determinará a la postre los vectores acústicos que se utilizarán a la hora de reconocer. En la Sección 2.4 se presentarán los distintos tipos de modelado acústico que actualmente se están empleando. Por su parte, la Sección 2.5 está dedicada al modelado del lenguaje que, junto con el acústico, compone la base estadística de los sistemas de RAH más utilizados hasta la fecha. Finalmente, en la Sección 2.6 se explica cómo se realiza la búsqueda entre las palabras del vocabulario para dar con la secuencia de las mismas más probable dado un conjunto de vectores acústicos.

2.2 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 distintos principios, en la actualidad las aproximaciones más utilizadas poseen un enfoque probabilístico basado en el teorema de Bayes, la teoría de la información, las técnicas de reconocimiento de patrones y la programación dinámica [BK65] [Ney90] [Ney93] [DHS00].

Desde esta concepción probabilística, 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, devuelva la o las secuencias de patrones que con mayor probabilidad lo representen.

El conjunto de observaciones acústicas, $\mathbf{O}$, también conocidas como vectores de características, consiste en una secuencia de vectores de parámetros que se extraen de la señal de audio captada previamente mediante un sensor o micrófono. Con dicha extracción se pretende utilizar únicamente aquellas cualidades de la señal de voz más representativas y útiles para el ámbito del RAH.

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

donde t es el índice temporal asociado a las observaciones acústicas, o, $t \in [1, T]$. 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, w, $n \in [1, N]$. La óptima secuencia de palabras reconocidas, siempre desde el enfoque probabilístico, sera aquélla que proporcione la mayor probabilidad a posteriori de estar asociada al conjunto de observaciones acústicas de entrada, $p(\mathbf{W}|\mathbf{O})$, lo que matemáticamente se expresa mediante

$$\mathbf{W} = \underset{\mathbf{W}}{arg\,max} \left\{ p(\mathbf{W}|\mathbf{O}) \right\}. \tag{2.3}$$

Dado que la maximización propuesta en (2.3) 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 este modo (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.

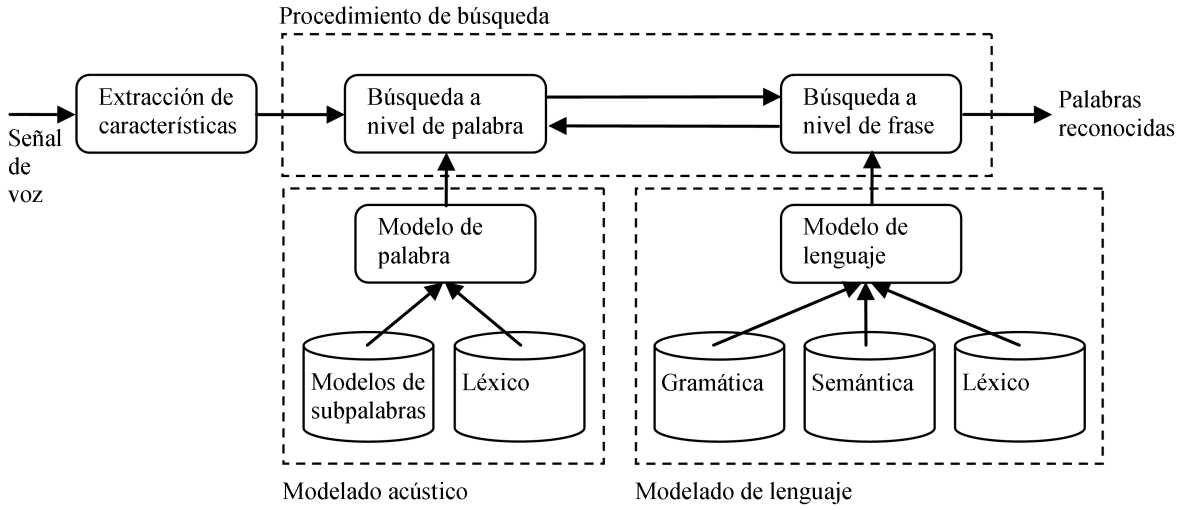


Figura 2.1: Esquema general de un sistema de Reconocimiento Automático del Habla, RAH, basado en un enfoque estadístico Bayesiano, donde quedan patentes los distintos bloques fundamentales que lo componen: "Extracción de características", "Modelado acústico", "Modelado de lenguaje" y "Procedimiento de búsqueda".

$$\mathbf{W} = \underset{\mathbf{W}}{arg\,max} \left\{ \frac{p(\mathbf{O}|\mathbf{W})p(\mathbf{W})}{p(\mathbf{O})} \right\}, \tag{2.4}$$

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 [RJ93]. De esta manera, la expresión que se debe evaluar para obtener la secuencia óptima de palabras pronunciada es

$$\mathbf{W} = \underset{\mathbf{W}}{arg\,max} \left\{ p(\mathbf{O}|\mathbf{W})p(\mathbf{W}) \right\}. \tag{2.5}$$

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, a saber: extracción de las características de la señal de voz, modelado acústico, modelado del lenguaje y procedimiento de búsqueda de la secuencia de palabras óptima, todos ellos representados de un modo esquemático en la Figura 2.1. A continuación se comenta brevemente cada uno de los bloques

- Extracción de características de la señal de voz. Los sistemas de RAH no utilizan directamente las muestras de audio como entrada, sino que éstas se preprocesan para obtener 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 más en detalle en la Sección 2.3.

- Modelado acústico. Este sistema describe la probabilidad de observar un conjunto de vectores acústicos dada una secuencia de palabras, $p(\mathbf{O}|\mathbf{W})$. Típ icamente los modelos acústicos de palabras, que suelen ser os más empleados, se construyen a

partir de modelos de unidades menores, o subpalabras, haciendo uso de una serie de reglas que rigen como unir estos últimos ("léxico"). En la Sección 2.4 se abordará con más profundidad este bloque.

- Modelado de lenguaje. Esta unidad cubre el léxico, 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 2.5 trata con más detalle dicho modelado.

- Procedimiento de búsqueda. Dicho bloque tiene como meta encontrar la frase óptima pronunciada, esto es, aquélla que posea la máxima probabilidad a posteriori dada la secuencia de vectores acústicos mediante el teorema de Bayes (2.5). Para ello, y tal y como se aprecia en la expresión anterior, se puede considerar que es necesario una doble búsqueda: una asociada al modelado acústico ("Búsqueda a nivel de palabra") y otra correspondiente al modelado de lenguaje ("Búsqueda a nivel de frase"). La Sección 2.6 proporciona una visión más profunda de esta unidad.

2.3 Extracción de Características.

El objetivo de las distintas técnicas de extracción de características es, a partir de la señal de audio, proporcionar unos vectores de parámetros que, idealmente, deberían cumplir las siguientes tres características

- 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 resultaría complicado 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 adecuen óptimamente a cada aplicación concreta. Adviértase que no todas deben ser tenidas en consideración del mismo modo. Así, por ejemplo, una buena parametrización para sistemas de RAH puede incluir ciertas características del 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, poco útiles para tareas de RAH, pero eficaces por ejemplo a la hora de reconocer locutores.

- Ser robusta, de tal forma que cualquier alteración sobre la señal de voz afecte de la menor forma posible al cálculo de los vectores de características. De este modo, los distintos sistemas de RAH que utilicen la correspondiente parametrización podrían tener un comportamiento satisfactorio aun cuando el desajuste entre los espacios de señal de reconocimiento y el empleado para obtener los modelos acústicos fuera sensible.

En la actualidad dos son los sistemas de extracción de características más comúnmente utilizados por la comunidad científica, [DM80]: los coeficientes *Linear Prediction Coefficients*, LPC, y los *Mel-Frequency Cepstral Coefficients*, MFCC. La parametrización

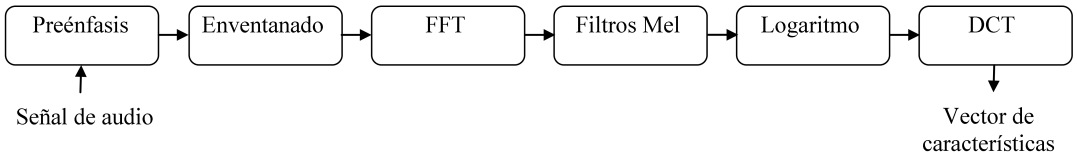


Figura 2.2: Esquema de la parametrización MFCC, donde queda patente los diferentes bloques por los que se ha de pasar la señal de voz hasta obtener el vector de características.

LPC surge al aplicar análisis de predicción lineal a la señal de voz y, en algunas ocasiones, de cara a obtener una representación más adecuada para los sistemas de RAH, se calcula posteriormente el *cepstrum* sobre los propios coeficientes LPC, dando lugar a los parámetros *cepstrum* LPC [OS75] [RJ93]. Por su parte, los coeficientes MFCC se obtienen al transformar el espectro de los coeficientes cepstrales de la señal de voz a la escala bark mediante una transformación Mel. El proceso completo incluye, tal y como se muestra en la Figura 2.2, una fase de preénfasis previa al enventanado para, posteriormente, extraer el espectro a partir de la transformada de Fourier y aplicar los filtros Mel. Finalmente se evalua el logaritmo y, de cara a decorrelar los coeficientes del vector de características final, se aplica la transformación *Discrete Cosine Transform*, DCT.

A pesar de que los coeficientes LPC y MFCC son los más extendidos, a lo largo del tiempo se han propuesto diferentes mejoras y nuevas técnicas en el campo de la extracción de características [Mor97]. Así, por ejemplo, se han desarrollado representaciones basadas en el modelo auditivo humano, como los coeficientes *Perceptual Linear Prediction*, PLP, [Her90], el modelo *Ensemble-Interval Histogram*, EIH, [Ghi92] [Ghi94] [RJ93], o los modelos auditivos asíncronos, como el de Seneff [JHDL95] o el *Synchronous Linear Prediction*, SLP, [JH96]. Todas estas representaciones basadas en el modelo auditivo humano proporcionan, especialmente en condiciones acústicas adversas, buenos resultados si se comparan con los alcanzados haciendo uso de los coeficientes cepstrum LPC [JW89] [JH96] [JHDL95]. Sin embargo el hecho de que el tiempo computacional necesario para la extracción de los parámetros basados en el modelo auditivo humano es, en general, elevado, se puede comprender que este tipo de parametrización 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.

Con el objetivo de proporcionar robustez a los sistemas de extracción de características clásicos han surgido a lo largo del tiempo diferentes soluciones, como por ejemplo: *modulation spectrum* [KMG98] [GK00], en la que mediante el uso de filtros paso bajo se trata de eliminar ciertas componentes de la señal de audio que pudieran resultar negativas de cara a una cierta aplicación de RAH concreta. El uso de modelos perceptuales de enmascaramiento es otra posible solución. En ella se oculta el ruido que pudiera afectar a la señal de voz [UIE94]. La utilización de técnicas más robustas que la transformada de Fourier para estimar el espectro de la señal de audio, como por ejemplo las transformadas wavelets [RV91] [ECY95], es otro campo de estudio. También se ha trabajado en 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 *Teager Energy Operador*, TEO, [Kai90], o en 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 a aquéllas que se encuentren menos contaminadas [BGH00].

A la hora de construir el vector acústico final que 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 [HHG97], o la velocidad de pronunciación [MFM97]. A su vez, y para modelar la correlación temporal existente entre los vectores acústicos próximos, habitualmente se suele hacer uso de la primera y segunda derivadas [Fur86], aunque para ello también se puede estudiar la información presente en fragmentos de señal más amplios que la ventana de análisis [Her98].

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 debería poseer; de ahí que en muchos casos se requiera de técnicas de compensación para paliar algunas de las distintas limitaciones que pueda tener.

2.4 Modelado Acústico.

El modelado acústico, como ya se ha adelantado, tiene como misión el determinar la probabilidad de observar un conjunto de vectores de características dada una secuencia de palabras, esto es, $p(\mathbf{O}|\mathbf{W})$. Para ello se emplean diversas técnicas de aprendizaje que hacen uso, por lo general, de amplios corpora de audio. Dado que el problema de modelar secuencias de una manera completa desde el punto de vista probabilístico puede llegar a ser computacionalmente inviable debido a que la complejidad crece de modo exponencial con la longitud de la propia secuencia, se han venido considerando distintas aproximaciones de independencia que hacen del modelado acústico una tarea más sencilla [Bau72] [Jel76] [Rab88].

Actualmente los modelos ocultos de Markov, *Hidden Markov Models*, HMMs, [Bak75] [Rab88], constituyen la solución más extendida entre los sistemas de RAH. Se trata de unos autómatas de estados finitos a cada uno de los cuales se les asocia una función de densidad de probabilidad, *probability density function*, pdf, que normalmente suele ser una mezcla de Gaussianas, *Gaussian Mixture Model*, GMM, aunque hipotéticamente podría ser cualquier otra distribución. Asimismo los estados se relacionan unos con otros mediante probabilidades de transición. Las aproximaciones de independencia anteriormente comentadas se materializan en este caso mediante dos consideraciones: el proceso estocástico de generación de observaciones sólo depende en cada momento de un estado del modelo, 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 [Gha02], mientras que el entrenamiento de las diversas variables que conforman los HMMs, esto es, las probabilidades de transición entre estados y los parámetros de las funciones de densidad de probabilidad asociadas a cada estado, se realiza generalmente mediante la estimación de máxima verosimilitud, *Maximum Likelihood*, ML, haciendo uso del algoritmo iterativo *Expectation Maximization*, EM, [DLR77].

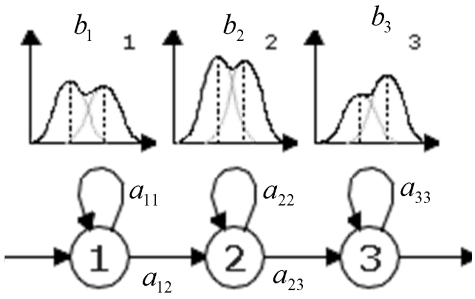


Figura 2.3: Ejemplo esquemático de un modelo oculto de Markov, *Hidden Markov Model*, HMM.

En la Figura 2.3 se presenta un ejemplo de modelo oculto de Markov de orden 1. Se pueden observar los diferentes estados, así como las probabilidades de transición de unos a otros, a_{ij}. Igualmente se han representado las consiguientes funciones de densidad de probabilidad asociadas a cada uno de los estados, b_i.

Tal y como se puede deducir de las aproximaciones de independencia ya 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 [HH94]. Así por ejemplo se desarrolló el modelado basado en la descomposición temporal, previamente empleada en codificación [Ata83], y en la que se considera que la señal de voz es una combinación de ciertas funciones base controladas por un parámetro, [BCDM88] [Del90] [Lle90]. Esta técnica, si bien no se ha seguido desarrollando por la comunidad científica, mantiene algunas similitudes conceptuales con otras extensiones de los HMMs, como los modelos segmentales y los de trayectorias, que posteriormente se pasan a describir someramente.

Los modelos segmentales buscan eliminar alguna de las debilidades de los HMMs de orden 1 mediante la concatenación de submodelos de fragmentos de voz, no teniendo que ser éstos necesariamente de la misma longitud y síncronos como sucede en el caso de los HMMs [ODK96] [Gla03]. Por su parte, los modelos de trayectorias tratan de parametrizar los descriptores de la señal de voz de manera que el término que controla la forma de la señal o de su evolución se pueda adaptar de una manera más fina a sus cambios [Gol94] [Sun95] [GN96].

Otra extensión de los HMMs empleada para el modelado acústico en los sistemas de RAH son los Campos de Markov, *Markov Random Fields*. Esta técnica, concebida inicialmente para tratamiento de imagen [GSC99] [Gra00], modela el espectro tiempo-frecuencia de la señal de voz, de modo que el índice temporal de los estados correspondientes pasa de ser unidimensional a bidimensional, ya que las relaciones de dependencia se definen en este caso como vecindades 2-dimensionales [LMP01] [ZG02].

Los modelos de Markov para la generación de observaciones, más conocidos como HMM2, 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, normalmente una mezcla de Gaussianas, tal y como ya se ha comentado, se sustituye por un nuevo HMM. Con ello se pretende modelar la variabilidad que existe en la evolución de los formantes de la señal de voz [WBB00]. Sin embargo, hasta el momento no se ha conseguido que los HMM2 sean lo suficientemente discriminativos [Web03] como para obtener unos resultados satisfactorios.

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 poder subsanar los errores introducidos por algunos tipos de datos a partir de otros. Para ello habrá que elegir adecuadamente las distintas fuentes de información. Así pues, se distinguen tantas técnicas de modelado conjunto como naturalezas de las señales que se pretenda combinar. De este modo, 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 [BHM96] [BDHM72] busca dotar al sistema de RAH de robustez ante ruidos de banda estrecha, ya que se hace prevalecer 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 [MHB99] [Hag00]. Basándose en el modo en que el ser humano percibe los sonidos, el uso conjunto de señales de voz registradas a partir de dos (binaural) o más sensores también se ha mostrado efectivo bajo ciertas condiciones [Wit01] [Kle02]. 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 minimizar sus debilidades [PPN$^+$03]; asimismo se pueden incorporar igualmente vectores de características asociados a distintas escalas temporales [HB00], ya que en esos casos existe información incluso de niveles más altos que el puramente acústico-fonético [Her98] [SBdlTR01]. Incorporar información visual a la señal de voz es otra clase de fusión que ya se planteó desde los primeros tiempos [SP54]; sin embargo hasta el momento no se han obtenido los resultados deseados debido en buena parte a que la lectura de labios aún no proporciona una aceptable tasa de acierto si no es bajo condiciones de experimentación muy controladas [PNG$^+$03].

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 [RFS01], 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 [US98] [KU03]. Parcialmente relacionados con los modelos de deformación elástica, los modelos de normalización del tracto vocal [Wak77] normalizan la escala frecuencial, tratando de compensar con ello las variaciones inter-locutor, que son en muchos casos las 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 [AKC94] [LR98] [Pit05].

Además de las extensiones consideradas anteriormente, cabe destacar como el mejor conocimiento de los HMMs [GJ97] [GB00], unido a la aparición de nuevas y más completas teorías sobre el aprendizaje estadístico, ha hecho que el modelado acústico sea actualmente una prometedora línea de investigación en el ámbito del RAH que proporciona cada día mejores y más completas extensiones de los clásicos HMMs [JJ94] [Dig92] [PT00] [JJT02].

2.5 Modelado de Lenguaje.

El objetivo del modelado de lenguaje es incorporar el conocimiento lingüístico a los sistemas de RAH, 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 [Lea79] [LHH$^+$89] [PTG$^+$92] [WY93]. Para ello se incluyen aspectos como el léxico, la semántica y la gramática. Todo ello tiene su traducción matemática en el cálculo de la probabilidad a priori de las distintas secuencias de palabras, $p(\mathbf{W})$, hablando siempre de los sistemas de RAH basados en aproximación estadística Bayesiana.

Dado que, al igual que sucedía con el modelado acústico, representar probabilísticamente una secuencia de palabras de un modo completo puede resultar inviable por cuestiones computacionales, se suele acotar la dependencia entre vocablos próximos. Así, suponiendo que las secuencias de palabras siguen un proceso de Markov de orden $(n-1)$ [vK92], 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 [JBM75] y actualmente es el más empleado, aunque no el único.

Las N-gramas pueden entrenarse, de la misma manera que el modelado acústico, mediante el criterio de máxima verosimilitud, ML, utilizando la perplejidad como criterio de evaluación [BJM83]. Para ello sólo se requieren bases de datos de texto y no de audio, lo que es una gran ventaja por ser aquéllas más fácilmente accesibles. Sin embargo, hay que tener en cuenta que si se incrementa el orden del modelo (n), buscando con ello una mayor especificidad de las N-gramas, el tamaño de la base de datos también se debe ampliar considerablemente, pudiéndose dar en muchas ocasiones la ausencia de varias de las posibles agrupaciones de n palabras, lo que generaría modelos erróneos. Para solventar este problema se suelen emplear métodos de suavizado, *smoothing*, [MHJ$^+$99], en los que los parámetros de los modelos de las unidades problemáticas se estiman a partir de los de orden inferior ($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 [Kat87] [NE91] [NEK94] [KNST94] [BPS$^+$92].

A su vez, a lo largo del tiempo se han ido desarrollando distintas técnicas y extensiones para mejorar el comportamiento de los modelos de lenguaje basados en N-gramas. De esta manera, se puede nombrar *language model cache* [KdM90], que, buscando una mayor especificidad, hace uso de las últimas palabras reconocidas para adaptar el modelo de lenguaje a la tarea de RAH en cuestión. Otra posible mejora consiste en agrupar palabras que aparecen habitualmente en el mismo orden para tratarlas como si de una única unidad se tratara [Jel91]. Una idea similar a la anterior, y ya comentada como un método de *smoothing*, consiste en agrupar vocablos 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 [BPS+92].

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 [Ney90] [WW91] [FL94]. Este tipo de gramáticas, si bien más potentes y útiles que las N-gramas en entornos restringidos, tienen el serio inconveniente de que su manejo puede llegar a ser inviable ante tareas complicadas debido a la complejidad de los árboles necesarios para tales casos. A su vez, y aunque las probabilidades de las reglas pueden calcularse de un modo automático, la generación previa de dichas reglas suele ser un proceso manual ya que hasta la fecha los sistemas de generación automática de las mismas no se encuentran actualmente muy desarrollados [SG91] [CDEB91], siendo por ello por lo que las gramáticas de estados finitos no están tan extendidas entre los sistemas de RAH como las N-gramas. 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 [MH82] [PW80] [Shi85] que, al igual que las de estados finitos, precisan de procesos manuales para generar las reglas, lo que las hace, en general, poco apetecibles de cara a su implementación en los sistemas de RAH.

2.6 Procedimiento de Búsqueda.

Mediante el procedimiento de búsqueda se obtiene la secuencia de palabras que maximiza la expresión (2.5), que viene dada, como ya se ha comentado con anterioridad, por el producto de los dos términos referentes a 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 proporcionara un mayor valor. 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 [Bel57], 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* [Jel69] y el de Viterbi [Vit67] [Vin71]. 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 hipotéticos estados que podrían haber generado el correspondiente vector de características. Una vez obtenida dicha lista se realizan, asíncronamente con el tiempo, las proyecciones desde cada estado hacia otra pequeña lista de estados elegida de un modo heurístico. Nótese que de esta manera el resultado final de la técnica depende en gran medida de dicha estimación heurística, lo que no deja de ser un inconveniente. En el algoritmo de Viterbi, por el contrario, 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 de esta manera el uso de métodos de podado, *pruning*, que minimizan todavía más la complejidad de cálculo de la optimización. El objetivo de los métodos de *pruning* consiste en reducir el número de estados sobre los que proyectar a la hora de realizar la búsqueda, de modo que únicamente los hipotéticos estados que previsiblemente van a formar parte de la secuencia de palabras óptima se activan (*beam search* [NMNP87] [ON95]). Si bien la utilización de técnicas de podado proporciona una clara reducción de cómputo, también se puede dar el hecho de que la secuencia de vocablos 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 podado.

Asimismo, cabe destacar que se pueden aplicar diversas técnicas para incrementar la eficiencia de los métodos de podado, como *language model look-ahead* [STN94], 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. Asimismo, 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 haciendo uso de modelos simplificados [NHUTO92], 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 [Fri97], la cuantización de los vectores de características [Boc93], o la partición del espacio de los vectores acústicos [NN96]. También se consigue una importante reducción del coste computacional paralelizando la cálculo de la verosimilitud mediante instrucciones *Single Instruction Multiple Data*, SIMD, [KSN00].

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 muy útil para agilizar el proceso de búsqueda. Así, inicialmente se utilizan modelos acústicos y de lenguaje más simplificados, que proporcionan no sólo la secuencia de palabras más verosímil, sino el conjunto de las N más probables, *N-best*, [SC90], o bien un grafo de palabras [SA91]. Posteriormente, las siguientes iteraciones del sistema de RAH se realizarán únicamente sobre estos últimos resultados aunque con modelos acústicos y de lenguaje cada vez más restrictivos.

Capítulo 3

Robustez en Reconocimiento Automático del Habla.

3.1 Introducción.

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 restricciones deseables consiste en que la señal de audio carezca de ruido tanto en la fase de entrenamiento como en la de reconocimiento. Lamentablemente en una situación ordinaria no se suele dar esta circunstancia, 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 Sección 2.2), y dejando a un lado el modelado del lenguaje, que depende de la tarea en cuestión, y el sistema de búsqueda, que suele considerarse inalterable en la mayoría de las ocasiones, se puede concluir que las técnicas de robustez pueden actuar principalmente bien sobre el modelado acústico, bien sobre la parametrización o extracción de características. Así, se distinguen tres tipos de métodos de robustez [Gon95] [Bel97], a saber: extracción robusta de características, adaptación de los modelos acústicos a la señal que se pretende decodificar y adaptación de la señal que se ha de reconocer a los modelos acústicos. Nótese que el objetivo último de las tres líneas de actuación es, conceptualmente, el mismo, esto es, reducir el desajuste entre los modelos acústicos y los vectores de características producido por el ruido, aunque el modo en que se trata de alcanzar difiere según la línea en cuestión.

La extracción robusta de características busca que los vectores acústicos se vean afectados lo menos posible por el ruido. Por su parte, la adaptación de modelos acústicos pretende transformar éstos últimos con la finalidad de acercarlos a las condiciones con que se extraen los vectores de características que se trata de reconocer; por último, la tercera línea de actuación posible es la adaptación de los vectores de características o normalización, que propone la solución inversa a la línea anterior, esto es, adecuar los vectores de características a los modelos acústicos entrenados bajo las condiciones de referencia. Cabe resaltar llegados a este punto que, en algunas ocasiones, la adaptación

de los vectores de características puede llegar a verse, de un modo general y bajo ciertas circunstancias, como una adaptación de modelos acústicos en la que se recalculan algunos de los parámetros definen a éstos últimos. Asimismo también se puede pensar en soluciones híbridas [NY94] [SL96], que se obtendrían a partir de la combinación de técnicas pertenecientes a distintos tipos de métodos de robustez. Por todo ello, y aunque taxonómicamente resulta interesante dividir la técnicas de robustez en tres posibles líneas, en la práctica no es sencillo hacerlo.

Por lo general, se suele considerar que la adaptación de modelos acústicos proporciona mejores tasas de RAH que cualquiera de las otras soluciones propuestas [NW95] por cuanto en dicho caso se puede representar estadísticamente la aleatoriedad del ruido, que es en último término la causante de la incertidumbre entre las correspondientes realizaciones ruidosas y limpias; hecho este responsable de buena parte 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 propias de la aplicación concreta en cada caso.

Este Capítulo, que versa sobre las técnicas de robustez más comúnmente empleadas por la comunidad científica en los sistemas de RAH, se estructura del siguiente modo: en la Sección 3.2 se resumen brevemente los métodos de extracción robusta de características más representativos. Los algoritmos clásicos enmarcados en la adaptación de modelos acústicos se estudian en la Sección 3.3. Finalmente, y ya en la Sección 3.4, se enumeran y comentan brevemente aquellos métodos incluidos en la adaptación de vectores de características más utilizados en la actualidad.

3.2 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 que representan la señal de voz. Con objeto de obtener técnicas de extracción de características que se vean lo menos afectadas posible por el ruido, se han investigado distintos tipos de soluciones, algunas de las más empleadas son: la utilización de ventanas de *liftering*, distancias basadas en la proyección cepstral o parametrizaciones obtenidas mediante criterios discriminativos o mediante el procesado en sub-bandas. Nótese que en este apartado no se mencionarán, por haber aparecido ya en la Sección 2.3, aquellas técnicas que, aunque nacidas para proporcionar robustez, están basadas en el modelo auditivo humano o en el cepstrum en escala Mel. Igualmente, y por la misma razón previamente esgrimida, tampoco se menciona en este apartado la inclusión de los parámetros dinámicas en el vector acústico final.

La idea del empleo de ventanas de *liftering* se basa en que el ruido generalmente no afecta del mismo modo a todos los coeficientes cepstrales. De esta manera, dichas ventanas pueden realzar aquellos parámetros menos sensibles al ruido, a la vez que reduce la importancia del resto, lográndose así un mejor comportamiento ante entornos acústicos adversos. Así, por ejemplo, en los coeficientes *cepstrum*-LPC son los de orden menor los

que, normalmente, se encuentran más afectados por el ruido, de modo que se podría aplicar ventanas de *liftering* del tipo seno remontado [JRW87] o *general exponential lifter* [JW89] para minimizar el efecto de dichos coeficientes.

El fundamento del uso de las técnicas de proyección cepstral se encuentra en que uno de los efectos más importantes que introduce el ruido blanco es la reducción de la norma de los vectores cepstrales. Así, mediante la medida de la proyección cepstral se enfatiza los picos espectrales de energía, que son los menos afectados por el ruido, haciendo que la distancia basada en proyección cepstral sea una medida más robusta que las distancias euclídeas [CC91] [JH96]. Si bien este tipo de técnicas son eficaces con ruido blanco, desgraciadamente no lo son tanto con otros tipos de ruidos [HN94].

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, [DH73] [Fuk90] el más extendido, habiéndose empleado con éxito en sistemas de RAH mediante distintas aproximaciones y métodos [YSvVH00] [SBdlTR01].

Si el ruido es de banda estrecha, hecho este que en algunos entornos es factible asumir, se puede emplear con satisfactorios resultados la parametrización mediante procesado de sub-bandas [HTP96] [BDHM72], que comprende un conjunto de técnicas que están relacionadas con el dominio frecuencial de la señal de voz y que se sustentan 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 extracción de características, muchas han sido las contribuciones en este campo, la mayoría de las cuales con la intención de combinar la información no corrupta por el ruido [Hag00] [HB00]. Cabe destacar asimismo que esta aproximación de procesado en sub-bandas está muy relacionada con las técnicas de *missing data* [MBB01] [CGJ$^+$01], en las que se utilizan métodos de marginalización para reconstruir una función de densidad de probabilidad con el fin de generar posteriormente con ella las observaciones necesarias, eliminando de este modo la dependencia con las variables aleatorias supuestamente corruptas por el ruido [VGCJ99] [Jos02].

Además de las distintas opciones planteadas en esta Sección, que cubren a grandes rasgos las técnicas clásicas de extracción robusta de características, en la actualidad es la parametrización ETSI *advanced* [ETS02], que emplea conjuntamente sustracción espectral, *Spectral Substraction*, SS, y filtrado de Wiener, la más empleada en entornos ruidosos debido a los excelentes resultados obtenidos con ella, de un modo especial con la base de datos *Aurora* 2.

Cabe destacar que, aunque los algoritmos que forman parte de esta línea de actuación asumen que son inmunes al ruido, y que por tanto no necesitarían hipotéticamente ningún otro método para proporcionar robustez adicional al sistema final de RAH, en la actualidad es común encontrar combinaciones de este tipo de técnicas con otras de adaptación de modelos o de vectores de características para dar lugar a sistemas más competitivos y robustos.

3.3 Adaptación de Modelos Acústicos a la Señal.

Las técnicas de adaptación de modelos acústicos a la señal tratan de acercar éstos al espacio de la señal que se pretende reconocer, reduciendo así el desajuste que pudiera haber entre ambos espacios. Dichos métodos, que se usan profusamente en aquellas situaciones en las que no se dispone de suficientes datos como para emplear las técnicas clásicas de entrenamiento, son capaces de compensar, siempre en el dominio de los modelos acústicos, la variación estadística que el ruido introduce en la señal de voz. De todos modos, en dichos casos también existe la posibilidad de contaminar señal limpia con ruido del entorno acústico concreto, eliminando así el inconveniente de la falta de datos [BFS99] [MOS01] [Mor96]; sin embargo esta última opción, si bien más viable que la grabación del corpus correspondiente, tampoco proporciona unos modelos perfectamente adaptados ya que no quedarían reflejados en ellos ciertas alteraciones de la voz, como las producidas por el estrés, el propio ruido... (efecto Lombard [Lom11]). Por todo ello, suele resultar más conveniente en la mayoría de las situaciones recurrir a la adaptación para obtener unos modelos acústicos que representen estadísticamente la señal de voz bajo las condiciones concretas de reconocimiento. En cualquier caso, cabe destacar que, 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 en la fase de adaptación y la que posteriormente se reconocerá.

De entre todas las técnicas de adaptación de modelos acústicos a la señal de voz, las más empleadas actualmente son *Maximum A Posteriori*, MAP, *Maximum Likelihood Linear Regression*, MLLR, *Parallel Model Component*, PMC, *Jacobian Adaptation*, JA, *Vector Taylor Series*, VTS, para adaptación de modelos acústicos (hay una versión que transforma los vectores de características) y selección de modelos acústicos. A pesar de que en las siguientes subsecciones se considera 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 sus debilidades.

3.3.1 *Maximum A Posteriori*, MAP.

En general, tal y como se ha comentado en la Sección 2.4, a la hora de estimar los parámetros que definen los modelos acústicos (en el caso de emplear HMMs, las probabilidades de transición entre estados y los términos que definen las pdfs correspondientes asociadas a cada estado), se suele recurrir a la estimación de máxima verosimilitud, ML. Sin embargo, en la técnica *Maximum A Posteriori*, MAP, [GL94], se incluye un conocimiento previo de los parámetros, de modo que se supone que son variables aleatorias regidas por una conocida función de densidad de probabilidad. Cabe reseñar que, para el caso concreto de modelos acústicos basados en HMMs, y considerando como homogéneas las supuestas pdfs a priori, las expresiones que estiman los diferentes parámetros de los modelos acústicos 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 ya que de este modo se puede resstimar posteriormente de forma

sencilla los distintos parámetros de los modelos acústicos, debido a que, como se puede comprobar, con dicha suposición los vectores de características 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 con modelos acústicos estimados con el criterio MAP son algo mejores que los logrados tras emplear el criterio ML, especialmente cuando la cantidad de datos disponibles es aceptable. A su vez, y para solventar algunas de sus debilidades, se han propuesto a lo largo del tiempo ciertas aproximaciones. Así, existe una versión *on line* [HL97], 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 *Extended Maximum A Posteriori*, EMAP, [SL87] [ZSM95].

3.3.2 *Maximum Likelihood Linear Regression*, MLLR.

Suponiendo nuevamente que los modelos acústicos se componen mediante HMMs con GMMs como pdfs asociadas a cada estado, la versión más extendida del algoritmo *Maximum Likelihood Linear Regression*, MLLR, es aquella en la que únicamente se modifican los vectores de medias de las correspondientes Gaussianas mediante una función afín [LW95]; aunque bien es cierto que existe también la posibilidad de compensar igualmente las matrices de covarianzas [Gal97b], 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 [DLR77].

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 supone una importante ventaja al reducir 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 al representar las funciones afines espcios demasiado genéricos. Por tanto, el punto clave para lograr buenos resultados de RAH empleando la técnica MLLR, especialmente cuando se dispone de pocos datos, viene de la correcta elección de las clases de regresión, lo que no deja de plantear un cierto compromiso.

Cabe destacar que, las distintas transformaciones afines también se pueden obtener mediante el estimador MAP en lugar del ML, lo que da lugar a la técnica *Maximum A Posteriori Linear Regression*, MAPLR, [CSL99], que proporciona un comportamiento en términos de RAH algo superior al obtenido por el método MLLR.

3.3.3 *Parallel Model Component*, PMC.

Parallel Model Component, PMC, [GY93] [Gal95] [Gal97a] es una técnica que obtiene los nuevos modelos acústicos combinando 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 acústicos, lo que se realiza a partir de una cierta función de desajuste, *mismatch function*. En este sentido, y dado que normalmente se supone que el ruido es aditivo en el dominio temporal, la combinación de los parámetros de los correspondientes modelos acústicos 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 puede apreciar, en la presunción de una determinada función de desajuste, de modo que si ésta no se corresponde con la real puede darse un cierto desajuste que desemboque en el cálculo de unos modelos acústicos erróneos. Asimismo, el coste computacional de esta técnica 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 [Gal95].

3.3.4 *Jacobian Adaptation*, JA.

Jacobian Adaptation, JA, [YZH02] es un algoritmo que proporciona una eficiente solución a la hora de adaptar modelos acústicos a un no muy elevado coste computacional, partiendo del hecho de que las diferencias acústicas entre el espacio de reconocimiento y el de referencia son pequeñas.

Esta técnica requiere de cuatro pasos, a saber: entrenar los modelos acústicos de referencia, a partir de los cuales se construirán posteriormente los adaptados a las nuevas condiciones; calcular las matrices jacobianas en el dominio cepstral, para lo que habrá que suponer, del mismo modo que para el algoritmo PMC, un determinado modelo de degradación de los vectores de características; el tercer paso consiste en obtener una estimación de la diferencia del ruido entre el espacio de reconocimiento y el de referencia, para finalmente, y a partir de las matrices jacobianas y la estimación de la diferencia del ruido, estimar 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 cierto modelo de degradación de los vectores de características entre el espacio de reconocimiento y el de referencia y, ya por último, 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 reales no se dan las aproximaciones anteriores, la comunidad científica ha propuesto a lo largo del tiempo diversas mejoras para el método JA. Así, en el caso de que el desajuste entre los espacios de entrenamiento y reconocimiento

sea elevado, se puede hacer uso de un *clustering* de modelos acústicos iniciales [SKA+00], de modo que se elija de entre ellos el que mejor se adecue a la situación propia de cada momento. Asimismo, también se puede descomponer el ruido en tres términos para modelar el efecto aditivo, la distorsión convolucional y la dependencia de cada locutor [SSNS02]. Por otra parte, para incrementar el rango de acción de la adaptación Jacobiana, y reducir igualmente 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 [CRBJ00] [SH00]. Nótese que para cada una de las extensiones anteriores se cumplen más estrictamente las aproximaciones previamente comentadas, ampliando de esta manera el rango de acción de la técnica.

3.3.5 *Vector Taylor Series*, VTS, para Adaptación de Modelos Acústicos.

La técnica *Vector Taylor Series*, VTS, para adaptación de modelos acústicos [Mor96] [KUK98] [ADKZ00] presupone un cierto modelo de degradación de la señal acústica, normalmente constituido por una distorsión convolucional y un ruido aditivo [Ace90], que vienen definidos por una serie de parámetros que se aproximan en el dominio cepstral mediante una serie de Taylor, generalmente truncada hasta orden uno. De esta manera, una vez estimados los parámetros del modelo de degradación, y aplicando la correspondiente función inversa, se transforman convenientemente los modelos acústicos.

Esta técnica, al igual que todas las que presuponen la existencia de un modelo de degradación concreto, tiene el inconveniente de basar todas sus expectativas 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 comentado, JA, no deja de ser en cierto modo un caso especial y más sencillo de la técnica VTS cuando el modelo de degradación se representa mediante un polinomio de orden uno.

3.3.6 Selección de Modelos Acústicos.

En algunas ocasiones, la variabilidad del espacio de reconocimiento es demasiado elevada como para recurrir a los algoritmos clásicos de adaptación de modelos acústicos tratados anteriormente. En estos casos proporciona mejores resultados poseer un conjunto de modelos acústicos de modo que cada uno de ellos represente, de la mejor manera posible, a un cierto subentorno básico, siempre más homogéneo que el espacio completo. De esta manera, a la hora de reconocer se elegirá en cada momento aquel modelo acústico que mejor se adecue a las circunstancias.

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 mediante *Principle Component Analysis*, PCA, modelos adaptados a las nuevas condiciones a partir de una combinación lineal de los correspondientes a diversos subentornos básicos [KPNN00]. Dado que los modelos acústicos poseen generalmente una gran cantidad de parámetros, la elección de los que entrarán en el cálculo del análisis PCA resulta 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.

3.4 Adaptación de la Señal a los Modelos Acústicos.

La tercera línea de acción considerada a la hora de proporcionar robustez a un sistema de RAH propone adaptar 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 con tanta eficacia como las técnicas de adaptación de los modelos acústicos. Sin embargo, el menor coste computacional y la reducida cantidad de datos precisada para proporcionar interesantes resultados hacen que, actualmente, las técnicas de adaptación de vectores de características constituyan una de las soluciones más utilizadas a la hora de proporcionar robustez a un sistema de RAH. Los algoritmos incluidos dentro de esta línea pueden agruparse en tres grandes clases [SRM97], 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 opciones, haciendo especial hincapié en los algoritmos más representativos en cada caso.

3.4.1 Filtrado paso alto o *high-pass filtering*.

Dentro de los métodos de adaptación 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, pues suele ser 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 *Cepstral Mean Normalization*, CMN, también conocido como *Cepstral Mean Substraction*, CMS, el procesamiento *RelAtive SpecTral Amplitude*, RASTA, o técnicas clásicas de filtrado.

El algoritmo CMN [HS94] [YZH02] [dlTFH01] 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 [HM94] se basa en un filtrado paso alto, o paso banda, aplicado bien en el dominio log-espectral [HMBK91] [HMH93], bien en el cepstral [MMJ93]. Nótese que en las dos técnicas anteriores se pretende compensar principalmente los efectos de la distorsión convolucional, ya que ambas se sustentan 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. De este modo, dichos algoritmos proporcionarán mejores resultados conforme estas dos aproximaciones se acerquen a la realidad. Conviene recordar, llegados a este punto, que en el dominio Mel-cepstrum (parametrización típica en los sistemas de RAH actuales) no se da la primera de las suposiciones 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, existe 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 [AH95], obteniéndose así una ligera mejora sobre el método clásico. También se puede entender como una extensión de la técnica CMN, la normalización de términos estadísticos de orden mayor, como por ejemplo la varianza. De esta manera surge la técnica *Cepstral Variance Normalization*, CVN, [VL98] [Mol03] en la que se compensa la varianza de las distintas realizaciones, haciéndolas unidad.

Ya para finalizar, las técnicas clásicas de filtrado en el dominio temporal [MS97] pueden ser muy útiles en aquellas situaciones en las que el ruido predominante sea de banda limitada [MC95]. De este modo, y para estas condiciones concretas, se puede hacer uso de filtros paso banda en los que se adapten las frecuencias de corte atendiendo en cada momento a la naturaleza del ruido existente [Hos01]; otra posible solución consiste, por ejemplo, en emplear filtros de Wiener tras captar la señal mediante un *array* de micrófonos [MS97].

3.4.2 Técnicas basadas en modelos o *model-based.*

Las técnicas de adaptación de vectores de características basadas en modelos se sustentan en la presunción de que el desajuste entre los espacios de entrenamiento y de reconocimiento se puede representar mediante un modelo matemático de degradación. Una vez definido dicho modelo, los distintos métodos estiman los parámetros que lo representan convenientemente y, tras aplicar de un modo apropiado la correspondiente función inversa, 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. Habitualmente se suele considerar dos tipos de modelos de degradación, a saber: suponer que la señal ruidosa es, en el dominio temporal, la suma de la limpia y un ruido aditivo; o bien modelar la señal contaminada como la limpia afectada por un ruido aditivo y una distorsión convolucional [Ace90], aproximación esta algo más completa y realista.

Dentro de los algoritmos que consideran como modelo de degradación la combinación de ruido aditivo y distorsión convolucional, los más representativos son *Vector Taylor Series*, VTS, para normalización [KUK98] [Mor96], *Codeword Dependent Cepstral Normalization*, CDCN, [Ace90] y *Vector Polynomial Approximations*, VPS, [SRM97] [RGMS96]. Por otra parte, y dentro de aquellos métodos que modelan la señal contaminada como la limpia afectada únicamente por ruido aditivo, se encuentran *Minimum Mean Square Error*, MMSE, [YZH02] [MOSS02] [EM85], *Statistical Compensation*, StatComp, [dlTFH01], ecualización del ruido, *noise equalization*, [GJ99], o sustracción espectral, *Spectral Subtraction*, SS, [Bol79] [LB92] [NY94]. A continuación se describen brevemente

cada una de las técnicas anteriormente presentadas.

La técnica VTS para normalización utiliza los mismos fundamentos básicos que su variante para adaptación de modelos acústicos, considerando además que la señal limpia se puede modelar mediante una mezcla de Gaussianas, GMM. De esta manera, a cada una de las componentes de la GMM se le asocia una determinada transformación proviniente de 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 del entorno acústico considerados en este caso: 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 *Minimum Mean Square Error*, MMSE.

El método CDCN estima los parámetros del modelo de degradación mediante el algoritmo EM, a la vez que supone, como la técnica VTS para normalización, que la señal limpia se puede modelar mediante una mezcla de Gaussianas. Con todo ello se obtiene un vector de transformación asociado a cada componente para compensar la correspondiente degradación del entorno acústico. De cara a obtener la estimación final del vector de características limpio se hace uso de todos los vectores de transformación, ponderándolos convenientemente. Por otra parte, y como modificación de la técnica básica CDCN, surgió posteriormente el algoritmo *Interpolated Signal to Noise Rate Dependent Cepstral Normalization*, ISDCN, [Ace90], que es la versión interpolada del método de normalización empírico *Signal to Noise Rate Dependent Cepstral Normalization*, SDCN, [Ace90]. 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, de modo que los vectores de transformación dependerán en este caso tanto de dicha relación como del modelo de degradación presupuesto. Así, ante situaciones altamente ruidosas, SNR baja, el vector de transformación final tratará de 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 deberá de compensar en mayor medida, es la distorsión convolucional.

En el algoritmo VPS se supone nuevamente que la señal limpia puede modelarse mediante una mezcla de Gaussianas y, para cada componente de la misma, se obtiene un término de corrección estimado como la diferencia entre la media de la correspondiente Gaussiana de la GMM 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, lo que supone, de alguna manera, que el efecto del entorno acústico genera una Gaussiana en el espacio 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 tiende a hacer para el algoritmo VTS, aunque en este caso se suelen mejorar ligeramente las prestaciones de 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 modo que, por ejemplo, se puede utilizar conjuntamente con técnicas de *arrays* de micrófonos, así como con otros métodos de normalización, como el algoritmo CMN, e incluso de adaptación de modelos acústicos, caso de la técnica MLLR [YZH02].

En el método StatComp la señal limpia se estima a través de la señal ruidosa a partir de la generación de muestras mediante el método de Monte Carlo [LC98]. Para ello es necesario determinar previamente las pdfs del ruido, que se suelen suponer Gaussianas, así como la de la señal limpia, obtenida tras el correspondiente modelado realizado con un cierto corpus de entrenamiento. Los experimentos [dlTFH01] 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. Para ello 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, *Voice Activity Detection*, VAD.

Ya para finalizar con los algoritmos de adaptación d señal basados en modelos, 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 ruidosa, de modo que para ello en muchas ocasiones es necesario contar con un VAD suficientemente fiable. Esta solución, si bien normalmente produce una señal mucho más agradable de escuchar para el oído humano, también puede llegar a generar una importante distorsión (ruido musical) que hace que la utilización de este algoritmo en sistemas de RAH no siempre resulte tan satisfactoria. Asimismo hay que tener siempre presente las aproximaciones que se asumen en este método, ya que si no se dan en la realidad el sistema final no alcanzará las prestaciones deseadas. Dichas aproximaciones son: considerar que el entorno acústico sólo introduce distorsión propia de ruido aditivo y que la fase de la señal de voz no se ve afectada. Por todo ello la comunidad científica ha desarrollado algunas extensiones para compensar las limitaciones anteriores, dando lugar a técnicas como *Constrained Spectral Subtraction*, CSS, [KSS00], o la sustracción espectral usando harmónicos espectrales, *spectral subtraction using spectral harmonics*, [BK03].

3.4.3 Técnicas empíricas o *empirical.*

Las técnicas basadas en compensación empírica por comparación directa de los vectores de características requieren, en la mayoría de los casos, de una fase de entrenamiento con señal estéreo, aunque también existen aproximaciones que no la precisan, obteniéndose en dicho caso, eso sí, unas tasas de RAH algo menos satisfactorias. En general, este tipo de algoritmos se componen de dos fases: en la primera de ellas, que se puede denominar de entrenamiento, se estiman trama a trama todos aquellos parámetros necesarios para la adaptación de los vectores de características, que se lleva a cabo en la segunda fase, también conocida como de compensación, y en l que se hace 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, el éxito de los mismos se sustenta principalmente en cuan próximos se encuentran los datos ruidosos empleados en la fase de entrenamiento con respecto a los que posteriormente se adaptarán.

Dentro de los métodos de compensación empírica que más frecuentemente se han venido aplicando hasta la fecha destacan: *Signal to Noise Rate Dependent Cepstral Normalization*, SDCN, [Ace90], ecualización de histograma, *histogram equalization*, [dlTPS$^+$05] [Mol03], *Probabilistic Optimum Filtering*, POF, [NW94], *multivariate Gaussian-based cepstral normalization*, RATZ, y sus variantes [Mor96] y *Stereo based Piecewise LInear Compensation for Environments*, SPLICE, y sus extensiones [DDA01].

En la técnica SDCN se divide el rango de la SNR en varias bandas, de modo que para cada una de ellasse obtiene un vector de transformación a partir de señal estéreo en la correspondiente fase de entrenamiento. Posteriormente, y ya en la fase de compensación, se determina en qué banda se encuentran los vectores de características que se pretende adaptar 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 [Ace90] que no es preciso modificar todos los coeficientes de los vectores acústicos, sino sólo aquéllos 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 se basa en adaptar los vectores acústico mediante el uso de una función de transformación no lineal monótona creciente, suponiendo, además de la naturaleza ya comentada de la función de transformación, la independencia de las distintas componentes de los vectores de características. Cabe destacar que esta última aproximación no permitirá actuar, por ejemplo, sobre los efectos de rotación que el entorno acústico pueda producir en el vector de características. El criterio que se sigue para estimar la función de transformación es que la pdf de los vectores acústicos adaptados se asemeje a la de la señal limpia o a una pdf determinada a priori. 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 [Mol03].

El algoritmo POF consiste en filtrar cada vector de características a partir de los anteriores tratando de minimizar el error entre la señal adaptada y la limpia. Por otra parte, el espacio limpio se divide en varias regiones, de modo que para cada una de ellas se estima un filtro diferente. 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 distintas regiones.

En el algoritmo RATZ se presupone que los vectores de características limpios se pueden modelar mediante una pdf Gaussiana, o más genéricamente, mediante una mezcla de Gaussianas, GMM. De esta manera, en la fase de entrenamiento, en la que se puede o no hacer uso de señal estéreo, se estima un vector de transformación asociado a cada una de las Gaussianas. Posteriormente, y ya en la fase de compensación, se determina el vector de transformación final asociado a cada vector de características ruidoso a partir

de la suma ponderada de todos los asociados a las distintas Gaussianas. De igual modo que otras técnicas, también en este caso se desarrollaron ciertas modificaciones sobre el algoritmo básico, dando lugar a otras tantas extensiones. De este modo apareció el método *Signal to Noise Rate multivariate Gaussian-based cepstral normalization*, SNR RATZ, [Mor96] que trata el coeficiente de la energía 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 *Interpolated multivariate Gaussian-based cepstral normalization*, IRATZ, [Mor96], que representa el espacio degradado mediante varios entornos básicos y estima los correspondientes vectores de transformación para cada uno de ellos de manera independiente, incluyendo así 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, el vector de transformación final asociado a cada vector de características que se pretende adaptar se obtiene sumando ponderadamente todos los vectores de transformación estimados. Del mismo modo que en el caso anterior, y utilizando como base el método SPLICE, se han desarrollado a lo largo del tiempo diversas mejoras, como la extensión SPLICE *with model selection* [DDA01], que divide el espacio degradado en varios entornos básicos, dando lugar así a transformaciones más específicas y robustas (sería la extensión complementaria a la desarrollada con la técnica IRATZ), o el método *dynamic* SPLICE [DDA01], que emplea la correlación temporal de la señal que se pretende normalizar basándose en la suposición de que también debe haber una cierta relación entre los vectores de transformación finales empleados.

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 que tratan de 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 adaptación de vectores de características, como CMN, o con algoritmos de adaptación de modelos acústicos, caso de MLLR, [YZH02]. Asimismo también se ha propuesto con éxito reentrenar los modelos acústicos en el espacio normalizado definido tras aplicar la técnica SPLICE [DAPH00] [DDA02], o conjugar la estimación del ruido acústico con el propio algoritmo SPLICE [DDA03], de manera que se intenta compensar la limitación que poseen los métodos de normalización empíricos cuando la señal empleada en la fase de entrenamiento representa un espacio acústico distinto del de reconocimiento.

Capítulo 4

Marco de Experimentación.

4.1 Introducción.

A la hora de evaluar distintas técnicas, no ya sólo en este trabajo, sino en el ámbito de las tecnologías del habla en general, resulta conveniente elegir un marco de experimentación que se adecúe del modo más fiel posible a la problemática que se pretenda solucionar. De esta manera, por ejemplo, no reúne las mismas características un corpus diseñado para reconocimiento de locutor que uno pensado para RAH, ya que en el primero de los casos es conveniente grabar diferentes sesiones de cada locutor durante distintos intervalos 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 podrían ser engañosos. En este sentido hay que ser especialmente cuidadoso a la hora de elegir los corpora sobre los que evaluar técnicas de adaptación de vectores de características basadas en modelos (*model-based*), ya que éstos presuponen un determinado tipo de degradación que, en el fondo, no deja de ser una aproximación del real.

En este trabajo se pretende estudiar principalmente el comportamiento de distintas técnicas de adaptación de vectores de características ante entornos acústicos reales y altamente dinámicos, de modo que quede registrado el mayor número de alteraciones posibles en la voz, tanto independientes del locutor, caso del ruido aditivo y la distorsión convolucional, como 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 reales de conducción. Además, el hecho de que el locutor sea el copiloto permite, en mayor o menor medida, que el efecto Lombard se manifieste 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 distintos trabajos externos no resultan sencillas. Por ello, se realizaron también experimentos con la base de datos *Aurora* 2, que ha venido siendo muy utilizada en los últimos tiempos y posee además gran cantidad de entornos acústicos diferentes, lo que la convierte en un banco de pruebas muy válido de cara a estudiar el comportamiento

de distintas técnicas de robustez. Sin embargo, el corpus *Aurora* 2 tiene el gran inconveniente de que la señal ruidosa se genera añadiendo artificialmente ruido aditivo a las realizaciones limpias, lo que hace que no se manifiesten algunas de las importantes alteraciones en la señal de voz que ya se ha comentado anteriormente. Lo mismo sucede con el corpus *Hiwire*, que igualmente ha sido empleado en este trabajo como banco de pruebas.

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 han de realizar las pruebas de hipótesis estadística convenientes, que proporcionan, con un cierto intervalo de confianza, la certidumbre de si los distintos algoritmos poseen o no un comportamiento diferenciado.

En el presente Capítulo se analizan en la Sección 4.2 los tres corpora sobre los que se va a desarrollar toda la experimentación del trabajo: *SpeechDat Car* en español, *Aurora* 2 y *Hiwire*. En la Sección 4.3 se presentan las técnicas de hipótesis estadísticas más utilizadas en RAH, considerando en cada caso las ventajas y limitaciones que presentan, y determinando finalmente la que se empleará a lo largo del trabajo. Ya por último, en la Sección 4.4, se incluyen los resultados de referencia obtenidos con las bases de datos *SpeechDat Car* en español, *Aurora* 2 y *Hiwire*. Dichos resultados serán los que posteriormente servirán para determinar las mejoras proporcionadas por las distintas técnicas presentadas.

4.2 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, tal y como ya se ha indicado, 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 [vdHBC+99] [MLD+00] ya que, al ser grabado en condiciones reales, introduce todos los efectos que el entorno acústico del vehículo puede producir, tanto los independientes del locutor, caso del 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 [HP00] 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 más empleada y contrastada, por lo que se pueden realizar fácilmente comparaciones con otras técnicas y trabajos externos similares. Adicionalmente, también se llevó a cabo una experimentación con el corpus *Hiwire* [SEP+07], también generado a partir de la adición artificial de ruido a señales limpias. A continuación se presentan las características más relevantes de los tres corpora empleados en este trabajo.

4.2.1 Base de datos *SpeechDat Car* en español.

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 principalmente la

base de datos *SpeechDat Car* en español. 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 [Ace90], sino también otro tipo de alteraciones dependientes del locutor producidas por el estrés o el propio ruido (efecto Lombard). Así pues, y dado que el espacio acústico representado en el corpus *SpeechDat Car* en español es extremadamente variable, se suelen distinguir siete entornos básicos atendiendo a las condiciones de conducción, a saber

- E1: vehículo detenido con el motor en funcionamiento.
- E2: vehículo circulando por ciudad con las ventanillas cerradas y el climatizador apagado (condiciones silenciosas).
- E3: vehículo circulando por la ciudad en condiciones ruidosas (ventanillas abiertas y/o climatizador encendido).
- E4: vehículo circulando a baja velocidad por pavimento en mal estado en condiciones silenciosas.
- E5: vehículo circulando a baja velocidad por pavimento en mal estado en condiciones ruidosas.
- E6: vehículo circulando a alta velocidad por pavimento en buen estado en condiciones silenciosas.
- E7: vehículo 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, denominado *CLose Talk*, CLK, (Shure SM-10A) se coloca junto a la boca del locutor, 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: CLK o de referencia que, por la proximidad del sensor a la boca del conductor, se puede considerar que capta la señal de voz libre de ruido y, por tanto, proporciona el límite de RAH al que se puede aspirar; mientras que el segundo canal, elegido de 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, en la medida de lo posible, 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 tras un estudio previo ya que presentaba los mejores resultados de RAH para los tres canales de campo lejano de la base de datos [LMO+02]. Las señales para los dos canales, CLK y HF, se muestrean a 16 KHz y se codificaron con 16 bits.

	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

Tabla 4.1: Número de frases y palabras para los dos canales (CLK o HF) de los corpora de reconocimiento y entrenamiento ("# Frases reconocimiento", "# Frases entrenamiento", "# Palabras reconocimiento", "# Palabras entrenamiento", respectivamente) 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 (T1), mientras que el de entrenamiento comprende diferentes tareas de la base de datos, no sólo dígitos. Se incluyen igualmente los datos de la parte del corpus de entrenamiento correspondiente a la tarea de reconocimiento ("# Frases entrenamiento T1" y "# Palabras entrenamiento T1")

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 ligeramente reducidas por no disponer del todas las señales. Así, el corpus de entrenamiento estará compuesto 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 empleado se compone en total de 1.086 frases por canal (CLK y HF), grabadas todas por locutores distintos de los empleados para el corpus de entrenamiento y correspondientes con la tarea de dígitos aislados y conectados (T1). 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, se incluye también en la tabla el número de frases y palabras del corpus de entrenamiento correspondientes a la tarea de reconocimiento para cualquiera de los dos canales ("# Frases entrenamiento T1" y "# Palabras entrenamiento T1").

La relación señal a ruido, SNR, del 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 diversas. Así, para el más benigno acústicamente hablando, 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 el estudio de las características del ruido presente en el canal HF, se incluye en la Figura 4.1 la densidad espectral de potencia media, *average Power Spectral Densities*, PSD, del ruido para dicho canal y los distintos entornos básicos. Para ello se aplicó el método de Welch [Wel67] 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 de automóvil [MC95]. Por otra parte, en la misma figura queda patente de un modo indirecto la relación proporcional existente entre la velocidad del vehículo y la potencia media del ruido [MS97]. A pesar de que los entornos básicos no se han definido atendiendo a la velocidad, si se tienen en cuenta únicamente las bajas frecuencias, que en este caso es

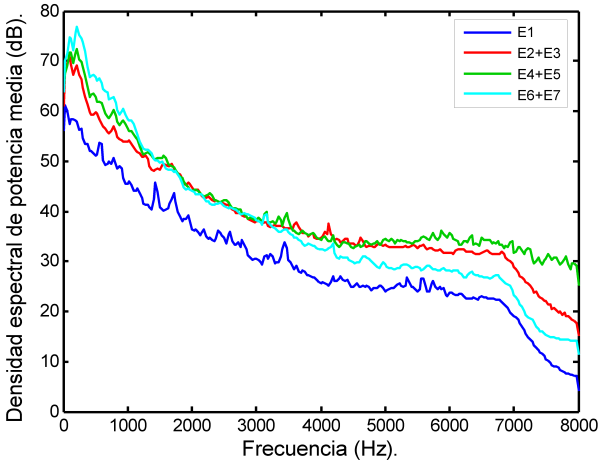


Figura 4.1: Densidad espectral de potencia media,*average Power Spectral Densities*, PSD, del ruido obtenida a partir del canal HF para los diferentes entornos básicos definidos para el corpus *SpeechDat Car* en español: E1, que se corresponde con la línea azul, E2 y E3, que se representan con la línea roja, E4 y E5, que se corresponde con línea verde y finalmente E6 y E7, cuya línea representativa es cian.

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 asociada a la unión de los entornos básicos E4 y E5 y posteriormente la correspondiente a la combinación de los entornos básicos E2 y E3, como podría haberse supuesto desde un principio atendiendo a la consabida relación entre la velocidad del vehículo y la potencia del ruido.

4.2.2 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, por ser ésta una base de datos referente y, por tanto, muy útil y valorada a la hora de comparar distintas técnicas y trabajos.

Aurora 2 se generó a partir de la base de datos de dígitos aislados y conectados en inglés *TIDigits* [LD93]. Dado que se pretendía disponer de un corpus que se correspondiera fielmente con las características frecuenciales típicas de terminales y equipamiento del área de las telecomunicaciones, las señales limpias del corpus *TIDigits* se submuestrearon 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 *International Telecommunication Union*, ITU, [ITU96]; dichos filtros "estándar" se denominarán en lo sucesivo G.712 y MIRS. Así pues, la base de datos *TIDigits* submuestreada y doblemente filtrada constituye finalmente el corpus limpio de *Aurora* 2, por lo que, utilizando la misma nomenclatura que en la subsección 4.2.1, se corresponde con el canal CLK.

Por su parte, el corpus ruidoso, o canal HF, se genera añadiendo artificialmente distintos tipos de ruido aditivo con diferentes SNRs a la señal limpia; dichos SNRs se calculan a partir de la señal no contaminada y el ruido 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 los entornos típicos en los que se suele hacer uso de terminales de telecomunicaciones, definiéndose finalmente ocho escenarios, a saber: metro o *subway*, muchedumbre o *babble*, vehículo o *car*, salón de exhibiciones o *exhibition hall*, restaurante o *restaurant*, calle o *street*, aeropuerto o *airport* y estación de tren o *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 o *multi-condition*, que 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 cuatro SNRs distintas: 20dB, 15dB, 10dB y 5dB. En ambos corpora de entrenamiento se emplea el 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, utilizando en todo momento el filtro "estándar" G.712. Para cada tipo de ruido y SNR se emplean 1.001 de las 4.004 frases distintas del corpus de reconocimiento seleccionado de la base de datos *TIDigits*. De este modo, el *set* A está constituido por 28.028 frases (1.001×7×4). Nótese que los tipos de ruido empleados en este caso son los mismos que los que forman parte del entrenamiento multi-condición, lo que quedará reflejado posteriormente en la experimentación.
- 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*, de modo que existirá un serio desajuste entre la señal de reconocimiento y cualquiera que sea el corpus de entrenamiento. Así pues, con este *set* se trata de observar el comportamiento de los sistemas de RAH cuando se reconocen señales contaminadas con ruido que no se ha visto 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: 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). Con el *set* C se pretende estudiar principalmente el comportamiento del sistema de RAH cuando la distorsión convolucional es distinta de la observada en el corpus de entrenamiento.

	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

Tabla 4.2: Número de frases para los dos corpora de entrenamiento ("# Frases entrenamiento limpio" y "# Frases entrenamiento multi-condición") y los tres *sets* de reconocimiento ("# Frases reconocimiento *set* A", "# Frases reconocimiento *set* B" y "# Frases reconocimiento *set* C") de la base de datos *Aurora* 2. En todos los casos la tarea se compone por dígitos continuos y aislados.

A modo de resumen se incluye en la Tabla 4.2 la composición, en número de frases, de los dos corpora de entrenamiento ("# Frases entrenamiento limpio" y "# Frases entrenamiento multi-condición") y los tres *sets* de reconocimiento ("# Frases reconocimiento *set* A", "# Frases reconocimiento *set* B" y "# Frases reconocimiento *set* C").

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 llegar a parecer, si únicamente se consideran las PSDs, que varios tipos de ruido son similares. Sin embargo esto no es así y, por ejemplo, los hay altamente estacionarios, como los correspondientes a *car* y *exhibition hall*, y los hay que se caracterizan precisamente por su falta de estacionaridad, como *street* o *airport*.

4.2.3 Base de datos *Hiwire*.

La tercera y última base de datos con la que se llevará a cabo algún tipo de experimentación es *Hiwire* [SEP+07]. Dicho corpus, que se diseñó como banco de pruebas de distintas técnicas de robustez, supone un entorno acústico compuesto por ruido aditivo. De este manera, y al igual que sucedía con la base de datos *Aurora* 2, las señales ruidosas se generan tras incluir artificialmente ruido aditivo, en este caso con tres diferentes niveles de SNR, dando lugar a otras tantas condiciones: bajo o *low*, medio o *medium* y alto o *high* nivel de ruido, que se corresponden aproximadamente con 10dB, 5dB y -5dB, respectivamente.

A diferencia de las bases de datos consideradas hasta el momento, la tarea del

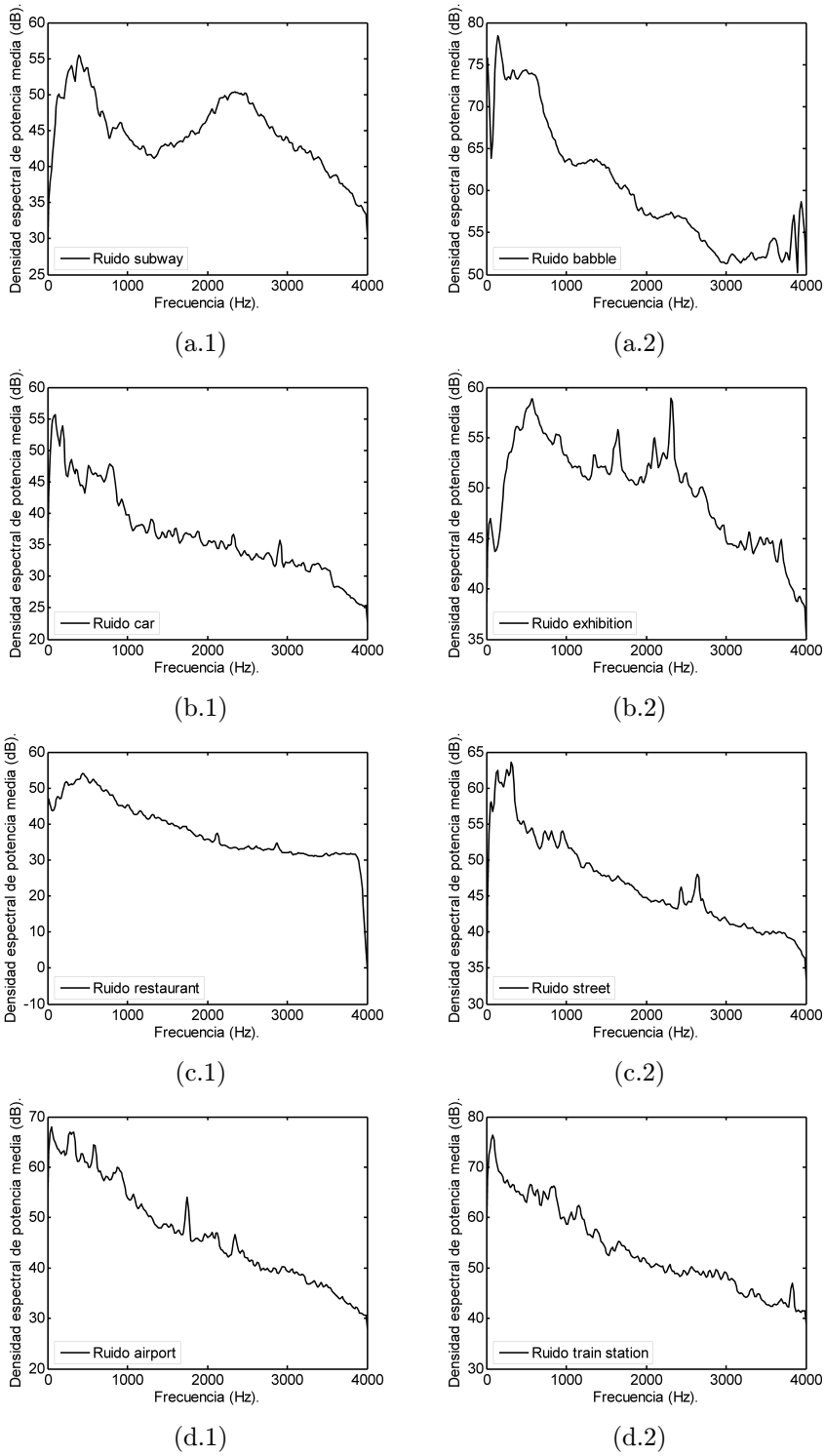


Figura 4.2: Densidad espectral de potencia media, *average Power Spectral Densities*, PSD, de los distintos ruidos presentes en el corpus *Aurora* 2: *subway* (a.1), *babble* (a.2), *car* (b.1), *exhibition hall* (b.2), *restaurant* (c.1), *street* (c.2), *airport* (d.1), *train station* (d.2).

corpus *Hiwire* no comprende dígitos, ya sean en español o inglés, sino que es algo más compleja, puesto que se trata de 100 frases de comandos y control aéreo pronunciadas en inglés por cada uno de los 81 locutores no nativos (31 franceses, 20 griegos, 20

italianos y 10 españoles) en un estudio (micrófono Plantronics USB-45 a 16 kHz). De esta manera, el vocabulario de la tarea consta de 133 palabras, y la perplejidad de la gramática es de 14.9. Por su parte, el ruido añadido se obtuvo a partir de grabaciones (micrófono AKG Q300) en la cabina de un Boeing 737 durante un trayecto ordinario, dando lugar, tal y como sería previsible, a un ruido altamente estacionario. Como ya se ha indicado, el hecho de introducir artificialmente la distorsión acústica supone un cierto déficit, dado que algunas alteraciones no quedan reflejadas en la señal de voz ruidosa final.

A partir del corpus definido se cosideraron dos posibles modos de trabajo. El primero de ellos, denominado *Robust Non-Native*, RNN, emplea todo el corpus para reconocimiento, tanto la parte formada por las señales limpias, como el compuesto por las ruidosas. Mediante este modo se pretende evaluar principalmente distintos métodos de extracción de características.

El segundo modo, o *Non-Native Adaptation*, NNA, divide el corpus en dos partes iguales de 50 alocuciones para cada locutor y nivel de ruido, definiendo de este modo un conjunto de frases de adaptación y otro de evaluación. Como se puede apreciar, la capacidad de maniobrabilidad en este segundo modo es sensiblemente mayor que en el anterior, por lo que se convierte en un banco de pruebas muy válido para evaluar diferentes métodos de robustez en general. De hecho, en este trabajo se dejará al margen el modo RNN, para centrarse en el NNA.

4.3 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 con presentar los resultados de la experimentación y cotejarlos directamente, sino que se ha de establecer de un modo estadístico, y bajo un cierto intervalo de confianza, hasta qué punto la diferencia de comportamiento entre las técnicas es significativa [GC89] e independiente de la base de datos seleccionada. A tal efecto en este trabajo se ha empleado la prueba de hipótesis estadística *z-test*.

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 [McN47] [GC89], *matched-pairs* [GC89] [PFF90] y *z-test* [GC89]. Por otra parte, si se deseara estudiar conjuntamente el comportamiento de más de dos algoritmos habría que recurrir a otro tipo de evaluaciones [Leh75].

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, considerando como información relevante únicamente el número de variables etiquetadas de distinta manera por ambas técnicas, mientras que desecha el resto. Se considera 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, pudiéndose afirmar, si dicha probabilidad es menor que una fijada, α, 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 vocablos próximos. Por todo ello, en muchas ocasiones se suele considerar la frase como variable discreta independiente, lo que genera otro tipo de problema, ya que cada frase etiquetada como errónea puede tener un número diferente de palabras erróneas, pudiendo dar lugar así a una comparación injusta entre los distintos algoritmos.

La prueba de hipótesis estadística *matched-pairs* compara el comportamiento de dos técnicas a partir del estudio de la diferencia del número de errores ocurridos entre los dos algoritmos. Dichos errores, que se obtienen en términos de unidades de distinta longitud e independientes entre sí, pueden ser de cualquier tipo, siempre que se recuenten de un modo consistente para ambas técnicas. En este caso se introduce una nueva variable aleatoria: la media de la diferencia del número de errores por unidad; de modo que la correspondiente hipótesis nula consiste en que la media de dicha variable sea nula. Con todo lo anterior, para rechazar la hipótesis nula se presupone que la variable aleatoria previamente definida y normalizada en varianza sigue una densidad de probabilidad normal de media nula y varianza unidad, $\mathcal{N}(0,1)$, de manera que se puede calcular mediante tablas la probabilidad de que dicha variable aleatoria normalizada tome un valor menor que el alcanzado a partir de los datos concretos de experimentación obtenidos tras evaluar ambas técnicas. Si dicha probabilidad es menor que una prefijada, α, se podrá aseverar que ambas técnicas presentan resultados diferenciados estadísticamente significativos con un intervalo de confianza de $1-\alpha$. Normalmente la elección de las unidades se realiza, en el ámbito del RAH, fraccionando las frases, de modo que se considera como límite bien una palabra correctamente reconocida por los dos sistemas que se pretenden comparar, 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. Asimismo la independencia de los errores cometidos en unidades próximas no deja de ser una premisa que, en muchas ocasiones, es difícilmente asumible.

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 p_1 y p_2 son independientes, la expresión $\sigma_d^2 = \sigma_1^2 + \sigma_2^2$, 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 utilizado en la experimentación 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 no deja de ser 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, $\mathcal{N}(0,1)$, siempre que n sea 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 la experimentación de ambas técnicas, denominado w. Si dicha probabilidad es menor que una prefijada, α, ($2p(W \geq |w|) < \alpha$, donde recuérdese que la función de densidad de probabilidad asociada a W responde a una normal de media nula y varianza unidad, $\mathcal{N}(0,1)$) se podrá afirmar que ambas técnicas presentan resultados diferenciados estadísticamente significativos con un intervalo de confianza de $1 - \alpha$.

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 p_1 y 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 p_1 y 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, obsérvese que las expresiones introducidas anteriormente para la evaluación de la prueba de hipótesis estadística *z-test* asumen, en el caso de que la covarianza entre p_1 y p_2 fuera negativa, unas condiciones aún más conservadoras, mientras que si, por el contrario, dicha covarianza fuera positiva, no se podría extraer conclusión fiable alguna a partir de los resultados calculados con las susodichas expresiones. De todo lo anterior se puede concluir pues que siempre que se use el método de prueba de hipótesis estadística *z-test* bajo la misma base de datos, como en

este trabajo, habrá 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 métodos asociados a RAH, parece introducir más dudas el uso de corpora diferentes, o el empleo de las pruebas de hipótesis estadísticas de McNemar o *matched-pairs*, que utilizar la técnica *z-test*, y todo ello a pesar de sus limitaciones.

4.4 Experimentación.

Tal y como se ha justificado con anterioridad, la experimentación realizada en este trabajo conjuga el uso de tres bases de datos: *SpeechDat Car* en español, *Aurora* 2 y *Hiwire*. Por otra parte, y tanto por necesidades de algunos de los algoritmos tratados como por proporcionar resultados con varias condiciones de experimentación, se han empleado diferentes unidades para el modelado acústico: fonemas y palabras, así como de parametrizaciones: estándar ETSI [ETS00] y ETSI *advanced* [ETS02]. En las siguientes subsecciones se presentan los resultados de referencia o *baselines*, para los tres corpora haciendo uso de las distintas condiciones de reconocimiento consideradas. Estos resultados permitirán posteriormente comparar el comportamiento de las distintas técnicas presentadas a lo largo del trabajo.

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

El método de parametrización empleado para este corpus es el estándar ETSI [ETS00], que proporciona, cada 10 ms y haciendo uso de una ventana de Hamming de 25 ms, un vector de características de 39 coeficientes: 12 estáticos MFCC más el logaritmo de la energía, junto con sus correspondientes primera y segunda derivadas.

En cuanto al modelado acústico, se han considerado dos opciones, representando en cada caso un tipo de unidad distinto, a saber, incontextuales o fonemas y palabras. El modelado acústico de unidades incontextuales se compone de 27 HMMs: uno asociado a cada fonema español más dos modelos de silencio, uno largo y otro corto para representar la pausa entre palabras. A su vez, cada HMM, salvo el asociado al silencio corto, consta de tres estados y la pdf correspondiente a cada uno de ellos se compone de una GMM de 16 componentes. Por su parte, el HMM correspondiente al silencio corto se construye a partir de un único estado cuya pdf asociada es una GMM de 16 componentes. De esta manera, y para el proceso de decodificación, cada palabra se modela mediante la concatenación de las correspondientes unidades fonéticas.

Por otra parte, el modelado acústico de palabras está compuesto por 12 HMMs, 10 de los cuales, los correspondientes con los dígitos que definen la tarea de reconocimiento, están constituidos por 16 estados, a cada uno de los cuales se les asocia como pdf correspondiente una GMM de 3 componentes. Además, el silencio largo se representa con un HMM de 3 estados y GMMs de 6 componentes para cada uno, mientras que 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 para algunos aspectos de la experimentación posteriores, que en el caso de los modelos acústicos

Entre.	Reco.	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

Tabla 4.3: Resultados de referencia en términos de WER (%), para los diferentes entornos básicos (E1,..., E7) de la base de datos *SpeechDat Car* en español utilizando la parametrización estándar ETSI y modelos acústicos para unidades fonéticas. Dichos modelos acústicos se pueden generar a partir de la señal limpia o la ruidosa (CLK o HF en la columna de "Entre.", respectivamente); HF† indica que se utilizan modelos acústicos específicos para cada entorno básico. La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que puede ser limpia (CLK) o ruidosa (HF).

de palabras, éstos se entrenan únicamente con la parte del corpus de entrenamiento perteneciente a la tarea de dígitos continuos y aislados, T1, (ver Tabla 4.1), mientras que los modelos acústicos de unidades incontextuales se obtienen haciendo uso de todo el corpus de entrenamiento. Para finalizar, 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 WER, para los distintos entornos básicos cuando se emplea la parametrización estándar ETSI y modelos acústicos de unidades fonéticas. MWER se corresponde con el WER medio, que se calcula a partir de las tasas de todos los entornos básicos de modo proporcional al número de palabras de los mismos (ver Tabla 4.1). Por otra parte, la columna marcada como "Entre." indica el canal de las señales empleadas para estimar los correspondientes modelos acústicos. Así, si se obtuvieron a partir de la señal limpia, la columna se marca con CLK; por el contrario, si la columna se nombra como HF indica que los modelos acústicos se entrenaron 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 mismo entorno básico. Nótese que estos últimos resultados proporcionan unas tasas ficticias puesto que en un caso real no se conoce a ciencia cierta el entorno básico al que pertenece la frase que se ha de decodificar. Ya para finalizar, la columna "Reco." indica que canal se emplea a la hora de reconocer: CLK, si es la señal limpia, o HF, si se hace lo propio con la señal ruidosa. En toda la experimentación presentada en este trabajo sobre la base de datos *SpeechDat Car* en español se aplica la técnica CMN tanto al corpus de entrenamiento como al de reconocimiento.

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 los distintos entornos básicos, generando un incremento significativo del WER en todos los casos ("Entre." CLK, "Reco." HF) con respecto a las tasas obtenidas cuando se emplea la señal limpia para modelar y reconocer ("Entre." CLK, "Reco." CLK). Por otra parte, utilizar modelos acústicos entrenados con toda la señal ruidosa, *matched condition*, ("Entre." HF, "Reco." HF) hace que el valor medio de WER, MWER, decaiga considerablemente con respecto al obtenido con modelos acústicos limpios ("Entre." CLK, "Reco." HF), aunque esta mejora no es consistente para todos los entornos básicos, como puede apreciarse en los menos ruidosos, como E1 y E2. Esto es debido a que los modelos acústicos ruidosos obtenidos

Entre.	Reco.	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

Tabla 4.4: Resultados de referencia en términos de WER (%), para los diferentes entornos básicos (E1,..., E7) de la base de datos *SpeechDat Car* en español utilizando la parametrización estándar ETSI y modelos acústicos para unidades de palabras. Dichos modelos acústicos se pueden generar a partir de la señal limpia o la ruidosa (CLK o HF en la columna de "Entre.", respectivamente); HF† indica que se utilizan modelos acústicos específicos para cada entorno básico. La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que puede ser limpia (CLK) o ruidosa (HF).

representan de un modo generalista a toda señal sucia y ésta representa a entornos básicos altamente heterogéneos, tal y como quedó patente en la Sección 4.2.1. Precisamente por todo ello el emplear modelos acústicos específicos para cada entorno básico ("Entre." HF†) proporciona, en todos los casos, los mejores resultados al decodificar 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 estándar ETSI y modelos acústicos de palabras. En este caso se puede apreciar igualmente el pernicioso efecto del ruido propio de los distintos entornos básicos ("Entre." CLK, "Reco." HF). Nótese, a partir de dicho resultado, como el modelado acústico de unidades fonéticas (16.21 % de MWER) es sensiblemente más robusto que el de palabras (21.49 % de MWER). Igualmente representativos son los resultados alcanzados con modelos acústicos entrenados con señal ruidosa, *matched condition*, ("Entre." HF, "Reco." HF), en los que el modelado de palabras proporciona importantes mejoras con respecto al modelado fonético (4.63 % de MWER comparado con 11.81 %, respectivamente). Ya para finalizar se puede apreciar como los mejores resultados en este caso se obtienen nuevamente cuando se emplean modelos acústicos ruidosos dependientes de cada entorno básico ("Entre." HF, "Reco." HF†), de modo que se alcanza 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). A partir de todo lo anterior se puede concluir que los modelos acústicos de palabra, mucho más específicos, proporcionan una importante mejora cuando las condiciones acústicas de las señales de reconocimiento y entrenamiento son similares, *matched conditions*, mientras que su comportamiento es sensiblemente inferior cuando no se dan estas circunstancias, *unmatched conditions*.

4.4.2 Experimentación con el corpus *Aurora* 2.

En la experimentación realizada con la base de datos *Aurora* 2 se utilizan las parametrizaciones estándar ETSI [ETS00] y ETSI *advanced* [ETS02] que, en ambos casos, proporcionan vectores de características de 39 componentes conformados por 13 parámetros estáticos (12 coeficientes MFCC más el logaritmo de la energía en el caso de la parametrización estándar ETSI, y 12 coeficientes MFCC modificados más el logaritmo de la energía para la parametrización ETSI *advanced*), junto con 26 dinámicos (la primera y segunda derivadas en ambos casos). 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

Tabla 4.5: Resultados de referencia en términos de exactitud por palabra, *word accuracy*, (%), para los diferentes *sets* (A, B y C) de la base de datos *Aurora* 2 utilizando la parametrización estándar ETSI y modelos acústicos de palabras generados a partir de los corpora de entrenamiento multi-condición o de señal limpia ("*multicondition training, multicondition testing*" y "*clean training, multicondition testing*", respectivamente).

En cuanto al modelado acústico seleccionado, éste está compuesto por modelos de palabras, de modo que cada uno de los 11 dígitos (en inglés el dígito 0 tiene dos posibles pronunciaciones) se representa con un HMM de 16 estados, asociándoles a cada uno de ellos una pdf compuesta por una GMM de 3 componentes. 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, que viene representado mediante un HMM de un estado con una GMM de 6 componentes. El modelo de lenguaje, al igual que para la experimentación realizada con el corpus *SpeechDat Car* en español, permite cualquier secuencia de dígitos.

En la Tabla 4.5 se presentan los correspondientes resultados de referencia en términos de exactitud por palabra, *word accuracy*, cuando se emplea la parametrización estándar ETSI. Como era de esperar, los resultados son sensiblemente más competitivos si se hace uso de los modelos acústicos multi-condición ("*multicondition training, multicondition testing*"), lo que es debido a que el desajuste entre los espacios de entrenamiento y reconocimiento es menor en este caso, especialmente para el *set* A. De la misma manera, también se aprecia claramente la relación existente entre la SNR y el comportamiento del sistema, de modo que ante SNRs reducidas, la exactitud por palabra decae hasta llegar a niveles dramáticos, y viceversa. Para finalizar, cabe destacar, tal y como queda patente en la Tabla 4.6, 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

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%

Tabla 4.6: Mejoras relativas en % obtenidas para los diferentes *sets* (A, B y C) de la base de datos *Aurora* 2 utilizando la parametrización estándar ETSI y modelos acústicos de palabras generados a partir de los corpora de entrenamiento multi-condición o de señal limpia, "*multi*" y "*clean*", respectivamente. El sistema de RAH referencia considerado de cara a calcular las mejoras relativas es HTK.

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	98,93	98,76	98,93	99,26	98,97	98,93	98,76	98,93	99,26	98,97	98,83	98,79	98,81	98,94
	20 dB	98,56	98,43	98,84	98,46	98,57	98,56	98,13	98,60	98,89	98,55	98,47	98,16	98,31	98,51
	15 dB	97,49	97,92	98,48	97,84	97,93	97,62	97,56	98,10	98,13	97,85	97,98	97,44	97,71	97,86
	10 dB	96,06	96,43	97,26	95,48	96,31	95,84	95,91	96,35	96,34	96,11	95,78	95,21	95,49	96,07
	5 dB	92,04	91,71	93,65	91,19	92,15	90,29	91,06	92,23	91,40	91,25	90,58	87,64	89,11	91,18
	0 dB	79,08	74,00	82,69	77,72	78,37	73,12	76,21	79,64	78,74	76,93	72,11	68,54	70,33	76,18
	-5dB	50,03	42,75	51,11	50,80	48,67	44,63	46,01	50,04	52,25	48,23	38,36	37,21	37,79	46,32
	Average	92,65	91,70	94,18	92,14	**92,67**	91,09	91,77	92,98	92,70	**92,14**	90,98	89,40	**90,19**	**91,96**

Aurora 2 Small Vocabulary	Clean 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	99,02	98,82	99,14	99,32	99,07	99,02	98,82	99,14	99,32	99,07	98,89	98,76	98,83	99,02
	20 dB	97,95	98,22	98,39	97,91	98,12	97,77	97,59	98,27	98,37	98,00	97,70	97,80	97,75	98,00
	15 dB	96,36	97,00	97,47	96,83	96,91	95,94	96,41	96,97	96,77	96,52	95,95	95,93	95,94	96,56
	10 dB	92,04	91,86	94,34	92,41	92,66	90,85	92,19	93,02	94,77	92,71	90,96	90,64	90,80	92,31
	5 dB	83,25	81,40	87,38	82,99	83,75	80,46	83,34	84,18	85,14	83,28	78,86	77,93	78,39	82,49
	0 dB	62,91	57,43	65,65	62,12	62,03	58,61	61,63	64,21	65,41	62,47	53,38	53,35	53,36	60,47
	-5dB	32,96	28,52	33,68	32,73	31,97	30,78	32,23	34,95	36,73	33,67	26,88	28,73	27,81	31,82
	Average	86,50	85,18	88,64	86,45	**86,69**	84,73	86,23	87,33	88,09	**86,60**	83,37	83,13	**83,25**	**85,97**

Tabla 4.7: Resultados de referencia en términos de exactitud por palabra, *word accuracy*, (%), para los diferentes *sets* (A, B y C) de la base de datos *Aurora* 2 utilizando la parametrización ETSI *advanced* y modelos acústicos de palabras generados a partir de los corpora de entrenamiento multi-condición y de señal limpia ("*multicondition training, multicondition testing*" y "*clean training, multicondition testing*", respectivamente).

referencia se obtienen con el sistema de RAH HTK [YEG+05] con las mismas condiciones de experimentación definidas previamente (extracción de características, modelado acústico y modelado de lenguaje).

En la Tabla 4.7 se presentan los resultados de referencia en términos de exactitud por palabra tras emplear la parametrización ETSI *advanced*. Se puede observar, si se comparan dichas tasas con las obtenidas al aplicar la parametrización estándar ETSI, una significativa mejora, especialmente cuando se hace uso de los modelos acústicos limpios, "*clean training, multicondition testing*". Esto es debido a las técnicas de robustez incluidas implícitamente en el algoritmo de extracción de características ETSI *advanced*: SS, filtrado de Wiener y ecualización ciega, principalmente, que reducen de una manera sensible el desajuste entre los espacios de entrenamiento y reconocimiento. Sin embargo, los resultados obtenidos cuando se utilizan los modelos acústicos entrenados con el

Aurora 2 Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	32,14%	37,91%	32,59%	34,54%
Clean	66,57%	73,27%	57,37%	67,41%
Average	49,35%	55,59%	44,98%	50,97%

Tabla 4.8: Mejoras relativas en % obtenidas para los diferentes *sets* (A, B y C) de la base de datos *Aurora* 2 utilizando la parametrización ETSI *advanced* y modelos acústicos de palabras generados a partir de los corpora de entrenamiento multi-condición y de señal limpia, "*multi*" y "*clean*", respectivamente. El sistema de RAH referencia considerado de cara a calcular las mejoras relativas es HTK.

corpus multi-condición, "*multicondition training, multicondition testing*", no representan una mejora tan importante puesto que el propio entrenamiento reduce, ya de por sí y en gran medida, el desajuste entre los modelos acústicos correspondientes y el corpus de reconocimiento. Para finalizar, se presenta la Tabla 4.8, en la que queda patente la importante mejora media que la parametrización ETSI *advanced* proporciona con respecto a los resultados del sistema de referencia HTK: 50.97 %.

4.4.3 Experimentación con el corpus *Hiwire*.

A pesar de que en este trabajo únicamente se va a trabajar en el modo NNA, en la siguiente subsección se incluirán los resultados de referencia o *baselines* de ambos modos (RNN y NNA). En los dos casos los vectores de características se obtuvieron a partir de los coeficientes MFCC (12 más C0), a los que se les añadió la primera y segunda derivada. Para proporcionar algo de robustez se aplicó el método CMN, lo que se realizó, al igual que la correspondiente parametrización, a partir de herramientas propias de HTK [YEG+05]. En cuanto al modelado del lenguaje, se utilizó una gramática de estados finitos.

Por otra parte, se entrenaron modelos acústicos independientes del locutor con herramientas de HTK y empleando todo el corpus de entrenamiento de la base de datos *TIMIT* [GLF+93], que comprende 5376 frases, obteniéndose 46 HMMs, representantes de otras tantas unidades fonéticas, cada uno de los cuales construido a partir de tres estados ocultos y 128 Gaussianas como pdf asociada a cada uno de ellos. De esta manera, y haciendo uso de dichos modelos acústicos, se obtuvieron los resultados de referencia para el modo RNN (Tabla 4.9), donde queda al descubierto el pernicioso efecto del ruido introducido. Obsérvese que se ha separado el origen de los locutores, así como las condiciones ruidosas, añadiendo a modo de comparación en este caso las tasas cuando se reconoce con la señal limpia.

Para el caso del modo NNA, los resultados de referencia se obtuvieron tras modificar los modelos acústicos empleados en el modo RNN con el método MLLR, tecnica esta que se considerada actualmente como un estándar a nivel de adaptación de locutor. De este modo, y a partir de las 50 alocuciones por locutor y condición acústica que componen el corpus de adaptación del modo NNA, se obtuvieron los correspondientes modelos acústicos específicos. Para ello se hizo uso de herramientas propias de HTK y se aplicó un árbol de 32 clases de regresión. Con todo lo anterior, en la Tabla 4.10 se presentan los resultados de referencia en términos de WER, en %, para las tres condiciones acústicas consideradas y distintas nacionalidades de los locutores. Comparando los resultados

RNN	Francés	Griego	Italiano	Español	MWER
Limpio	6.96	10.17	11.35	7.74	8.95
Bajo	63.45	51.02	51.90	43.08	54.92
Medio	84.58	76.65	70.75	68.67	77.17
Alto	97.88	99.00	96.52	98.29	97.88

Tabla 4.9: Resultados de referencia en términos de WER (%) obtenidos para el modo *Robust Non-Native*, RNN, de la base de datos *Hiwire*, para los distintos niveles de ruido (limpio, bajo, medio y alto). Se utiliza una parametrización próxima al ETSI estándar y modelos acústicos fonéticos generados a partir de la base de datos *TIMIT*.

RNN	Francés	Griego	Italiano	Español	MWER
Bajo	35.44	27.69	24.37	19.28	28.67
Medio	66.01	54.16	45.45	42.74	54.92
Alto	96.13	96.66	87.92	92.51	93.74

Tabla 4.10: Resultados de referencia en términos de WER (%) obtenidos para el modo *Non-Native Adaptation*, NNA, de la base de datos *Hiwire*, para los distintos niveles de ruido (bajo, medio y alto). Se utiliza una parametrización próxima al ETSI estándar y modelos acústicos fonéticos para cada locutor y condición acústica. Dichos modelos acústicos se obtuvieron a partir del algoritmo MLLR.

incluidos en la Tabla 4.9, se puede observar claramente como la técnica MLLR aporta una cierta mejora ante los niveles de ruido menos adversos, mientras que la ésta es apenas perceptible cuando la SNR es aproximadamente -5dB.

Capítulo 5

Adaptación MMSE: Visión Unificada.

5.1 Introducción.

Tal y como se ha indicado en el Capítulo 3, muchas son las técnicas empleadas a lo largo del tiempo para adaptar los vectores de características a los modelos acústicos proporcionando así robustez a los sistemas de RAH, y todas ellas, aunque cada una de un modo distinto, se sustentan en mayor o menor medida en un conocimiento o suposición de la degradación que el entorno acústico produce en los distintos coeficientes de los vectores de características.

Por otra parte, y también en el Capítulo 3, se ha realizado una taxonomía pormenorizada de los métodos más comunes de adaptación de vectores de características, clasificándolos 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 pdf a priori de la variable que se pretende obtener, lo que en ciertas situaciones no es sencillo. De esta manera, algoritmos tan dispares como CMN, CDCN, SDCN, VTS, VPS, RATZ o SPLICE, entre otros, provienen, en primera aproximación, de la misma base teórica.

En este Capítulo se realiza primeramente (Sección 5.2) 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 serán los que constituyan los vectores de características empleados en las distintas técnicas de adaptación propuestas en este trabajo. A continuación, y ya en la Sección 5.3, se presenta un desarrollo teórico unificado para las técnicas de compensación empírica basadas en el criterio MMSE más utilizadas en la actualidad por la comunidad científica: CMN, que aunque se ha considerado con anterioridad como 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 como punto de partida para la presentación de los distintos métodos empí-

ricos propuestos en este trabajo. La Sección 5.4 está dedicada al algoritmo de adaptación empírica *Multi-Environment Model-based LInear Normalization*, MEMLIN, que, haciendo igualmente uso del criterio MMSE, fue desarrollado como respuesta a ciertas limitaciones de las técnicas RATZ y SPLICE. En la Sección 5.5 se incluyen los resultados de RAH obtenidos tras aplicar los métodos de normalización empíricos CMN, RATZ, SPLICE y MEMLIN con la base de datos *SpeechDat Car* en español, quedando patente en dicha experimentación el buen comportamiento del algoritmo MEMLIN con respecto al resto de los métodos empíricos basados en el criterio MMSE considerados.

5.2 El Efecto del Ruido.

De cara a estudiar los efectos que el entorno acústico produce en la señal de voz, se suele proponer normalmente un modelo de degradación simplificado que simule el comportamiento del mismo. De este modo, el esquema más ampliamente utilizado hasta la fecha 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 [Ace90]. Así, todas las posibles alteraciones que el entorno acústico pueda introducir quedan representadas mediante la combinación de un ruido aditivo y una distorsión convolucional. Dado que, tal y como se ha comentado en el Capítulo 2, 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 decodificador, 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 asociado a los distintos vectores acústicos en el dominio MFCC: $\mathbf{x}_t$, que es el vector de características limpio, $\mathbf{n}_t$, 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

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

donde DFT es la transformada discreta de Fourier, *Discrete Fourier Transform*, e IDFT es la transformada discreta de Fourier inversa, *Inverse Discrete Fourier Transform*. Llegados a este punto, y obviando 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, t, $\mathbf{h}_t$ y $\mathbf{n}_t$ puedan tener distintas expresiones, dando lugar a lo que se denomina incertidumbre entre los vectores de características limpios y ruidosos, o, dicho de otro modo, que distintos vectores acústicos ruidosos pueden provenir de la misma trama limpia y viceversa. Obsérvese que este hecho supone, en última instancia, un gran hándicap para aquellas técnicas de normalización de vectores de características basadas en la aplicación de una función de transformación dependiente del vector acústico contaminado, puesto que para cada uno de ellos se asociará siempre la

misma trama como estimación del vector acústico limpio, lo que, como ya se ha indicado, no se corresponde con la realidad.

En la Figura 5.1 se pueden apreciar, de forma cualitativa, 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. Asimismo, el hecho de eliminar las tramas de silencio en este estudio se debe a que se pretende evitar el sesgo que una base de datos con gran cantidad de pausas podría introducir en los resultados. 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 *log-scattergram* correspondiente (Figura 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 pasa a ser mayor de 25 ms (Figura 5.1.b.2). En esta ocasión queda patente no sólo el incremento de la incertidumbre, 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 el primer coeficiente MFCC no son independientes de la longitud de la respuesta impulsional, siendo éstos más graves cuanto mayor es dicha longitud. Asimismo, se observa que la distorsión convolucional no produce únicamente, como en muchas ocasiones se asume, un desplazamiento constante para los distintos coeficientes de los vectores de características en el dominio MFCC, sino que también la varianza, aunque en menor grado, se ve afectada.

En la Figura 5.1.c, la señal contaminada correspondiente se genera añadiendo artificialmente ruido aditivo con una SNR de 0 dB. Se puede apreciar en este caso como el histograma de la señal limpia, Figura 5.1.c.1, se ha visto modificado drásticamente de una forma no lineal, reduciendo la varianza y compactando los dos modos que poseía, reduciéndolos prácticamente a uno. Por su parte, la incertidumbre (Figura 5.1.c.2) es mucho mayor a la observada en los dos casos anteriores, ampliando sensiblemente el rango de posibles valores del primer coeficiente MFCC de las tramas limpias asociado a un mismo valor para el mismo 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 asociado al coeficiente ruidoso 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,

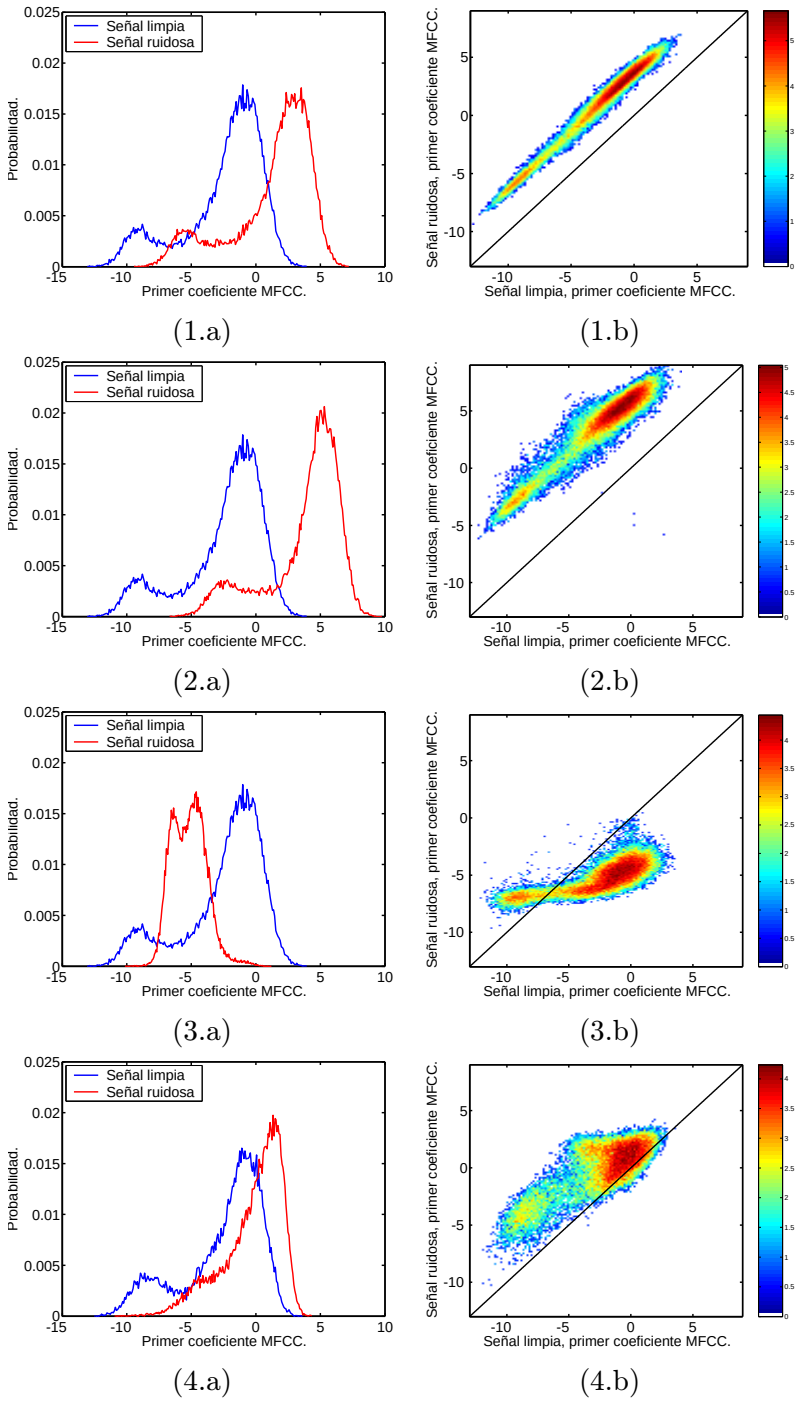


Figura 5.1: *Log-scattergrams* e histogramas del primer coeficiente MFCC de los vectores de características de voz limpia y ruidosa para distintos entornos acústicos. Las señales limpias se corresponden con el corpus de entrenamiento del entorno básico E4 de la base de datos *SpeechDat Car* en español. Por su parte, las señales contaminadas se obtienen tras 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 (b). En ambos casos la respuesta impulsional se obtuvo a partir de medidas en el habitáculo de un vehículo. El tercer entorno acústico asume únicamente ruido aditivo con SNR de 0 dB (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 los *log-scattergrams* representa la función identidad $x = y$.

aunque sin llegar al nivel del entorno acústico tratado anteriormente, donde la relación señal a ruido, 0 dB, es sensiblemente menor a la que se tiene en este caso: 8.05 dB 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 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 confirmar que en los entornos reales se dan conjuntamente los efectos ya comentados asociados tanto a la distorsión convolucional como al ruido aditivo.

5.3 Técnicas de Adaptación de Vectores de Características Empíricas Basadas en MMSE.

Independientemente de que una técnica de adaptación de vectores de características se pueda incluir en un grupo u otro dentro de la taxonomía que se ha presentado en el Capítulo 3, esto es: filtrado paso alto, basada en modelos o empírica, el estimador empleado para obtener el vector normalizado es, en muchas ocasiones, Bayesiano, de modo que es necesario suponer una determinada pdf 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 otras técnicas clásicas de estimación ya que, en cierto modo, se acotan los posibles valores de la variable objeto de estudio.

De entre los estimadores Bayesianos, es común la elección de aquél que minimiza el error cuadrático medio sobre todas las realizaciones, *Mean Square Error*, MSE. A dicho estimador Bayesiano se le denomina *Minimum Mean Square Error*, MMSE, 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 mediante técnicas basadas en el estimador MMSE, al igual que si se usara cualquier otro estimador Bayesiano, dependerán tanto de las realizaciones de que se disponga, como de la elección de la pdf a priori, hecho este último que a menudo no resulta sencillo.

Tal y como se ha adelantado, varias son las técnicas de adaptación de vectores de características que hacen uso del estimador Bayesiano MMSE: CDCN, CMN, SDCN, VTS, VPS, RATZ, SPLICE... Dado que parte del presente trabajo está centrado en el desarrollo de este tipo de técnicas, a continuación se presenta una visión teórica unificada de los algoritmos que, basados en dicho estimador, son los más empleados en la actualidad, esto es, RATZ y SPLICE. Asimismo se añade al estudio la técnica

CMN por ser prácticamente un estándar de facto en casi todos los sistemas de RAH. Posteriormente, en éste y sucesivos Capítulos, se derivarán las expresiones de las distintas técnicas que se han desarrollado a lo largo de esta tesis partiendo de la misma evolución teórica que se va a exponer a continuación.

Como se ha indicado anteriormente, el estimador óptimo que minimiza el MSE Bayesiano, MMSE, en este caso es la media de la pdf del vector de características limpio dado el ruidoso. Sea pues el vector acústico ruidoso para un instante de tiempo, t ($t \in [1, T]$), $\mathbf{y}_t$, y el correspondiente limpio para el mismo instante de tiempo, $\mathbf{x}_t$. De esta manera, el vector acústico limpio estimado, $\hat{\mathbf{x}}_t$, se obtendrá 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 el operador $E[\ ]$ representa la esperanza y $f(\mathbf{x}|\mathbf{y}_t)$ es la pdf de $\mathbf{x}$ dado $\mathbf{y}_t$. Llegados a este punto, el modo en que se aproxime tanto $\mathbf{x}$, denominado modelo del espacio de señal, como la pdf a priori $f(\mathbf{x}|\mathbf{y}_t)$, identificado en lo sucesivo como modelo de probabilidad condicionada entre espacios de señal, definirá los diversos algoritmos de normalización de vectores de características basados en el criterio MMSE.

La técnica CMN no asume ninguna expresión específica para el modelo de probabilidad condicionada entre espacios de señal, $f(\mathbf{x}|\mathbf{y}_t)$, mientras que el modelo del espacio de señal 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 $\mathbf{x}$, de modo que el método contempla únicamente un desplazamiento general entre los vectores de características. Así pues, y bajo estas premisas, la expresión (5.3) para el algoritmo 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, y se minimiza con respecto a $\mathbf{r}$

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

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

donde el operador $Tra[\ ]$ es la traza y $(\ )^T$ traspone el vector o matriz correspondientes. El desarrollo para obtener la expresión (5.6) a partir de (5.5) se encuentra en el Anexo 5.6, situado al final del presente Capítulo. Se puede apreciar que lo que propone esta técnica es suprimir del vector acústico degradado su propia media a la vez que se añade la de las tramas limpias, de modo que finalmente la media de los vectores de características normalizados coincida con la de los limpios. Otra posible realización de este mismo algoritmo consiste en sustraer a los vectores de características limpios del corpus de entrenamiento su propia media, de modo que los correspondientes modelos acústicos se entrenan con tramas ya normalizadas en media, haciéndose por tanto innecesario estimar $E[\mathbf{x}]$ a la hora de adaptar la señal ruidosa; en ese caso el vector de desplazamiento será

únicamente $\mathbf{r} = E[\mathbf{y}]$. Por su parte, y pensando ya en una aplicación práctica en tiempo real, la estimación de $E[\mathbf{y}]$ para el instante de tiempo t se suele realizar mediante métodos iterativos considerando los $t-1$ vectores de características ruidosos anteriores. En la actualidad, 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.

Obsérvese que la versión de la técnica 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. Dicha limitación se ve en muchas ocasiones unida a la de considerar que la distorsión convolucional es invariable en el tiempo y que, por tanto, se puede compensar con un vector de desplazamiento $\mathbf{r}$ fijo, lo que además, y de un modo indirecto, supone que el efecto de la distorsión convolucional 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 en presencia de ruido convolucional controlado (Sección 5.2). Estas claras limitaciones que posee la técnica CMN las trata de compensar el algoritmo RATZ incluyendo restricciones en el modelado de la probabilidad condicionada entre espacios de señal, $f(\mathbf{x}|\mathbf{y}_t)$, y modificando igualmente el modelado del espacio de señal.

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:

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

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

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 modelado del espacio de señal 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. Nótese que la transformación propuesta consiste nuevamente en un desplazamiento, al igual que para el algoritmo CMN, 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 el método RATZ se transforma en

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

donde, como se puede apreciar, se ha incluido la dependencia con la Gaussiana del modelo limpio, s_x, en el modelado de la probabilidad condicionada entre espacios de señal, lo que se manifiesta finalmente en el cálculo de $p(s_x|\mathbf{y}_t)$, que 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 estima a partir de un pseudo-modelo GMM que representa el espacio ruidoso y que se genera a partir de (5.7) y (5.8) asumiendo que el ruido produce

un efecto aditivo en el dominio MFCC [Mor96], lo que no deja de ser una aproximación que en muchas ocasiones se aleja de la realidad.

Por su parte, el cálculo del vector de desplazamiento, $\mathbf{r}_{s_x}$, se realizará en una fase de entrenamiento previa haciendo uso de señal estéreo (aunque también existe una versión "ciega" [Mor96] 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 con respecto a $\mathbf{r}_{s_x}$ el previamente definido error cuadrático medio asociado a la Gaussiana correspondiente, ξ_{s_x}

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

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

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 (5.12) se calcula 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 5.6 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 que sus vectores de transformación son más específicos, la utilización del algoritmo RATZ mejora sensiblemente las prestaciones de los sistemas de RAH con respecto a aquéllas obtenidas tras emplear la técnica CMN. Sin embargo, el modo en que se estima $p(s_x|\mathbf{y}_t)$ puede producir serios desajustes debido a que el modelo de degradación presupuesto para transformar cada Gaussiana del espacio limpio al ruidoso en el dominio MFCC no es del todo realista. Para eliminar dicho modelo de degradación, el algoritmo SPLICE asume dos aproximaciones: la primera de ellas consiste en modelar mediante una GMM el espacio ruidoso en lugar del limpio:

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

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

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 modelo del espacio de señal 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, aunque 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

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

donde se ha incluido en el modelado de probabilidad condicionada entre espacios de señal el término asociado a la Gaussiana del modelo del espacio ruidoso, s_y. Por su parte, $p(s_y|\mathbf{y}_t)$ 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 haciendo uso de (5.13) y (5.14) como

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

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 entrenamiento 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 con respecto a $\mathbf{r}_{s_y}$ el error cuadrático medio asociado a la Gaussiana correspondiente, ξ_{s_y}

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

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

El desarrollo teórico para obtener la expresión (5.18) a partir de (5.17) se puede consultar en el Anexo 5.6 del presente Capítulo. A partir de lo anterior se aprecia como en la técnica SPLICE 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 asociados al espacio ruidoso, redundando así en un mejor comportamiento en términos de RAH. 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, por mucho que 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, modelar el espacio ruidoso con pocas Gaussianas puede hacer que los vectores de desplazamiento de los métodos RATZ o SPLICE no sean lo suficientemente específicos. Para solucionar este inconveniente se suele dividir el espacio acústico ruidoso en varios entornos básicos, e, atendiendo a propiedades acústicas similares (SNR, componentes espectrales...), de modo que para cada uno de ellos se estiman los vectores de desplazamiento correspondientes de forma independiente. Adviértase que esta modificación se ha de trasladar convenientemente a los modelados del

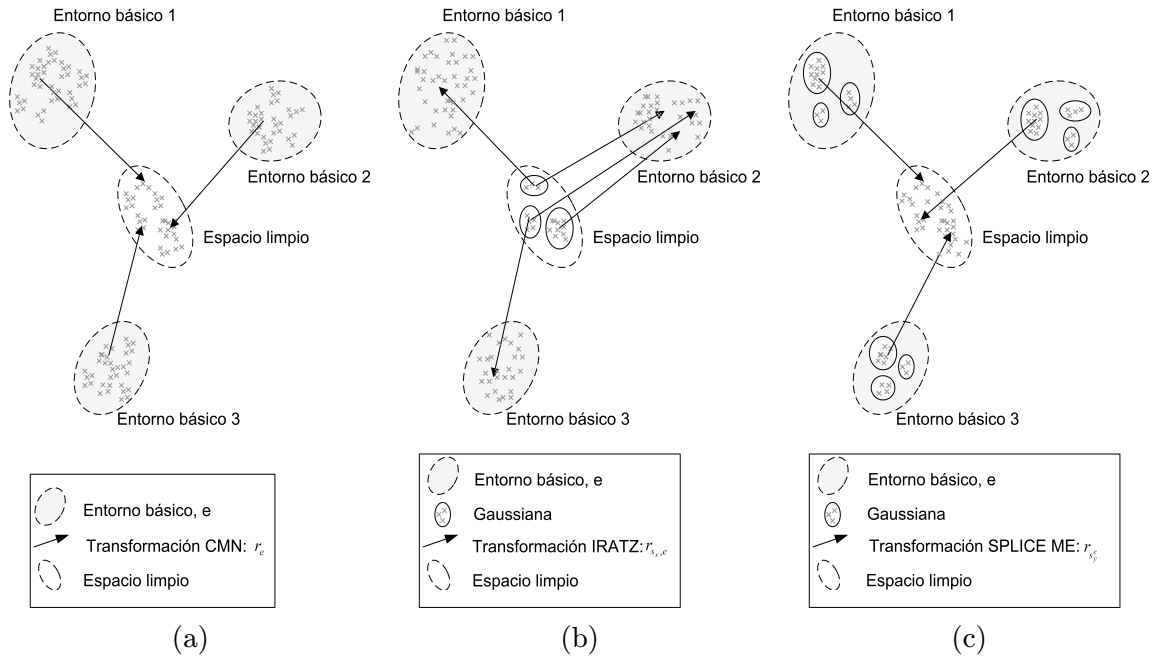


Figura 5.2: Representación gráfica de las versiones multi-entorno de las técnicas CMN (a), RATZ (b) y SPLICE (c), donde $\mathbf{r}_e$, $\mathbf{r}_{s_x,e}$ y $\mathbf{r}_{s_y^e}$ son los vectores de desplazamiento asociados a los respectivos algoritmos para cada entorno básico e.

espacio de señal y de la probabilidad condicionada entre espacios de señal, introduciendo así nuevas dependencias y dando lugar a unas soluciones más robustas: *Interpolate* RATZ, IRATZ, [Mor96] y SPLICE con selección del modelo de entorno, SPLICE *with environmental model selection* [DDA01], respectivamente. 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 cada entorno básico y Gaussiana dado el vector acústico ruidoso, $p(e, s_x|\mathbf{y}_t)$ para la técnica IRATZ o $p(e, s_y^e|\mathbf{y}_t)$, donde s_y^e es el índice de la Gaussiana del entorno básico e, 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 multi-entorno de las técnicas CMN, RATZ y SPLICE, 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).

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

Se ha podido comprobar en la Sección 5.3 que el método RATZ, a pesar de hacer uso de unas transformaciones más selectivas que el algoritmo CMN, posee una cierta

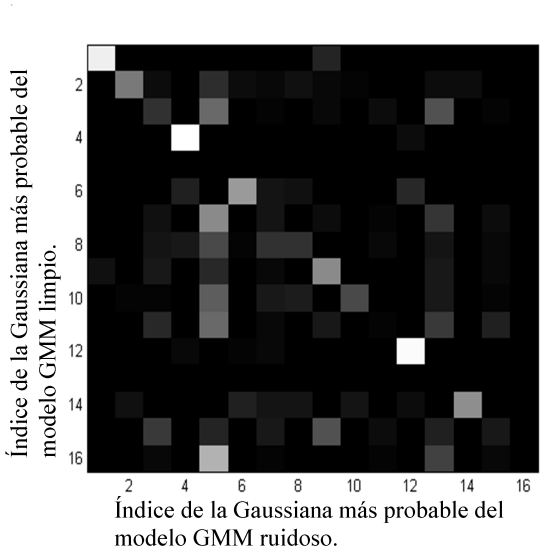


Figura 5.3: Histograma en dos dimensiones de los pares de Gaussianas más probables obtenidos a partir de la señal estéreo del corpus de entrenamiento del entorno básico E4 de la base de datos *SpeechDat Car* en español. El eje de las abscisas representa el índice de la Gaussiana del modelo ruidoso, mientras que el eje de las ordenadas incluye los índices de la componente del modelo limpio. Ambos modelos constan de 16 Gaussianas. Cuanto más clara sea la representación, mayor es el número de pares de vectores de características asociados a esa pareja concreta de Gaussianas.

debilidad a la hora de estimar la probabilidad a posteriori de la Gaussiana del modelo limpio dado el vector de características ruidoso. Este problema, tal y como ha quedado igualmente patente en la Sección 5.3, 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. Así, 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 de las componentes 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 relacionaran los vectores de características limpios producidos por una determinada componente con los correspondientes ruidosos. Este hecho, si bien se podía intuir de los *log-scattergrams* presentados en la Sección 5.3, queda totalmente reflejado 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. Se puede apreciar como, considerando una componente del modelo limpio como la más probable, pueden ser múltiples las componentes 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 respectivamente para entrenar los distintos vectores de desplazamientos asociados a cada Gaussiana cubran gran espacio, lo que proporcionaría un entrenamiento poco específico de dichos vectores de desplazamiento.

Para solventar el problema de la 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 *Multi-Environment Model-based LInear Normalization*, MEMLIN, [BLMO04a] 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 modelado del espacio de señal se aproxima mediante una función lineal del vector de características ruidoso; aunque en este caso dicha función depende del entorno básico y de las distintas componentes de las GMMs que modelan los espacios limpio y ruidosos: $\mathbf{x} \approx \Psi(\mathbf{y}_t, s_x, s_y^e) = \mathbf{y}_t - \mathbf{r}_{s_x,s_y^e}$, donde $\mathbf{r}_{s_x,s_y^e}$ es el vector de desplazamiento entre los vectores acústicos $\mathbf{y}_t$ y $\mathbf{x}$ asociado al par de Gaussianas s_x y s_y^e. Nótese que también en el método MEMLIN la transformación propuesta únicamente incluye un factor de desplazamiento.

Haciendo uso de las tres aproximaciones anteriores, la expresión (5.3) se transforma en este caso en

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

donde se puede apreciar como se ha incluido en el modelado de probabilidad condicionada entre espacios de señal las variables correspondientes al entorno básico, e, y a las Gaussianas de los modelos limpio y degradados, s_x y s_y^e, respectivamente. Por su parte, $p(e|\mathbf{y}_t)$ 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 adaptación de la señal degradada a partir de 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, $\mathbf{r}_{s_x,s_y^e}$, en una fase de entrenamiento previa no supervisada independiente para cada entorno básico haciendo uso de señal estéreo.

La probabilidad a posteriori del entorno básico, $p(e|\mathbf{y}_t)$, se calcula de modo recursivo 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. En el caso de que se pueda asumir que los entornos básicos no se suceden muy rápidamente a lo largo del tiempo, el valor de β debería 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 obtiene igualmente a partir de las expresiones (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 $\mathbf{r}_{s_x,s_y^e}$, así como la 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 y 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 como (5.24)

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

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

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). Si se desea consultar el desarrollo teórico para obtener la expresión (5.25) a partir de (5.24), éste se encuentra en el Anexo 5.6 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, en primera aproximación, 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 no supervisada es $p(s_x|e, s_y^e)$, que se puede obtener mediante frecuencia relativa, solución *hard*, o bien empleando (5.7) (5.8), (5.19) y (5.20), 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 s_x y s_y^e es el más probable para todas las parejas de vectores de características del corpus de entrenamiento del entorno básico e, mientras que $N_{s_y^e}$ es el número de veces que la Gaussiana más probable del modelo ruidoso es s_y^e para todos los vectores acústicos degradados del corpus de entrenamiento del entorno básico e.

Por otra parte, el 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 existe una gran cantidad de 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 ser posible disponer de suficientes datos, en cuyo caso la opción *soft* proporciona una solución más consistente. En este trabajo todos los experimentos realizados con la técnica MEMLIN se llevaron a cabo haciendo uso de la opción *hard*. Adviértase, llegados a este punto, como el algoritmo MEMLIN es, al igual que los métodos RATZ o SPLICE, totalmente no supervisado, esto es, en ningún momento, ni en la fase previa de entrenamiento, ni en la de adaptación, se hace necesario conocer o estimar la trascripción de la señal.

A modo de resumen se incluye en la Figura 5.4 una representación gráfica del algoritmo MEMLIN. En ella se puede apreciar claramente el radio de acción del correspondiente vector de desplazamiento para este método, r_{s_x,s_y^e}, sensiblemente más selectivo que los asociados a las técnicas CMN, RATZ y SPLICE (Figura 5.2). Nótese asimismo la diferencia entre los correspondientes espacios de proyección, que en el caso del algoritmo MEMLIN posee mucha menor incertidumbre, lo que supone una importante ventaja si se compara con el resto de técnicas tratadas. 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

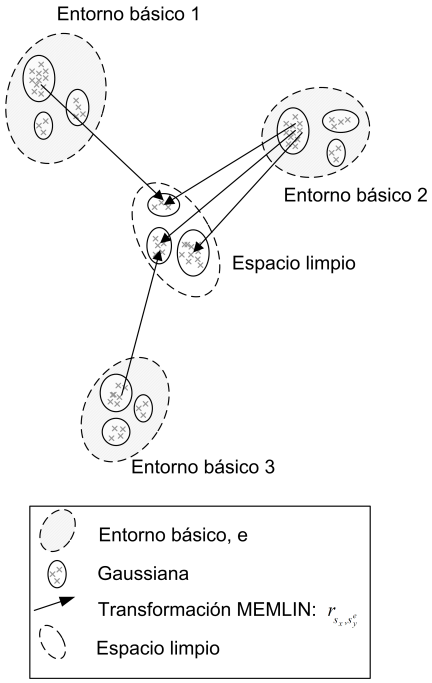


Figura 5.4: Representación gráfica de la técnica MEMLIN, donde $\mathbf{r}_{s_x,s_y^e}$ es el vector de desplazamiento asociado al par de Gaussianas s_x y s_y^e.

limpio, respectivamente, mientras que en el algoritmo MEMLIN el espacio de proyección se circunscribe a nivel de Gaussiana.

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

La experimentación comparativa de las distintas técnicas de adaptación de vectores de características empíricas tratadas en las Secciones 5.3 y 5.4 se realizó con la base de datos *SpeechDat Car* en español, que, como ya se indicó en la Sección 4.2, está dividida, además de por canales, en dos corpora: entrenamiento y reconocimiento, que se componen a su vez de diversos entornos básicos. Así pues, a la hora de realizar el proceso de entrenamiento no supervisado previo, 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 adaptación de los vectores acústicos degradados mediante las correspondientes técnicas, se aplicará el método CMN. Para esta experimentación se utilizó la parametrización estándar ETSI y se modelaron acústicamente 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 se estiman los diversos parámetros necesarios para las distintas técnicas de normalización, para lo que, tal y como ya se ha comentado, se hace uso del corpus de entrenamiento estéreo ("Entrenamiento"). 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 adaptada empleando los modelos acústicos que representan al espacio limpio.

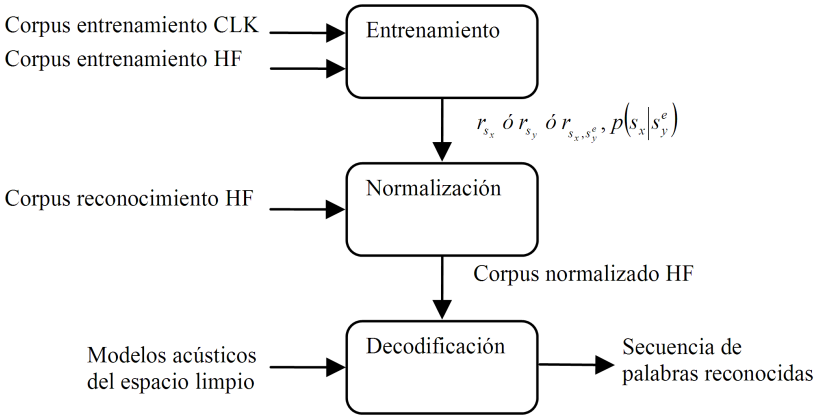


Figura 5.5: Esquema general de la experimentación realizada para las técnicas de adaptación de vectores de características empíricas: IRATZ, SPLICE con selección de modelo de entorno y MEMLIN. Se distinguen tres pasos, a saber: la fase previa de entrenamiento no supervisado, "Entrenamiento", para la que se supone en este caso el uso de señal estéreo. La segunda fase se corresponde con la estimación del vector acústico limpio, "Normalización". Finalmente, la última fase consiste en la decodificación de la señal normalizada haciendo uso de los modelos acústicos del espacio limpio, "Decodificación".

En la Tabla 5.1 se pueden apreciar los mejores resultados obtenidos para las distintas técnicas de adaptación de vectores de características. Junto a la señal decodificada, que aparece en la columna marcada como "Reco.": "HF IRATZ", "HF SPLICE ME", que hace referencia al método SPLICE con selección de modelo de entorno y "HF MEMLIN", se incluye el número de componentes que conforman las GMMs necesarias para cada algoritmo (se realizó el barrido con 8, 16, 32, 64 y 128 componentes cuyos resultados completos se pueden consultar en el Apéndice 5.7 de este mismo Capítulo). Cabe destacar que para la técnica MEMLIN, y mientras no se indique lo contrario, en lo sucesivo el número de Gaussianas empleadas para modelar el espacio limpio será el mismo que el utilizado para representar cada entorno básico, e. Los modelos acústicos se obtuvieron a partir de la señal limpia, "CLK" en la columna "Entre.". 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

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

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 CMN, 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

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

Tabla 5.1: Mejores resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando distintas técnicas de adaptación de vectores de características. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) 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.

limpias, que en principio es el límite al que se debe aspirar. 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 CMN; del mismo modo, al aplicar el algoritmo 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 %), aunque resulta algo inferior si se compara con la técnica MEMLIN (MWER de 6.05 %).

Por otra parte, y para determinar si se puede afirmar o no que los resultados anteriormente presentados 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 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. Si, por el contrario, se comparan los resultados proporcionados por las técnicas SPLICE ME y MEMLIN, se aprecia que $w = 1{,}53 < 1{,}96$, con lo que no se puede afirmar 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 convenientemente en la Sección 4.3.

En la Figura 5.6 se muestra la mejora media de WER, MIMP, para las distintas técnicas comparadas cuando se realiza un barrido del número de componentes de las GMMs empleadas en cada método (8, 16, 32, 64 y 128). El hecho de que a la hora de comparar los diversos algoritmos se haya representado el MIMP con respecto al número de Gaussianas por entorno básico se debe a que dicho parámetro da una idea aproximada del coste computacional del proceso de adaptación, muy dependiente de la cantidad de exponenciales evaluadas. Así, cabe destacar, de cara a completar las características de la comparación propuesta, que, aunque el número de vectores de desplazamiento asociados a cada Gaussiana del espacio ruidoso es mayor en el caso del algoritmo MEMLIN que

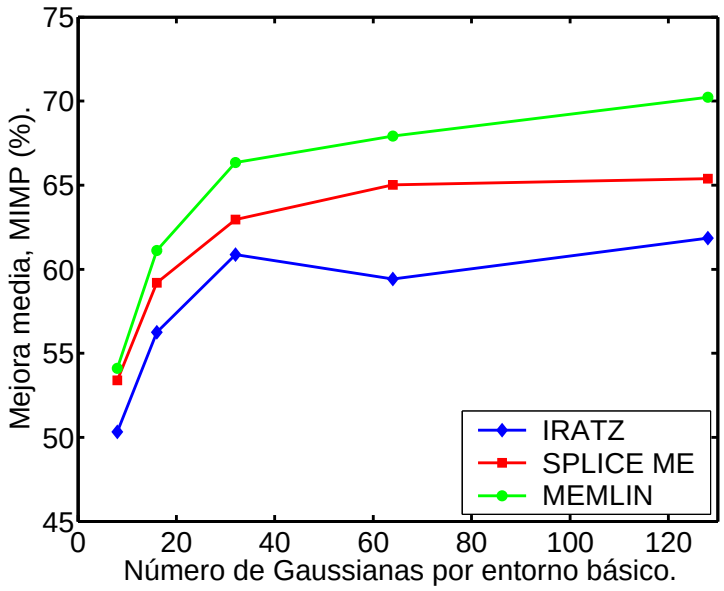


Figura 5.6: Mejora media del WER, MIMP, para las técnicas IRATZ, SPLICE con selección del modelo de entorno (SPLICE ME) y MEMLIN, atendiendo al número de componentes con que se modela cada entorno básico. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia.

en el del resto de técnicas, el coste computacional en la fase de normalización es casi idéntico, puesto que sólo se han de evaluar las Gaussianas asociadas al espacio ruidoso. Recuérdese que el método IRATZ, a pesar de modelar el espacio limpio, transforma cada una de las Gaussiana 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, lo que es 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 componentes 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 lugar por tanto a transformaciones más específicas, se manifiesta en una mejora en las tasas de RAH.

Ya para finalizar, las Figuras 5.7.a y 5.7.b presentan los histogramas comparativos y los *log-scattergrams* construidos a partir del primer coeficiente MFCC de los vectores acústicos de voz provinientes de las señales limpia y ruidosa (a) y limpia y normalizada (b) para el corpus de reconocimiento del entorno básico E4 de la base de datos *SpeechDat Car* en español. La señal normalizada se obtuvo mediante el algoritmo MEMLIN utilizando 128 Gaussianas por entorno básico. Comparada con la Figura 5.7.a.2, la incertidumbre tras el proceso de adaptación (Figura 5.7.b.2) se ha visto sensiblemente reducida, acercando además el *log-scattergram* correspondiente hacia la función identidad $x = y$, que define la normalización óptima. Por su parte, en la Figura 5.7.b.1 queda

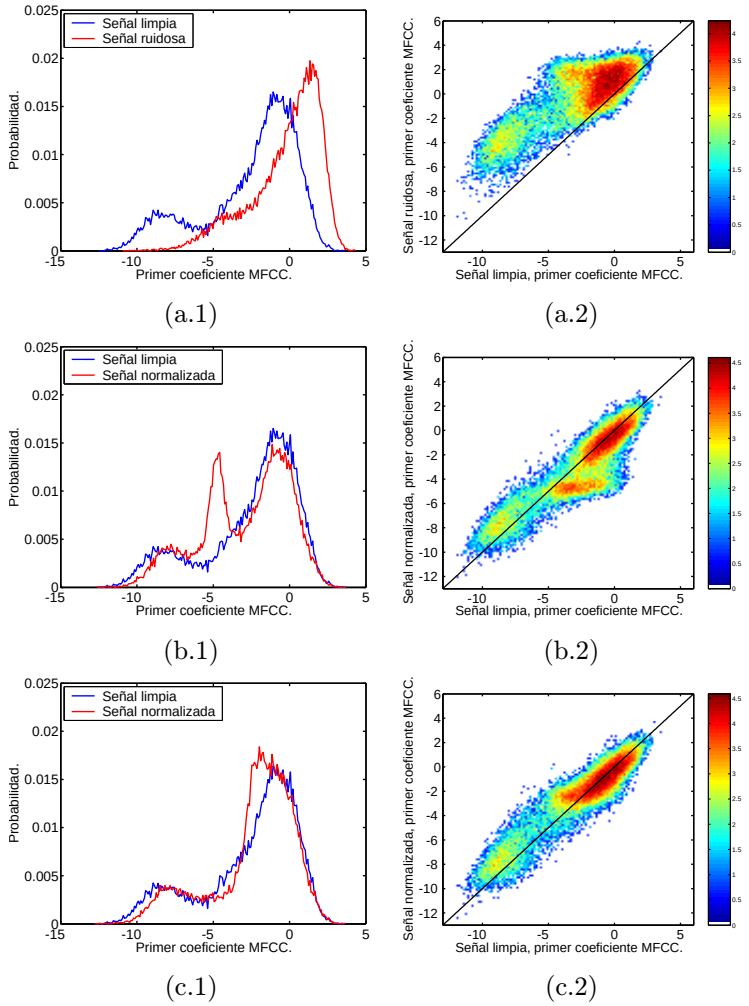


Figura 5.7: *Log-scattergrams* e histogramas del primer coeficiente MFCC de los vectores de características de voz limpia y ruidosa (a), o limpia y normalizada usando la técnica MEMLIN con 128 Gaussianas por entorno básico (b). En la figura (c) se representa el *log-scattergram* y el histograma obtenidos tras aplicar una variante del método MEMLIN, en la que la fase de entrenamiento 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-scattergrams* representa la función $x = y$.

patente como el histograma de la señal adaptada es similar al de la señal limpia salvo por un importante pico que aparece en torno a -5, 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 *Voice Activity Detector*, VAD, en el proceso de normalización, tanto en la fase previa de entrenamiento como en la posterior transformación del vector acústico ruidoso. Para asegurar esta afirmación se modificó la fase de entrenamiento del mismo mod en que se ha comentado previamente, de modo que en ella se emplearon sólo los vectores de características de voz, previamente identificados mediante un VAD evaluado sobre la señal limpia. La Figura 5.7.c representa el *log-scattergram* y el histograma construidos a partir del primer coeficiente MFCC de las tramas de voz de la señal limpia y normalizada del entorno básico E4 con la nueva extensión de la fase de entrenamiento para la técnica MEMLIN (128 Gaussianas por entorno básico). Se puede observar como el pico en el histograma desaparece. Cabe

destacar que, a pesar de las diferencias, tanto conceptuales como en cuanto a las tasas de RAH presentadas con anterioridad, los *log-scattergrams* 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.

A pesar del satisfactorio comportamiento en términos de MWER de la técnica MEMLIN, cuyos valores son menores que los obtenidos con los algoritmos CMN, IRATZ y SPLICE ME, se puede afirmar a modo de conclusión que, analizando dicho método en profundidad, se observan principalmente dos aproximaciones que pueden afectar en gran medida al comportamiento final de la técnica: por una parte la elección del modelo del espacio de señal ($\mathbf{x} \approx \mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e) = \mathbf{y}_t - \mathbf{r}_{s_x,s_y^e}$), 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 del espacio limpio dada otra del modelo del entorno básico ruidoso correspondiente sea en todo momento la misma, independientemente del vector acústico, lo que no se corresponde con la realidad. La primera limitación se estudiará de un modo directo en el Capítulo 6, mientras que para proporcionar un modelo de la probabilidad entre Gaussianas más realista se propondrá, en el Capítulo 7, una nueva solución basada en GMMs.

5.6 Anexo A.

Dado que la técnica MEMLIN es, de los cuatro algoritmos de adaptación 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 superíndice Tr para indicar que se trata del corpus de entrenamiento, tal y como sí estaba recogido en la Sección 5.4. 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

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

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

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

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

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

O, lo que es lo mismo

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

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

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

5.7 Anexo B.

En este Anexo se presentan los resultados en términos de WER (%) obtenidos para los diferentes entornos básicos (E1,..., E7) de la base de datos *SpeechDat Car* en español utilizando distintas técnicas de adaptación de vectores de características (IRATZ, SPLICE con selección de modelo de entorno, que se identifica como SPLICE ME, y MEMLIN). Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos entrenados con señal limpia. A su vez, y junto al nombre de las distintas técnicas, se ha incluido el número de Gaussianas con que se modelan los correspondientes entornos básicos (8, 16, 32, 64 y 128 Gaussianas, aunque para el caso del método MEMLIN también se incluye la opción con 4 Gaussianas).

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF IRATZ 8	4.31	8.74	6.71	9.27	12.11	10.16	16.67	8.93	50.33
CLK	HF IRATZ 16	4.22	8.49	5.73	8.27	10.96	8.41	14.29	8.07	56.26
CLK	HF IRATZ 32	4.22	7.89	6.15	7.14	9.34	7.94	12.59	7.41	60.87
CLK	HF IRATZ 64	3.93	9.09	6.85	8.02	9.15	7.46	10.54	7.62	59.42
CLK	HF IRATZ 128	3.74	8.83	6.15	7.77	9.06	7.30	8.50	7.27	61.84

Tabla 5.2: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características IRATZ. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica IRATZ. Junto al nombre de la técnica aparece el número de Gaussianas con que se modeló el espacio limpio. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF SPLICE ME 8	3.81	8.75	6.85	8.15	12.39	9.68	12.59	8.49	53.38
CLK	HF SPLICE ME 16	3.24	8.32	6.99	6.89	10.87	8.41	11.22	7.65	59.20
CLK	HF SPLICE ME 32	3.15	8.15	6.71	6.64	9.34	7.94	9.52	7.10	62.96
CLK	HF SPLICE ME 64	2.96	7.98	6.71	5.89	8.96	7.30	9.86	6.81	65.02
CLK	HF SPLICE ME 128	2.96	8.06	6.29	6.14	8.77	7.46	9.18	6.75	65.39

Tabla 5.3: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características SPLICE ME. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica SPLICE ME. Junto al nombre de la técnica aparece el número de Gaussianas con que se modelaron los distintos entornos básicos. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF MEMLIN 4	3.45	10.03	7.69	11.78	13.92	11.27	18.37	10.05	42.56
CLK	HF MEMLIN 8	3.16	8.49	6.43	9.27	11.91	9.05	14.97	8.39	54.10
CLK	HF MEMLIN 16	3.26	8.06	5.45	7.64	10.01	7.78	12.92	7.37	61.12
CLK	HF MEMLIN 32	2.49	7.80	5.03	6.64	9.25	6.51	11.22	6.61	66.35
CLK	HF MEMLIN 64	2.40	7.72	5.31	6.52	8.77	6.35	9.18	6.39	67.91
CLK	HF MEMLIN 128	2.30	7.46	4.62	6.39	8.77	5.40	8.16	6.05	70.22

Tabla 5.4: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características MEMLIN. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica MEMLIN. Junto al nombre de la técnica aparece el número de Gaussianas con que se modelaron los distintos entornos básicos, así como el espacio limpio. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Capítulo 6

Mejoras en el Modelado del Espacio de Señal.

6.1 Introducción.

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 espacio de señal ($\mathbf{x} \approx \mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e) = \mathbf{y}_t - \mathbf{r}_{s_x,s_y^e}$), que presupone para dicho método, una transformación lineal del vector de características ruidoso para cada par de Gaussianas, s_x y s_y^e, compuesta por un término de pendiente unidad y un vector de desplazamiento, $\mathbf{r}_{s_x,s_y^e}$. Este tipo de transformaciones compensa, a nivel de pares de Gaussianas, las modificaciones que el entorno acústico pudiera generar en la media de los vectores acústicos, pero no así las alteraciones producidas en 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 Gaussianas, que muchos vectores de características ruidosos correspondientes a segmentos de voz acaben siendo transformados hacia el silencio del espacio limpio, hecho este que ya quedó patente en ciertos histogramas presentados en la Sección 5.5.

Para compensar las limitaciones mostradas por el modelado del espacio de señal 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.

Considerando un patrón más realista de $\mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e)$ se 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 incluir un término de pendiente que pudiera ser distinto de la unidad, lo que da lugar a la técnica *Polynomial Multi-Environment Model-based LInear Normalization*, P-MEMLIN, [BLM+07]. Asimismo, también se puede hacer uso

de una transformación no lineal para cada par de Gaussianas, generándose así el método *Multi-Environment Model-based HIstogram Normalization*, MEMHIN, [BLMO04b].

Por otra parte, la segunda alternativa propuesta para mejorar el modelo del espacio de la señal consiste en emplear transformaciones dependientes de los sonidos, de tal manera que únicamente se definan vectores de desplazamiento 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 ello es necesario dividir ambos espacios, limpio y ruidoso, en fonemas, representando cada uno mediante una GMM. A esta nueva técnica se la denomina *Phoneme Dependent Multi-Environment Model-based LIneal Normalization*, PD-MEMLIN, [BLMO05c] [BLMO05a] 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 (Sección 6.2) el algoritmo P-MEMLIN haciendo uso de la misma base teórica empleada para explicar las distintas técnicas expuestas en el Capítulo 5. A continuación, en la Sección 6.3, se plantea el desarrollo teórico de la técnica MEMHIN, cerrando de esta manera la primera línea de actuación propuesta para modificar el modelo del espacio de señal. El método PD-MEMLIN se analiza convenientemente en la Sección 6.4, donde se podrá observar que, en el fondo, se trata de una generalización de la técnica MEMLIN. Hasta el momento todas las técnicas de adaptación de vectores de características presentadas precisan, en su fase de entrenamiento no supervisado, señal estéreo. Para evitar esta posible limitación, en la Sección 6.5 se presenta una fase de entrenamiento para el algoritmo PD-MEMLIN en la que no es necesario señal estéreo. Finalmente, 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 6.6. 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. Sin embargo, y a partir de la experimentación con la base de datos *SpeechDat Car* en español realizada, se puede concluir que los métodos P-MEMLIN y MEMHIN no proporcionan importantes mejoras con respecto al algoritmo MEMLIN, aunque sus comportamientos ante ruido aditivo sí son considerablemente más satisfactorios.

6.2 Técnica *Polynomial* MEMLIN, P-MEMLIN.

La utilización de un modelado del espacio de señal 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 este concepto en la técnica CMN se puede ver como aplicar conjuntamente dicho método con normalización cepstral de varianza, *Cepstral Variance Normalization*, CVN, [VL98] [Mol03]. Por su parte, la técnica SPLICE posee igualmente una extensión en la que se introduce un término de pendiente en el modelado del espacio de señal [DMGA05]. Las mejoras en ambos casos, 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 ralizado en la

Sección 5.2 sobre la distorsión que el ruido aditivo introduce en los vectores de características. A continuación se trata la correspondiente extensión del algoritmo MEMLIN en la que se considera como modelado del espacio de señal compuesto un polinomio de orden uno del vector acústico ruidoso, permitiendo 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, [BLM+07].

Tal y como se ha adelantado, lo que se pretende con el método P-MEMLIN 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 de $\mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e)$ 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 del espacio de la señal, respectivamente, siendo ambos función de los distintos pares de Gaussianas de los modelos limpio y ruidoso, s_x y s_y^e. 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 *Discrete Cosine Transform*, DCT, empleada en las parametrizaciones consideradas en este trabajo. Una vez presentada la expresión (6.1), cabe destacar que el modelo de la probabilidad condicionada entre los espacios de señal para la técnica MEMLIN sigue siendo igualmente válida en este caso, por lo que se va a hacer uso de las expresiones (5.7) (5.8), (5.19) y (5.20). A partir de todo lo anterior, la expresión (5.3) se transforma en este caso en

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

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, esto es, mediante 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 del espacio de señal: $\mathbf{A}_{s_x,s_y^e}$ y $\mathbf{b}_{s_x,s_y^e}$, se estiman mediante señal estéreo de forma independiente 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 $\mathbf{b}_{s_x,s_y^e}$ consistirá en que tanto la media como la desviación típica de las pdfs asociadas al par de Gaussianas s_x y s_y^e de la señal limpia y de la obtenida mediante el modelo del espacio de señal coincidan. Cabe destacar que

dicho criterio coincide con el MMSE cuando el modelo lineal del espacio de señal consta únicamente de un vector de desplazamiento, caso de la técnica MEMLIN, por ejemplo. Con todo ello, se puede observar que las correspondientes expresiones óptimas para las variables $\mathbf{A}_{s_x,s_y^e}$ y $\mathbf{b}_{s_x,s_y^e}$ son

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

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

donde el operador $\sqrt{\ }$ realiza la raíz cuadrada elemento a elemento a la matriz o vector sobre el que se aplique; 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. Mientras que $\mu^x_{s_x,s_y^e}$ y $\mu^y_{s_x,s_y^e}$ son los vectores de 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 del siguiente modo, donde z puede ser x o y

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

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

donde el operador $diag[\ ]$ hace nulos todos los elementos distintos de la diagonal de la correspondiente matriz sobre el que se aplique. 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 5.7 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 las matrices de covarianza fueran idénticas para todos los coeficientes y pares de Gaussianas $\left(\sqrt{\mathbf{\Sigma}^x_{s_x,s_y^e}} = \sqrt{\mathbf{\Sigma}^y_{s_x,s_y^e}} \forall s_x, s_y^e\right)$, las matrices $\mathbf{A}_{s_x,s_y^e}$ óptimas se corresponderían con la identidad y las expresiones para el vector $\mathbf{b}_{s_x,s_y^e}$ coincidirían con las de los vectores 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. Asimismo es importante recalcar que el método P-MEMLIN sigue siendo no supervisado, con la ventaja que esto supone.

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

A pesar de que la técnica P-MEMLIN, tal y como se ha podido apreciar, 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 algo más completo que sea capaz de compensar órdenes estadísticos mayores, igualando, 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 *Multi-Environment Model-based HIstogram Normalization*, MEMHIN, [BLMO04b] en la que se propone un nuevo modelo del espacio de señal que, manteniendo la dependencia con cada par de Gaussianas, está basado en ecualización de histograma, *histogram equalization*.

La ecualización de histograma, que inicialmente se propuso en realce de imagen [GW87], ya se ha aplicado con anterioridad en sistemas de RAH para proporcionar robustez, tal y como se introdujo en el Capítulo 3. Así pues, en su realización más básica se considera una función de transformación no lineal monótona creciente para adaptar 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 [Mol03] [dlTPS+05]. Por su parte, el método MEMHIN introduce una función de transformación basada en ecualización de histograma para cada par de Gaussianas, s_x y s_y^e, 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 la aleatoriedad del ruido propio del entorno acústico, no se ajusta con exactitud a la realidad. La segunda aproximación es considerar que los coeficientes de los vectores de características son independientes entre sí, 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 del espacio de señal, permaneciendo inalterado el modelo de probabilidad condicionada entre espacios de señal. De esta manera 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 correspondientes igualmente a s_x y s_y^e. De este modo, el nuevo modelo del espacio de señal 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 de los vectores acústicos limpios asociados 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 de los vectores de características degradados asociados 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 siguen siendo válidas (expresiones (5.7), (5.8), (5.19) y (5.20)). A partir de todo lo anterior, el estimador Bayesiano MMSE para el vector de características limpio para la técnica MEMHIN será

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

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 en un proceso de entrenamiento previo haciendo uso de señal estéreo para cada entorno básico, $(\mathbf{X}_e^{Tr}, \mathbf{Y}_e^{Tr}) = \{(\mathbf{x}_1^{Tr,e}, \mathbf{y}_1^{Tr,e}); ...; (\mathbf{x}_{t_e}^{Tr,e}, \mathbf{y}_{t_e}^{Tr,e}); ...; (\mathbf{x}_{T_e}^{Tr,e}, \mathbf{y}_{T_e}^{Tr,e})\}$, 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 de modo independiente. 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 hecho esto, $\mathbf{C}_{x,s_x,s_y^e}$ 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. Obsérvese, llegados a este punto, como el método MEMHIN sigue manteniendo el carácter no supervisado de las técnicas anteriores.

6.4 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, numerosas vectores de características de voz ruidosos se proyecten hacia el silencio del espacio limpio, tal y como se producía en las técnicas MEMLIN, P-MEMLIN o MEMHIN. Para solventar este hecho, obteniendo 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 *Phoneme-Dependent Multi-Environment Model-based LInear Normalization*, PD-MEMLIN [BLMO05c] [BLMO05a].

El uso de algoritmos de adaptación empíricos dependientes del fonema no es del todo novedoso, ya que anteriormente se han desarrollado métodos como *Phone-Dependent Cepstral Normalization*, PDCN, [LSA94] cuya filosofía es similar a la planteada en la técnica de normalización empírica denominada *Fixed Codeword-Dependent Cepstral Normalization*, FCDCN, [AS90] 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 [Bea92], necesita, a la hora de adaptar 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, no va a necesitar hacer uso de un sistema de RAH para proporcionar una cierta hipótesis del fonema correspondiente a cada vector acústico degradado, aunque sí hace uso nuevamente de 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 del 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 asociados a cada fonema se modelan del mismo modo mediante una GMM

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

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

  siendo s_x^{ph} la Gaussiana correspondiente al modelo del espacio limpio del fonema ph, y $\mu_{s_x^{ph}}$, $\mathbf{\Sigma}_{s_x^{ph}}$, y $p(s_x^{ph})$ 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 aproxima mediante una función lineal del ruidoso, siendo 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, se expresa a partir del modelo del espacio de señal: $x \approx \mathbf{\Psi}(\mathbf{y}_t, s_x^{ph}, s_y^{e,ph}) = \mathbf{y}_t - \mathbf{r}_{s_x^{ph},s_y^{e,ph}}$, 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, que definen tanto el modelo del espacio de señal como el modelo de la probabilidad condicionada entre espacios de señal, el vector acústico limpio se estima mediante el criterio MMSE como

$$\begin{aligned}\hat{\mathbf{x}}_t &= \int_{\mathbf{X}} \sum_{e} \sum_{ph} \sum_{s_y^{e,ph}} \sum_{s_x^{ph}} \mathbf{x} p(\mathbf{x}, s_x^{ph}, e, s_y^{e,ph}, ph|\mathbf{y}_t) p(s_x^{ph}, e, s_y^{e,ph}, ph|\mathbf{y}_t) d\mathbf{x} \\ &= \mathbf{y}_t - \sum_{e} \sum_{ph} \sum_{s_y^{e,ph}} \sum_{s_x^{ph}} \mathbf{r}_{s_x^{ph}, s_y^{e,ph}} p(e|\mathbf{y}_t) p(ph|\mathbf{y}_t, e) \\ &\times \; p(s_y^{e,ph}|\mathbf{y}_t, e, ph) p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph}), \end{aligned} \tag{6.13}$$

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 $s_y^{e,ph}$, 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 correspondiente modelo de probabilidad entre Gaussianas, esto es, la probabilidad 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 estima 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 a partir 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 métodos previamente presentados, de modo iterativo a partir de las expresiones (6.9) y (6.10)

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

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 estima mediante las expresiones (6.9) y (6.10)

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

Ya para acabar con los términos que se han de obtener en la fase de normalización, la probabilidad a posteriori de la Gaussiana $s_y^{e,ph}$, 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

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

Tal y como se ha comentado con anterioridad, 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 $t_{e,ph} \in [1, T_{e,ph}]$. De cara a crear dicho corpus es necesario asignar un fonema concreto a cada par de vectores de características de un determinado entorno básico, e. Para ello se realiza un proceso de segmentación forzada en términos de fonemas sobre la señal limpia del corpus de entrenamiento mediante el algoritmo de Viterbi. A partir de lo anterior, y a la hora de estimar el vector de desplazamiento $\mathbf{r}_{s_x^{ph}, s_y^{e,ph}}$ (6.18), se sigue un proceso similar al considerado para la técnica MEMLIN, obteniendo aquella expresión que minimiza el error cuadrático medio asociado al par de Gaussianas s_x^{ph} y $s_y^{e,ph}$ ($\xi_{s_x^{ph}, s_y^{e,ph}}$), que se define como 6.17

$$\begin{aligned}\xi_{s_x^{ph}, s_y^{e,ph}} &= \frac{1}{T_{e,ph}} \sum_{t_{e,ph}} p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}, e, ph) p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{Tr,e,ph}, e, ph) \\ &\times Tra\left[\left(\mathbf{x}_{t_{e,ph}}^{e,ph} - \mathbf{\Psi}(\mathbf{y}_t, s_x^{ph}, s_y^{e,ph})\right)\left(\mathbf{x}_{t_{e,ph}}^{Tr,e,ph} - \mathbf{\Psi}(\mathbf{y}_t, s_x^{ph}, s_y^{Tr,e,ph})\right)^T\right], \quad (6.17)\end{aligned}$$

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

donde $p(s_x^{ph}|\mathbf{x}_{t_{e,ph}}^{Tr,e,ph}, e, ph)$ es la probabilidad a posteriori de la Gaussiana 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 (6.19) se puede calcular a partir de las expresiones (6.11) y (6.12). 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 5.6 para la técnica MEMLIN, se puede consultar en el Anexo 6.7 de este mismo Capítulo.

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

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, $p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph}) \simeq p(s_x^{ph}|e, ph, s_y^{e,ph})$, y al igual que en métodos ya presentados, el modelo de probabilidad entre Gaussianas se puede estimar de dos maneras: mediante frecuencia relativa, solución *hard*, cuya expresión es

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

donde $C_N(s_x^{ph}|s_y^{e,ph})$ es el número de veces que el par de Gaussianas s_x^{ph} y $s_y^{e,ph}$ es el más probable para todas las parejas de vectores de características del corpus de entrenamiento del entorno básico e y el fonema ph; mientras que $N_{s_y^{e,ph}}$ es el número de veces que la

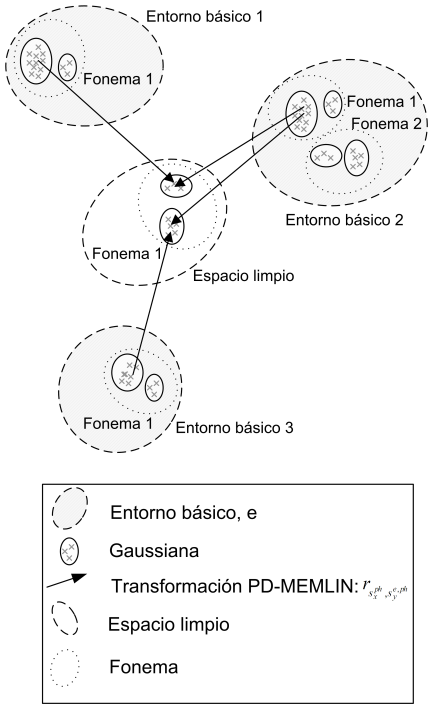


Figura 6.1: Representación gráfica de la técnica PD-MEMLIN, donde $\mathbf{r}_{s_x^{ph},s_y^{ph}}$ es el vector de desplazamiento asociado al par de Gaussianas s_x^{ph} y $s_y^{e,ph}$.

Gaussiana $s_y^{e,ph}$ es la más probable del modelo ruidoso para todos los vectores acústicos degradados del corpus de entrenamiento del entorno básico e y el fonema 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

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

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 representativo para todos los fonemas, especialmente si los vectores acústicos se han representado con un número elevado de componentes. 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*, que es más robusta ante este tipo de problemas. Por otra parte, y a modo de resumen, se incluye una representación gráfica del método PD-MEMLIN, Figura 6.1. Obsérvese que 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 es factible forzar que cada unidad fonética esté representada 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, aunque para ello es necesario, tal y como se ha ha podido observar anteriormente, un proceso de entrenamiento supervisado.

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

Hasta el momento, la fase de entrenamiento previa para los diversos algoritmos de adaptación de vectores de características empíricos presentados ha precisado de señal estéreo. Sin embargo, en muchas ocasiones, no es posible disponer de ella, por lo que es preciso desarrollar un procedimiento que, a partir de un corpus de entrenamiento compuesto únicamente por señal ruidosa, sea capaz de estimar las distintas variables necesarias. A este tipo de técnicas se las suele denominar "ciegas". Con esta misma mentalidad se definió, por ejemplo, la versión "ciega" del método RATZ [Mor96].

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 [BLMO05b]. La obtención de las correspondientes expresiones para la técnica MEMLIN es inmediata si se considera que los espacios limpio y ruidosos constan de un único fonema; aunque, como se indicará más adelante, es posible que, por la naturaleza del propio procedimiento desarrollado, no se obtuvieran tan buenos resultados en términos de RAH como los alcanzados con la versión "ciega" del método PD-MEMLIN.

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, $(\mathbf{Y}_{e,ph}^{Tr}) = \{(\mathbf{y}_1^{Tr,e,ph}; ...; \mathbf{y}_{t_{e,ph}}^{Tr,e,ph}; ...; \mathbf{y}_{T_{e,ph}}^{Tr,e,ph}\}$, con $t_{e,ph} \in [1, T_{e,ph}]$. En este caso el fonema asociado a cada vector acústico se obtiene mediante segmentación forzada de la señal ruidosa en términos de fonema a partir del algoritmo de Viterbi. Asimismo se dispone de las GMMs correspondientes con las que se modelan los distintos fonemas, tanto para los entornos básicos ruidosos, como 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 que define el modelado del espacio de señal, $\mathbf{r}_{s_x^{ph}, s_y^{e,ph}}$, y el modelo de probabilidad entre Gaussianas, que se sigue considerando independiente del vector de características degradado, $p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph}) \simeq p(s_x^{ph}|e, ph, s_y^{e,ph})$. 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 calcula, en primera aproximación, las dos variables anteriormente comentadas, dando lugar a $p_0(s_x^{ph}|e, ph, s_y^{e,ph})$ y $\mathbf{r}_{0,s_x^{ph}, s_y^{e,ph}}$. La expresión

para $p_0(s_x^{ph}|e,ph,s_y^{e,ph})$ se obtiene a partir de la distancia de Kullback-Leibler [KL51] modificada a tal efecto, de modo que proporcione 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 $s_y^{e,ph}$ minimizando el efecto que el ruido haya podido tener sobre ellas, en el cálculo de la distancia de Kullback-Leibler modificada no se tendrán en cuenta los vectores de medias de las Gaussianas, puesto que se supone que son ellos los más afectados por el ruido que pudiera darse en los distintos entornos básicos. Así, la distancia de Kullback-Leibler 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

$$\begin{aligned} d_{KL}(s_y^{e,ph},s_x^{ph}) &= p(s_y^{e,ph})log\left(\frac{p(s_y^{e,ph})}{p(s_x^{ph})}\right) \\ &+ \frac{p(s_y^{e,ph})}{2}\sum_i\left(log\left(\frac{\mathbf{\Sigma}_{s_x^{ph}}(i,i)}{\mathbf{\Sigma}_{s_y^{e,ph}}(i,i)}\right)+\frac{\mathbf{\Sigma}_{s_y^{e,ph}}(i,i)}{\mathbf{\Sigma}_{s_x^{ph}}(i,i)}-1\right), \end{aligned} \quad (6.22)$$

donde $\mathbf{\Sigma}_{s_x^{ph}}(i,i)$ y $\mathbf{\Sigma}_{s_y^{e,ph}}(i,i)$ son el término i-ésimo de las diagonales de las matrices de covarianzas de las Gaussianas s_x^{ph} y $s_y^{e,ph}$ respectivamente. En el Anexo 6.8 de este mismo Capítulo se puede consultar el desarrollo teórico necesario para, a partir de la definición de la distancia Kullback-Leibler, obtener la expresión correspondiente para dos Gaussianas.

Tal y como se puede apreciar, la distancia de Kullback-Leibler 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 denominada pseudo-verosimilitud entre dos Gaussianas, $pl_{KL}(s_y^{e,ph},s_x^{ph})$, que, teniendo en cuenta estos dos hechos, se calcula como

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

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

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

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)

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

Llegados a este punto, y con el proceso de inicialización ya concluido, se repitieron los experimentos expuestos en la Sección 5.5 para comprobar la bondad de dicho proceso en términos de RAH. Así pues, se empleó la base de datos *SpeechDat Car* en español, parametrización estándar ETSI y modelos acústicos fonéticos. Por su parte, los vectores de características ruidosos se normalizaron haciendo uso de la técnica PD-MEMLIN

utilizando únicamente $p_0(s_x^{ph}|e,ph,s_y^{e,ph})$ y $\mathbf{r}_{0,s_x^{ph},s_y^{e,ph}}$. 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 [DLR77], siendo (6.26) la correspondiente expresión para la iteración m-ésima, $\mathbf{r}_{m,s_x^{ph},s_y^{e,ph}}$, con $m>0$. Cabe destacar que en el Anexo 6.9 de este mismo Capítulo se ha incluido el desarrollo teórico completo necesario para obtener dicha expresión mediante la aplicación del algoritmo EM.

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

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

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

A la hora de mejorar la estimación de $p_0(s_x^{ph}|e,ph,s_y^{e,ph})$ en su correspondiente fase de ajuste, se utiliza señal de entrenamiento pseudo-estéreo. La parte limpia de dicha señal se obtiene tras adaptar los vectores de características del corpus de entrenamiento ruidoso mediante la técnica PD-MEMLIN utilizando en cada caso únicamente las transformaciones asociadas al correspondiente fonema "correcto", $\hat{ph}$. Dicho fonema "correcto" se obtiene mediante segmentación forzada de la señal ruidosa en términos de fonema a partir del algoritmo de Viterbi. A esta pseudo-técnica, por comodidad en la nomenclatura, se le denomina *Known* PD-MEMLIN, KPD-MEMLIN. Así, los vectores acústicos pseudo-limpios que completan la señal de entrenamiento pseudo-estéreo, $\hat{\mathbf{x}}_{t_e^{ph}}^{Tr,e,ph}$, se obtendrán como

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

De esta manera, y una vez definida la señal de entrenamiento pseudo-estéreo $(\hat{\mathbf{X}}^{Tr}_{e,ph}, \mathbf{Y}^{Tr}_{e,ph}) = \{(\hat{\mathbf{x}}^{Tr,e,ph}_1, \mathbf{y}^{Tr,e,ph}_1); ...; (\hat{\mathbf{x}}^{Tr,e,ph}_{t_{e,ph}}, \mathbf{y}^{Tr,e,ph}_{t_{e,ph}}); ...; (\hat{\mathbf{x}}^{Tr,e,ph}_{T_{e,ph}}, \mathbf{y}^{Tr,e,ph}_{T_{e,ph}})\}$, se puede obtener una nueva iteración para $p(s^{ph}_x|e, ph, s^{e,ph}_y)$ empleando las expresiones (6.20), o (6.21), según si se usa la decisión *hard* o *soft* respectivamente. Obsérvese que la obtención de señal pseudo-estéreo a partir del método KPD-MEMLIN se puede repetir iterativamente tantas veces como se considere preciso.

Para comprobar la funcionalidad del empleo de señal pseudo-estéreo, se repitió el mismo experimento de RAH anteriormente comentado. En este caso se hace uso de una única iteración del algoritmo EM para estimar el vector de desplazamiento, obteniendo $\mathbf{r}_{1,s^{ph}_x,s^{e,ph}_y}$, mientras que el modelo de probabilidad entre Gaussianas se estima mediante señal pseudo-estéreo tras la aplicación de la técnica KPD-MEMLIN con $\mathbf{r}_{0,s^{ph}_x,s^{e,ph}_y}$ y $p_0(s^{ph}_x|s^{e,ph}_y, e, ph)$. Se puede apreciar que, con todo ello, se alcanza una mejora media en términos de WER, MIMP, de 50.23 %, lo que certifica, de un modo cuantitativo, el potencial de la fase de entrenamiento "ciega" basada en señal pseudo-estéreo. A raíz de este resultado se decidió el empleo de dicha fase de entrenamiento para ajustar también la estimación del vector de desplazamiento haciendo uso de la expresión (6.18). De este modo, si el modelo de probabilidad entre Gaussianas se estima mediante señal pseudo-estéreo tras aplicar iterativamente tres veces la técnica KPD-MEMLIN y, por otra parte, la primera iteración del vector de desplazamiento obtenida mediante el algoritmo EM, $\mathbf{r}_{1,s^{ph}_x,s^{e,ph}_y}$, se ajusta con dos nuevas iteraciones adicionales con señal pseudo-estéreo, la MIMP para el experimento considerado anteriormente alcanza el 58.68 %, valor este que, como se podrá observar más tarde, se encuentra muy cercano del que se obtendría si se aplicara la correspondiente fase de entrenamiento con señal estéreo. Estos resultados muestran que la combinación del algoritmo EM con la fase de entrenamiento "ciega" basada en la técnica KPD-MEMLIN pueden proporcionar satisfactorias estimaciones de los vectores de desplazamiento y de los modelos de probabilidad entre Gaussianas. Nótese que el método de entrenamiento propuesto sigue siendo supervisado. Por su parte, y tal y como ya se ha comentado previamente, las expresiones de las distintas variables 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. Sin embargo, en ese caso no podría aplicarse el método KPD-MEMLIN, que es el causante en gran medida del satisfactorio comportamiento de la fase final de entrenamiento propuesta, por lo que es probable que no se alcanzara un comportamiento tan satisfactorio. En consecuencia, el desarrollo de una técnica de entrenamiento "ciega" no supervisada para el algoritmo MEMLIN se ha convertido en una de las principales líneas futuras de investigación.

Por último, y 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, donde "Inicialización p" e "Inicialización r" se corresponden con las expresiones (6.24) y (6.25), respectivamente. El sistema nombrado como "EM" se emplea para obtener la primera iteración de ajuste para el vector de desplazamiento, $\mathbf{r}_{1,s^{ph}_x,s^{e,ph}_y}$. Por su parte, el bloque denominado "KPD-MEMLIN" hace referencia a la adaptación de los vectores de características ruidosos utilizando el conocimiento a priori del correspondiente fonema "correcto", dando lugar así a la señal pseudo-estéreo (6.28). Por último en el bloque identificado como "Entrenamiento estéreo" se estiman las nuevas variables (vectores de

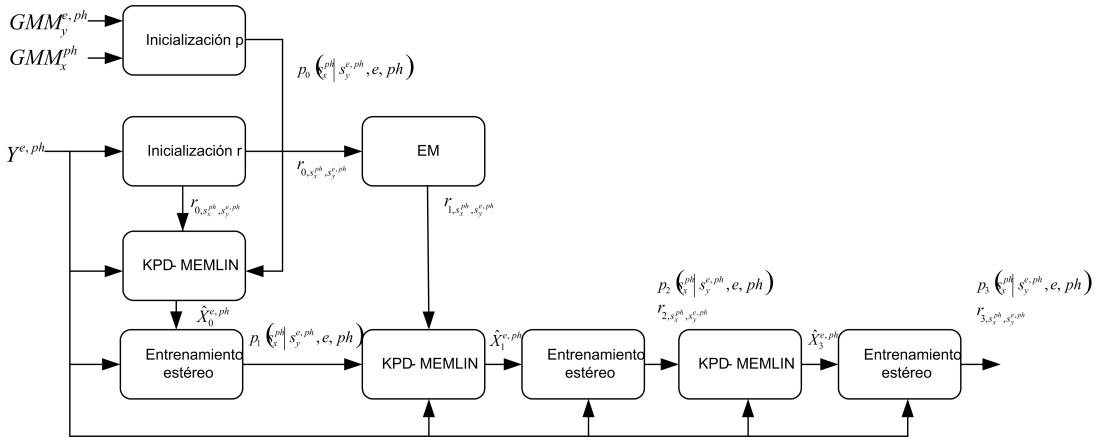


Figura 6.2: Representación gráfica del proceso de entrenamiento "ciego" de la técnica PD-MEMLIN que se va a emplear en este trabajo. A partir del bloque "Inicialización r" se obtiene la primera estimación del modelo de probabilidad entre Gaussianas, $p_0(s_x^{ph}|s_y^{e,ph}, e, ph)$ (6.24), del mismo 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 $r_{1,s_x^{ph},s_y^{e,ph}}$ (6.26). La obtención de la señal pseudo-estéreo a partir de los vectores de características ruidosos se realiza mediante el sistema identificado como "KPD-MEMLIN", 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).

desplazamiento y modelos de probabilidad entre Gaussianas) a partir de las expresiones (6.18) y (6.20) o (6.21).

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

La experimentación realizada con las técnicas de adaptación empíricas P-MEMLIN, MEMHIN y PD-MEMLIN, ésta última utilizando tanto la fase de entrenamiento con señal estéreo, como su versión "ciega", se llevó a cabo sobre la base de datos *SpeechDat Car* en español. A la hora de realizar la necesaria fase de entrenamiento previa se hará uso de los distintos corpora de entrenamiento asociados a cada entorno básico, utilizando bien señal estéreo, bien únicamente los vectores acústicos ruidosos, según el caso. Por otra parte, y una vez normalizados los vectores acústicos degradados con las correspondientes técnicas, se aplicará el método CMN. Para esta experimentación se utilizó la parametrización estándar ETSI y modelos acústicos fonéticos, de modo que bajo estas condiciones, los resultados de referencia se corresponden con los que se encuentran en la Tabla 4.3. Se puede apreciar asimismo que todos los parámetros que definen la experimentación en este caso coinciden con los aplicados en la Sección 5.5, de modo que los resultados presentados en ambas secciones son totalmente comparables. Por último, la Figura 5.5 sigue siendo válida para explicar los tres pasos precisados para realizar la experimentación.

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

Tabla 6.1: Mejores resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando distintas técnicas de adaptación de vectores de características. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con las técnicas MEMLIN, P-MEMLIN o MEMHIN. 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.

6.6.1 Resultados obtenidos con las técnicas P-MEMLIN y MEMHIN

En la Tabla 6.1 se pueden apreciar los mejores resultados para las técnicas de adaptación de vectores de características MEMLIN, cuyos resultados ya se presentaron en la Sección 5.5 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 pueden consultar en los Anexos 5.7 y 6.10). 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, además del WER medio, MWER, la mejora media de WER, MIMP, en tanto por ciento, y calculada del mismo modo que ya se comentó en el Capítulo 5.5 (ver expresión (5.29)).

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.

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 del WER para los distintos métodos en función del número de Gaussianas con que se modela cada entorno básico, Figura 6.3, sí se puede apreciar que los dos primeros proporcionan un mejor comportamiento

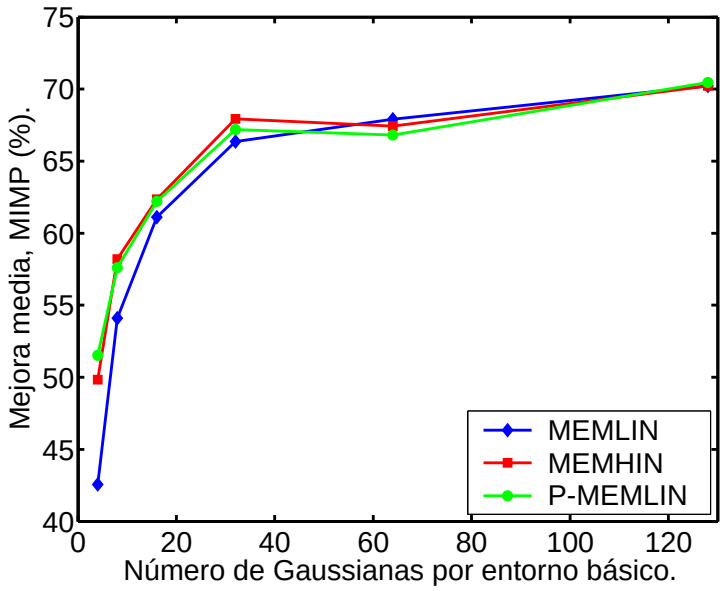


Figura 6.3: Mejora media del WER, MIMP, para las técnicas MEMLIN, MEMHIN y P-MEMLIN, atendiendo al número de componentes con que se modela cada entorno básico. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia.

cuando el número de componentes es reducido. Así, por ejemplo, si se aplica la técnica MEMLIN tras modelar 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 mayores: 51.53 % y 49.84 %, respectivamente; aunque, eso sí, dicha mejora queda reducida rápidamente conforme el número de componentes por entorno básico se eleva por encima de 8. Este comportamiento se debe a que la importancia de la compensación de la varianza de los vectores de características, que es lo que pretenden las técnicas P-MEMLIN y MEMHIN, es mayor cuando el modelado de los entornos básicos y el espacio limpio se realiza con un número reducido de Gaussianas. En ese caso los espacios representados por cada par componentes son mucho más variables y las transformaciones aprendidas en la fase de entrenamiento resultan más sensibles a la diferencia de varianza entre los vectores de características asociados al correspondiente par de componentes. En estas situaciones, un modelo más complejo del espacio de señal proporciona interesantes mejoras. Otro de los factores que puede hacer que técnicas como P-MEMLIN o MEMHIN tengan un mejor comportamiento es el tipo de ruido. Así, por ejemplo, el ruido aditivo, tal y como se ha podido apreciar en la Sección 5.2, afecta en buena medida a la varianza de los vectores de características, por lo que su compensación se adecua 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 comparando el comportamiento de los algoritmos MEMLIN y MEMHIN [BLMO04b] atendiendo a distintas SNRs. Para ello se empleó la base de datos *SpeechDat Car* en español, generando en este caso los nuevos corpora ruidosos tras incluir artificialmente ruido aditivo de vehículo a los correspondientes corpora limpios. Dicho ruido fue obtenido de la propia base de datos *SpeechDat Car* en español. Para este nuevo experimento se utilizó la parametrización estándar ETSI,

modelos acústicos fonéticos y 8 ó 16 Gaussianas para modelar los nuevos entornos básicos

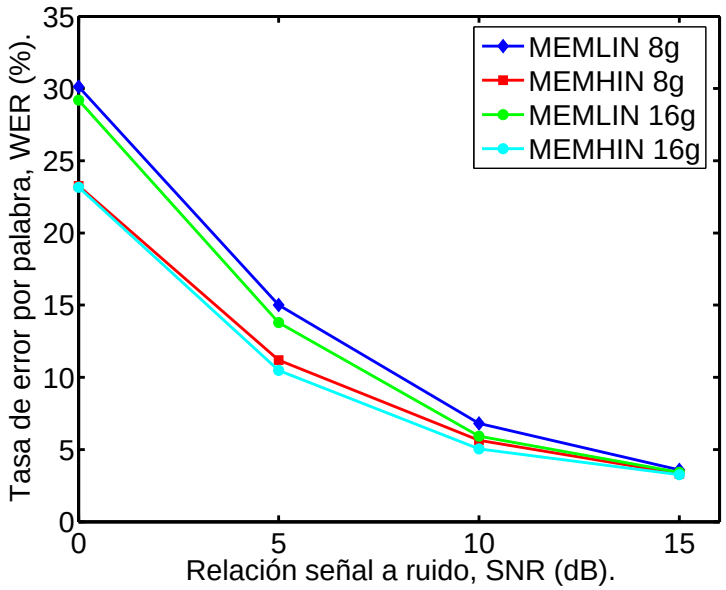


Figura 6.4: Mejora media del WER, MIMP, para las técnicas MEMHIN y MEMLIN, atendiendo al SNR. Se ha empleado la parametrización estándar ETSI, modelos acústicos fonéticos generados a partir de la señal limpia y 8 ó 16 Gaussianas para modelar los distintos entornos básicos y el espacio limpio. La señal 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.

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 (SNRs reducidas).

6.6.2 Resultados obtenidos con la técnica PD-MEMLIN

A continuación se comparan los resultados obtenidos con las técnicas MEMLIN y PD-MEMLIN. Para este último método se entrenan transformaciones para 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 adaptar un vector de características. Mediante este término se da una idea aproximada del coste computacional precisado 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 modela con el mismo número de Gaussianas, como

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

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

Tabla 6.2: Mejores resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando distintas técnicas de adaptación de vectores de características. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con las técnicas MEMLIN y PD-MEMLIN. Junto al nombre de las diferentes técnicas aparece el número de transformaciones por entorno básico precisado en $\log_{10}$, TpE. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

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 la cantidad de fonemas considerados ($n_{ph} = 1$, para la técnica MEMLIN). Para esta experimentación se utiliza el mismo número de componentes para modelar cada fonema del espacio limpio y de cada entorno ruidoso básico, pudiendo ser 2, 4, 8, 16 ó 32, y cuyos resultados completos se incluyen en el Anexo 6.10 de este mismo Capítulo.

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 (16), 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, como en otras ocasiones, 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 $w = 1{,}73 < 1{,}96$, 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 ente estos dos últimos métodos sí es estadísticamente significativo con un intervalo de confianza del 95 %. De todos modos, se recuerda una vez más que estos resultados se han de tratar con suma cautela dadas las limitaciones de la propia prueba, ya comentadas en la Sección 4.3.

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 el método PD-MEMLIN proporciona una significativa mejora relativa con respecto al algoritmo MEMLIN, independientemente del número de transformaciones por entorno básico que se emplee. Por su parte, la Figura 6.6 muestra el histograma y el *log-scattergram* del primer coeficiente MFCC de los vectores de características de voz limpios del entorno básico E4 y los correspondientes adaptados mediante el método

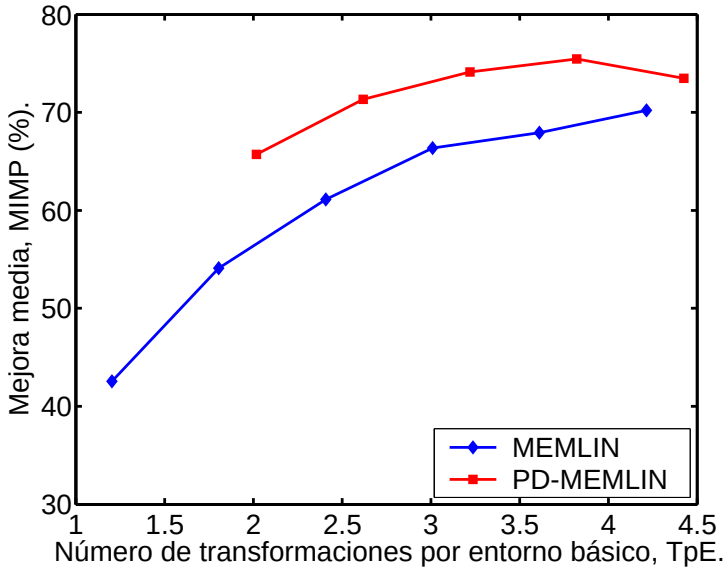
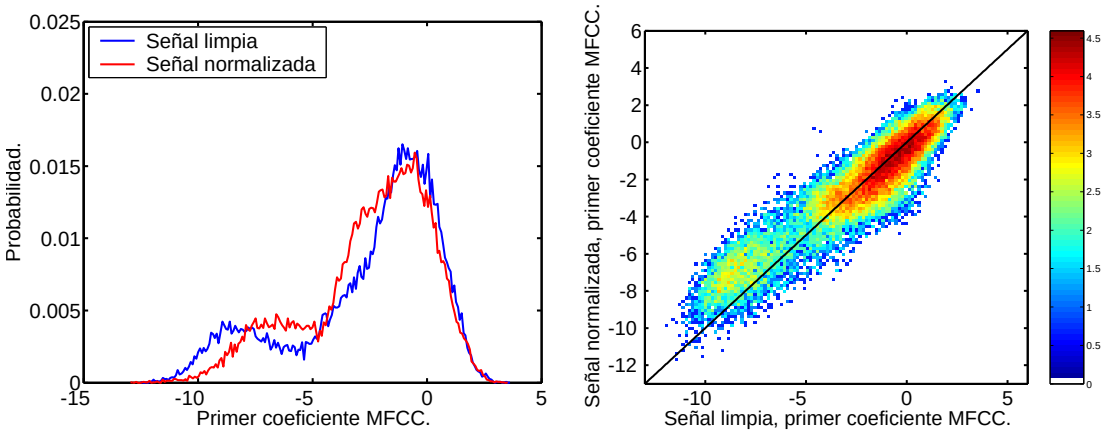


Figura 6.5: Mejora media del WER, MIMP, para las técnicas MEMLIN y PD-MEMLIN, atendiendo al número de transformaciones por entorno básico en $\log_{10}$. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia.



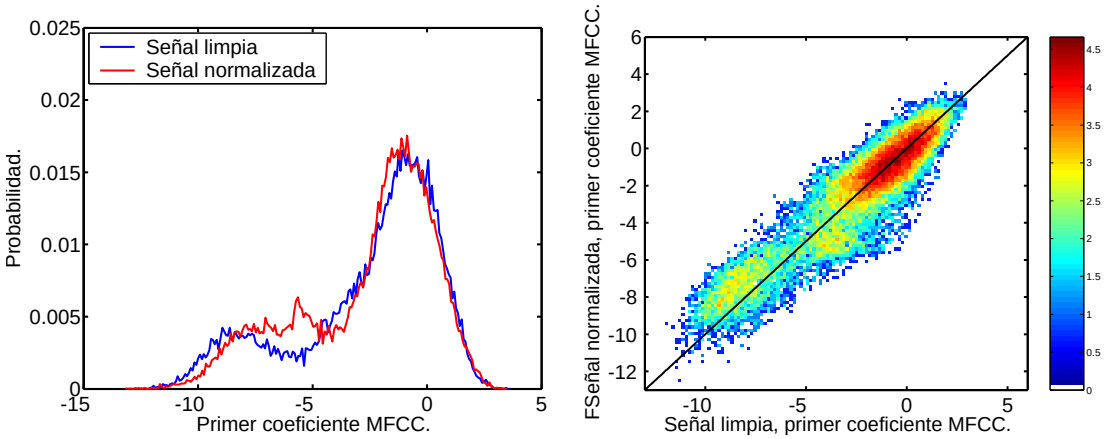
(a.1) (a.2)

Figura 6.6: *Log-scattergram* e histograma del primer coeficiente MFCC de los vectores de características de voz limpia y normalizada usando la técnica PD-MEMLIN con 16 Gaussianas por fonema y entorno básico. 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 el *log-scattergram* representa la función $x = y$.

PD-MEMLIN empleando 16 Gaussianas por fonema. A partir de esta Figura se puede concluir que las transformaciones propuestas por el algoritmo PD-MEMLIN corrigen el problema de la proyección de gran cantidad de vectores de características ruidosos hacia el silencio del espacio limpio, tal y como sucedía en la técnica MEMLIN (ver Figura 5.7.b); a su vez se observa una importante reducción de la incertidumbre, así como la eliminación de buena parte de los efectos que el entorno acústico introduce en los coeficientes de

Entre.	Reco.	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 KPD-MEMLIN 3.82	0.96	2.57	2.52	1.75	2.10	1.27	1.02	1.84	99.37

Tabla 6.3: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la pseudo-técnica de adaptación de vectores de características KPD-MEMLIN. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la pseudo-técnica KPD-MEMLIN. Junto a su nombre 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-scattergram* e histograma del primer coeficiente MFCC de los vectores de características de voz limpia y normalizada usando la pseudo-técnica KPD-MEMLIN con 16 Gaussianas por fonema y entorno básico. 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 el *log-scattergram* representa la función $x = y$.

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, llevándolo a nivel de componente asociada a un fonema dado, y produciendo así 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 aplicando la pseudo-técnica KPD-MEMLIN (ver Sección 6.5). Sus resultados más significativos se muestran en la Tabla 6.3, en la que además aparecen las tasas de reconocimiento obtenidas con la señal limpia ("Entre." CLK, "Reco." CLK), incluidas a modo de comparación. Dicha experimentación se realizó con la base de datos *SpeechDat Car* en español haciendo uso de la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir del corpus de entrenamiento de señal limpia ("Entre." CLK). Las correspondientes GMMs entrenadas para los distintos fonemas se componen de 16 Gaussianas para cada entorno básico y el espacio limpio. Asimismo, y por completar

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

Tabla 6.4: Tasa media de fonemas correctos, *Mean Correct Phoneme*, MCP, en % obtenidas con la base de datos *SpeechDat Car* en español para los diferentes entornos básicos (E1,..., E7) utilizando el algoritmo PD-MEMLIN (HF PD-MEMLIN) y la pseudo-técnica KPD-MEMLIN (HF KPD-MEMLIN). Se ha empleado la parametrización estándar ETSI y GMMs para las unidades fonéticas entrenadas con señal limpia. Para ambos métodos se modelan los fonemas con 16 Gaussianas para todos los entornos básicos y el espacio limpio.

el estudio de la pseudo-técnica KPD-MEMLIN, también se incluyen los consiguientes *log-scattergram* e histograma (Figura 6.7). Ambos se obtuvieron a partir del primer coeficiente MFCC de los vectores acústicos de voz provinientes de las señales limpia y normalizada para el corpus de reconocimiento del entorno básico E4 de la base de datos *SpeechDat Car* en español.

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 adaptados para el entorno básico E4 no se ha visto reducida considerablemente con respecto a la obtenida con la técnica PD-MEMLIN bajo las mismas condiciones de experimentación (Figura 6.6). Esto es debido a que las normalizaciones empleadas proyectan los vectores acústicos ruidosos al espacio limpio a nivel de fonema, que de por sí posee una cierta incertidumbre. Esto puede confirmarse mediante la definición y posterior estudio de la tasa media de fonemas correctos, *Mean Correct Phoneme*, MCP. Así, la tasa MCP se obtiene como relación de fonemas correctamente reconocidos, considerando como fonema correcto para cada vector acústico aquél proporcionado por la segmentación forzada de la señal limpia en términos de fonema a partir del algoritmo de Viterbi. Por su parte, el fonema reconocido será el que mayor verosimilitud proporcione haciendo uso de las GMMs del espacio limpio con que se modelan los distintos fonemas. De esta manera, en la decisión final no influye modelo de lenguaje o vocabulario alguno. Los resultados de la tasa MCP obtenidos con el algoritmo PD-MEMLIN y la pseudo-técnica KPD-MEMLIN para los distintos entornos básicos del corpus de reconocimiento de la base de datos *SpeechDat Car* en español se muestran en la Tabla 6.4, componiéndose cada GMM asociada a entorno básico y fonema de 16 Gaussianas. Nótese como la pseudo-técnica KPD-MEMLIN mejora, en media, los resultados proporcionados por el método PD-MEMLIN más allá del 10 %, a pesar de que, como ya se ha constatado anteriormente, la incertidumbre apenas se ve reducida. De todo lo anterior se puede concluir que el hecho de no reducir la incertidumbre entre los coeficientes de los vectores acústicos limpios y adaptados no implica que la normalización no sea satisfactoria a nivel de RAH.

A partir de las Tablas 6.3 y 6.4, se puede concluir que las transformaciones dependientes de fonema consideradas para el método PD-MEMLIN son consistentes con respecto a los modelos acústicos, ya que los vectores de características se proyectan desde el espacio ruidoso al limpio a nivel de fonemas. Asimismo, y gracias a los estudios realizados a partir de la pseudo-técnica KPD-MEMLIN, se puede definir una futura línea de trabajo

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

Tabla 6.5: Mejores resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando distintas técnicas de adaptación de vectores de características. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con las técnicas MEMLIN, PD-MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega". Junto al nombre de las diferentes técnicas 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.

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 verificación e identificación de locutor dependientes del texto en entornos acústicos adversos. En este sentido ya se han realizado unas primeras pruebas preliminares [BLR+06] [BLMO05a] [BLR+05], 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 locutor, no es deseable. A pesar de ello se obtuvieron importantes mejoras en ambas tareas.

6.6.3 Resultados obtenidos con 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 presentada en la Figura 6.2. Para realizar la correspondiente adaptación con la técnica PD-MEMLIN, independientemente de su fase de entrenamiento, y al igual que la subsección 6.6.2, se entrenaron y emplearon transformaciones para los 25 fonemas españoles más el silencio, pudiéndose modelar con 2, 4, 8, 16 ó 32 componentes. Nótese que las condiciones de la experimentación son las mismas que las definidas previamente para la subsecciones 6.6.2 y 6.6.1, 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 debe calcular para adaptar cada vector de características ruidoso. Asimismo adviértase que los resultados completos se incluyen en el Anexo 6.10 de este mismo Capítulo.

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

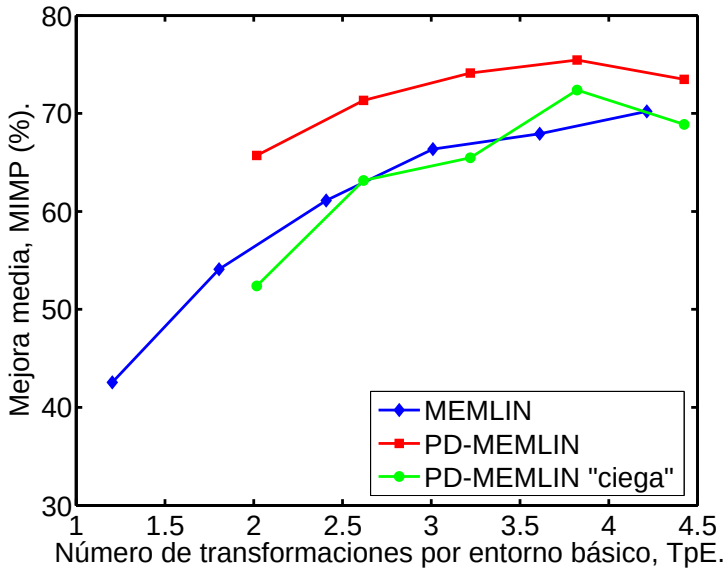


Figura 6.8: Mejora media del WER, MIMP, para las técnicas MEMLIN, PD-MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega", atendiendo al número de transformaciones por entorno básico en $\log_{10}$. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia.

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 4.3.

Asimismo, y para estudiar la tendencia del comportamiento de los distintos algoritmos considerados en esta subsección, en la Figura 6.8 se representan las correspondientes mejoras medias de WER, MIMP, 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, que los logrados por el método MEMLIN para los distintos valores de TpE estudiados. Todo ello con la ventaja añadida de que no es necesario disponer se señal estéreo para realizar la fase de entrenamiento previa, aunque eso sí, ésta es supervisada, lo que en ciertas circunstancias podría ser un inconveniente. De todos modos, las mejoras obtenidas con esta técnica aún quedan lejos del mejor resultado alcanzado hasta el momento por el método PD-MEMLIN con fase de entrenamiento con señal estéreo.

A modo de resumen se incluyen en la Tabla 6.6 los resultados más significativos,

Entre.	Reco.	TpE	MWER (%)	MIMP (%)
CLK	HF MEMLIN	4.21	6.05	70.22
CLK	HF MEMHIN	4.21	6.05	70.22
CLK	HF P-MEMLIN	4.21	6.02	70.47
CLK	HF PD-MEMLIN	3.82	5.30	75.44
CLK	HF PD-MEMLIN "ciega"	3.82	5.74	72.40

Tabla 6.6: Mejores resultados medios obtenidos con la base de datos *SpeechDat Car* en español en términos de WER y mejora media en WER (MIMP) (%) utilizando distintas técnicas de adaptación de vectores de características. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con las técnicas MEMLIN, MEMHIN, P-MEMLIN, PD-MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega". Junto al nombre de las diferentes técnicas, en la columna referenciada como "TpE", aparece el número de transformaciones por entorno básico en $\log_{10}$ precisadas en cada caso.

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 para cada caso el TpE requerido. 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 emplean señal estéreo en su correspondiente fase de entrenamiento. Sin ambargo, en ambos casos es necesaria la trascripción del corpus de entrenamiento, lo que no deja de ser una limitación con respecto al resto de algoritmos.

6.7 Anexo C.

En este Anexo se muestra el desarrollo teórico necesario para estimar la matriz diagonal asociada al término de pendiente, $\mathbf{A}_{s_x,s_y^e}$, y el vector que representa el término independiente, $\mathbf{b}_{s_x,s_ye}$, del modelo del espacio de señal asociado 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 superíndice Tr que aparecía en la nomenclatura de la Sección 6.2 para indicar que se trataba del corpus de entrenamiento. El criterio elegido para estimar las dos variables anteriormente comentadas consiste, tal y como se indicó en la Sección 6.2, en igualar tanto la media como la desviación típica de los vectores de características limpios y los obtenidos mediante el modelo del espacio de señal $(\mathbf{\Psi}(\mathbf{y}_t, s_x, s_y^e) = \mathbf{A}_{s_x,s_y^e}\mathbf{y}_t - \mathbf{b}_{s_x,s_y^e})$. para cada par de Gaussianas, s_x y s_y^e. 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 $\sqrt{\ }$ realiza la raíz cuadrada elemento a elemento de la matriz correspondiente. 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 $\mathbf{z}$ se definen del siguiente modo

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

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

donde el operador $diag[\ ]$ hace nulos todos los elementos de la matriz correspondiente distintos de la diagonal. Al considerar que no hay dependencia alguna entre las componentes de los vectores de características, el sistema de ecuaciones propuesto mediante las expresiones (C.1) y (C.2) se puede ver como tantos sistemas independientes 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

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

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

donde i representa el índice de la componente i-ésima, ya sea de los vectores de medias o de 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 presentadas previamente en la Sección 6.2, (6.3) y (6.4), respectivamente.

6.8 Anexo D.

En este Anexo se incluye el desarrollo teórico necesario para estimar el correspondiente vector de desplazamiento, $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$, correspondinte al modelo del espacio de señal para la técnica PD-MEMLIN. Para ello se hace uso de la minimización del error cuadrático medio asociado a cada par de Gaussianas, s_x^{ph} y $s_y^{e,ph}$, identificado como $\xi_{s_x^{ph},s_y^{e,ph}}$ (D.1). Sea pues 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 6.4.

$$\begin{aligned}\xi_{s_x^{ph},s_y^{e,ph}} &= \frac{1}{T_{e,ph}} \sum_{t_{e,ph}} p(s_x|\mathbf{x}_{t_{e,ph}}^{e,ph}, e) p(s_y^e|\mathbf{y}_{t_{e,ph}}^{e,ph}, e) \\ &\times Tra\left[\left(\mathbf{x}_{t_{e,ph}}^{e,ph} - \Psi(\mathbf{y}_{t_{e,ph}}^{e,ph}, s_x^{ph}, s_y^{e,ph})\right)\left(\mathbf{x}_{t_{e,ph}}^{e,ph} - \Psi(\mathbf{y}_{t_{e,ph}}^{e,ph}, s_x^{ph}, s_y^{e,ph})\right)^T\right], \quad \text{(D.1)}\end{aligned}$$

donde el modelo del espacio de señal es $\Psi(\mathbf{y}_{t_{e,ph}}^{e,ph}, s_x^{ph}, s_y^{e,ph}) = \mathbf{y}_{t_{e,ph}}^{e,ph} - \mathbf{r}_{s_x^{ph},s_y^{e,ph}}$. 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

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

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).

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

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 para $\mathbf{r}_{s_x^{ph},s_y^{e,ph}}$

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

6.9 Anexo E.

En este Anexo se calcula la distancia de Kullback Leibler para dos funciones de densidad de probabilidad Gaussianas, p y q, $KL(p,q)$. Dado que p y q son continuas, la expresión que se desea obtener será, por definición,

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

donde se asume que p y q son pdfs de una variable vectorial, $\mathbf{x}$, y que, para este estudio, serán

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

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

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 del vector $\mathbf{x}$. Si se evalúa el término logarítmico de la expresión (E.1), introdciendo (E.2) y (E.3) se tiene que

$$\begin{aligned} log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) &= log\left(\frac{c_p}{c_q}\right) + \frac{1}{2} log\left(\frac{|\mathbf{\Sigma}_q|}{|\mathbf{\Sigma}_p|}\right) \\ &- \frac{1}{2}\left((\mathbf{x}-\mu_p)^T\mathbf{\Sigma}_p^{-1}(\mathbf{x}-\mu_p) - (\mathbf{x}-\mu_q)^T\mathbf{\Sigma}_q^{-1}(\mathbf{x}-\mu_q)\right). \end{aligned} \tag{E.4}$$

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

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

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

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

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

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

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

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

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

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, respectivamente, 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

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

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

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

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

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

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

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

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

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

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.

6.10 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: *Expectation*, E, y *Maximization*, M. En el paso E se obtiene 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 $t_{e,ph} \in [1, T_{e,ph}]$. Apréciese que se ha eliminado con respecto a la Sección 6.5 el superíndice Tr por simplificar la notación. Por otra parte, se dispone de las GMMs que modelan los vectores de características limpios y ruidosos para cada entorno y fonema: (6.9) (6.10), (6.11) y (6.12). A partir de todo lo anterior 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

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

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 $Q(\phi, \phi_{new})_e^{ph}$ (F.3), en la que $\phi = \{\mathbf{r}_{s_x^{ph},s_y^{e,ph}}\}$ se corresponde con el vector de desplazamiento que se dispone en cada iteración, esto es, el obtenido en la iteración precedente, mientras que $\phi_{new} = \{\mathbf{r}_{new,s_x^{ph},s_y^{e,ph}}\}$ será el nuevo vector de desplazamiento calculado.

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

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 se asume que $p(\mathbf{y}_{t_{e,ph}}^{e,ph}, s_x^{ph}, s_y^{e,ph}|\phi_{new}, e, ph) \simeq p(s_x^{ph}, s_y^{e,ph})p(\mathbf{y}_{t_{e,ph}}^{e,ph}|s_x^{ph}, s_y^{e,ph}\phi_{new}, e, ph)$, la expresión (F.3) se transforma en

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

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.

$$\begin{aligned}\mathbf{r}_{new,s_x^{ph},s_y^{e,ph}} &= \frac{\delta(Q(\phi,\phi_{new})_e^{ph})}{\delta(\mathbf{r}_{new,s_x^{ph},s_y^{e,ph}})} \\ &= \sum_{t_{e,ph}} p(s_x^{ph},s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph},\phi,e,ph)\frac{\delta(-\frac{1}{2}log|\mathbf{\Sigma}_{s_y^{e,ph}}|-\frac{1}{2}\mathbf{\Omega}^T\mathbf{\Sigma}_{s_y^{e,ph}}\mathbf{\Omega})}{\delta(\mathbf{r}_{new,s_x^{ph}},s_y^{e,ph})} = \mathbf{0}, \quad \text{(F.5)}\end{aligned}$$

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

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

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

$$\begin{aligned}p(s_x^{ph},s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph},\phi,e,ph) &\simeq p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph},e,ph)p(s_x^{ph}|\mathbf{y}_{t_{e,ph}}^{e,ph},s_y^{e,ph},\phi,e,ph) \\ &= p(s_y^{e,ph}|\mathbf{y}_{t_{e,ph}}^{e,ph},e,ph)\frac{p(s_y^{e,ph})p(\mathbf{y}_{t_{e,ph}}^{e,ph}|s_x^{ph},s_y^{e,ph},\phi)}{\sum_{s_x^{ph}}p(s_y^{e,ph})p(\mathbf{y}_{t_{e,ph}}^{e,ph}|s_x^{ph},s_y^{e,ph},\phi)}, \quad \text{(F.8)}\end{aligned}$$

que coincide con la expresión presentada en la Sección 6.5.

6.11 Anexo G.

En este Anexo se presentan los resultados en términos de WER (%) obtenidos para los diferentes entornos básicos (E1,..., E7) de la base de datos *SpeechDat Car* en español utilizando distintas técnicas de adaptación de vectores de características (P-MEMLIN, MEMHIN, PD-MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega"). Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia para las unidades fonéticas. A su vez, y junto al nombre de las distintas técnicas, se ha incluido el número de componentes con que se modelan los correspondientes entornos básicos (4, 8, 16, 32, 64 y 128 Gaussianas), o bien el número de las mismas con que se representa cada fonema para los diferentes espacios, si se trata de las técnicas PD-MEMLIN y PD-MEMLIN con fase de entrenamiento "ciega" (2, 4, 8, 16 y 32).

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF P-MEMLIN 4	3.26	9.61	6.57	10.40	11.82	8.73	14.97	8.76	51.53
CLK	HF P-MEMLIN 8	3.07	8.06	6.71	8.15	11.63	8.41	11.91	7.88	57.60
CLK	HF P-MEMLIN 16	2.88	7.89	6.57	7.27	9.91	7.46	11.22	7.21	62.21
CLK	HF P-MEMLIN 32	2.88	7.72	5.17	6.52	8.96	6.35	9.18	6.49	67.18
CLK	HF P-MEMLIN 64	2.59	7.72	5.73	7.02	8.96	6.35	8.50	6.55	66.82
CLK	HF P-MEMLIN 128	2.30	7.80	4.90	5.89	8.39	5.71	7.48	6.02	70.47

Tabla 6.7: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características P-MEMLIN. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica P-MEMLIN. Junto al nombre de la técnica aparece el número de Gaussianas con que se modelaron los distintos entornos básicos. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF MEMHIN 4	3.16	9.86	6.57	10.40	11.63	10.48	15.99	9.00	49.84
CLK	HF MEMHIN 8	3.07	7.98	6.57	8.02	11.63	7.78	12.59	7.79	58.20
CLK	HF MEMHIN 16	2.97	7.98	6.15	7.14	9.91	7.46	11.56	7.19	62.34
CLK	HF MEMHIN 32	2.40	7.80	5.31	6.52	8.77	6.51	8.50	6.39	67.92
CLK	HF MEMHIN 64	2.49	7.72	5.73	7.02	8.77	6.19	8.16	6.46	67.43
CLK	HF MEMHIN 128	2.21	7.89	5.17	6.02	8.29	5.56	7.82	6.05	70.22

Tabla 6.8: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características MEMHIN. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica MEMHIN. Junto al nombre de la técnica aparece el número de Gaussianas con que se modelaron los distintos entornos básicos. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF PD-MEMLIN 2	3.16	8.66	4.90	6.52	9.63	5.08	9.52	6.70	65.72
CLK	HF PD-MEMLIN 4	2.59	8.40	5.45	5.39	7.53	3.81	8.84	5.90	71.32
CLK	HF PD-MEMLIN 8	2.49	7.89	5.73	5.26	6.67	3.65	6.48	5.49	74.11
CLK	HF PD-MEMLIN 16	1.73	8.23	5.45	4.64	6.86	3.02	7.14	5.30	75.44
CLK	HF PD-MEMLIN 32	2.21	7.98	4.48	5.39	7.91	4.29	5.78	5.58	73.48

Tabla 6.9: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características PD-MEMLIN. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica PD-MEMLIN. Junto al nombre de la técnica aparece el número de Gaussianas con que se modelaron los distintos fonemas para cada entorno básico. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF PD-MEMLIN "ciego" 2	3.16	9.43	7.41	8.90	12.96	6.83	15.65	8.63	52.39
CLK	HF PD-MEMLIN "ciego" 4	3.36	7.98	5.03	6.89	10.87	5.24	12.59	7.07	53.15
CLK	HF PD-MEMLIN "ciego" 8	3.36	7.54	5.03	6.77	9.82	5.40	11.56	6.74	65.49
CLK	HF PD-MEMLIN "ciego" 16	2.59	6.43	4.34	6.14	8.39	4.44	9.86	5.74	72.40
CLK	HF PD-MEMLIN "ciego" 32	2.68	7.29	4.90	6.39	8.96	4.13	12.59	6.25	68.88

Tabla 6.10: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características PD-MEMLIN con fase de entrenamiento "ciega". Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica PD-MEMLIN con fase de entrenamiento "ciega". Junto al nombre de la técnica aparece el número de Gaussianas con que se modelaron los distintos fonemas para cada entorno básico. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Capítulo 7

Mejoras en el Modelado de Probabilidad Condicionada entre Espacios de Señal.

7.1 Introducción.

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 del espacio de señal, ya se ha tratado convenientemente en el Capítulo 6. Ahora conviene estudiar el modelo de la probabilidad condicionada entre espacios de señal, que, principalmente, se manifiesta en el término correspondiente a la probabilidad entre Gaussianas, $p(s_x|\mathbf{y}_t, e, s_y^e)$. Adviértase que dicho término posee gran importancia puesto que determina, a nivel de Gaussiana, la región de proyección del vector acústico ruidoso dentro del espacio limpio y, por tanto, la incertidumbre en la que se va a mover el vector de características normalizado, que estará en función de las varianzas de las distintas Gaussianas que modelan el espacio limpio. Hasta el momento, en los distintos métodos presentados en este trabajo, el término de probabilidad entre Gaussianas se ha estimado siempre eliminando la dependencia con $\mathbf{y}_t$, lo que supone considerar que el vector acústico limpio asociado al correspondiente degradado ha sido generado por una Gaussiana, s_x, que únicamente depende de la componente s_y, menospreciando, en cierto modo, la aleatoriedad introducida por el ruido del entorno acústico concreto.

Para compensar las deficiencias del modelado de la probabilidad condicionada entre espacios de señal considerado en las técnicas presentadas hasta el momento, se propone modelar mediante una GMM aquellos 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 [BLN+06] [BML+07a], o s_x^{ph} y $s_y^{e,ph}$ si se trata del algoritmo PD-MEMLIN [BLM+06]. Estas nuevas GMMs se estiman en la fase de entrenamiento previa mediante señal estéreo y, del mismo modo que proporcionan una interesante mejora de los resultados en términos de RAH, también son responsables de incrementar el coste computacional considerablemente. Sin embargo, tal y como se indicará en este Capítulo, dicha limitación puede ser minimizada si se reduce el número de pares de Gaussianas computadas.

En este Capítulo se presenta primeramente (Sección 7.2) un estudio sobre los efectos, tanto cualitativos como cuantitativos, que la probabilidad entre Gaussianas introduce en las técnicas MEMLIN y PD-MEMLIN. A raíz de los resultados presentados no sólo se podrá afirmar que el término en cuestión posee una importancia capital, sino que además el margen de mejora en términos de RAH al que se puede llegar a aspirar es muy importante. Una vez constatadas las limitaciones de la aproximación del modelo entre Gaussianas aplicada hasta el momento en las técnicas MEMLIN y PD-MEMLIN, se procede a exponer en la Sección 7.3 la solución propuesta en este sentido, que consiste, como ya se ha adelantado, en introducir unas nuevas GMMs para representar los vectores de características ruidosos asociados a cada par de Gaussianas: s_x y s_y^e en el caso del método MEMLIN, o s_x^{ph} y $s_y^{e,ph}$ si se trata del algoritmo PD-MEMLIN. En la Sección 7.4 se propone la adaptación de las expresiones presentadas en la Sección anterior para las técnicas MEMLIN y PD-MEMLIN. Los resultados de RAH obtenidos tras aplicar el nuevo modelado de probabilidad condicionada entre espacios de señal a 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 7.5. En ellos queda patente el buen comportamiento de las dos extensiones propuestas, no sólo con respecto a los métodos de adaptación de vectores de características empíricos basados en el criterio MMSE más utilizados en la actualidad por la comunidad científica (CMN, RATZ y SPLICE), sino también si se comparan con la técnicas MEMLIN y PD-MEMLIN.

7.2 Efectos del Modelado de la Probabilidad Condicionada entre Espacios de Señal.

Desde un primer momento se pensó que el modelado de la probabilidad condicionada entre espacios de señal podía desempeñar un papel capital en las técnicas de adaptación de vectores de características presentadas en este trabajo. En realidad, conceptualmente hablando, este término tiene la capacidad de determinar, a nivel de Gaussiana, la región de proyección del vector de características ruidoso dentro del espacio limpio y, por tanto, el rango de incertidumbre en el que se puede mover el vector acústico normalizado, que estará en función de las varianzas de las Gaussianas que modelan el espacio limpio. Esta suposición, sin embargo, sólo proporciona una cierta idea cualitativa de la importancia del modelado de la probabilidad condicionada entre espacios de señal. Para certificarla, y además determinar cuantitativamente cuan importante es dicho término, se realizaron sendos experimentos de RAH para las técnicas MEMLIN y PD-MEMLIN.

Para ello se modificaron ambos métodos de modo que 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 $p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph}) \simeq p(s_x^{ph}|\mathbf{x}_t, ph)$, si se trata del método PD-MEMLIN. De esta manera, el cálculo de las dos nuevas variables se realizará haciendo uso de (5.7) y (5.8) o de (6.11) y (6.12), respectivamente, tal y como se indica a continuación

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

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

Entre.	Reco.	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	HF MEMLIN 128	1.63	3.52	1.82	1.50	2.29	0.79	0.35	1.99	98.36
CLK	HF PD-MEMLIN 16	1.25	3.78	2.66	2.13	3.53	1.27	1.36	2.50	94.83

Tabla 7.1: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando distintas técnicas de adaptación de vectores de características. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la limpia (CLK), ruidosa (HF) o ruidosa normalizada con las técnicas MEMLIN o PD-MEMLIN cuando se emplea señal limpia para determinar el modelado de la probabilidad entre Gaussianas. Junto al nombre de las diferentes técnicas aparece el número de Gaussianas con que se modelaron bien los correspondientes espacios (MEMLIN), bien los distintos fonemas (PD-MEMLIN). Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Con este experimento, tanto más alejado de la realidad conforme mayor sea el número de componentes con que se modele el espacio limpio, se pretende conocer el límite de los métodos MEMLIN y PD-MEMLIN al hacer óptimo el modelado de la probabilidad condicionada entre espacios de señal. 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, las correspondientes tasas obtenidas cuando se reconoce la señal limpia ("Entre." CLK, "Reco." CLK) y ruidosa ("Entre." CLK, "Reco." 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 estándar ETSI y modelos acústicos fonéticos generados a partir de la señal de entrenamiento limpia ("Entre." CLK). 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; mientras que para el algoritmo modificado PD-MEMLIN se han utilizado 16 Gaussianas para componer la GMM asociada a cada fonema. Nótese que la experimentación propuesta posee los mismos parámetros que las expuestas en las Secciones 5.5 y 6.6, 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 condicionada entre espacios de señal es muy elevado, tanto que permite acercarse a la adaptación perfecta (100 % de MIMP). De este modo se certifica, ya de un modo cuantitativo, la suposición que sobre la importancia de este término 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, además de la probabilidad entre Gaussianas, que influye de un modo importante y que en esta ocasión no se ha optimizado: 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 6.6 al presentarse la pseudo-técnica KPD-MEMLIN.

Por otra parte, en la Figura 7.1 se presentan los histogramas y *log-scattergrams*

obtenidos a partir del primer coeficiente MFCC de los vectores de características de voz de la señal limpia y la normalizada tras aplicar las dos extensiones propuestas anteriormente. 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. Nótese que en las Figuras 7.1.a se vuelven a incluir, a modo de comparación, las representaciones obtenidas a partir de las correspondientes señales limpia y ruidosa, pudiéndose apreciar nuevamente el efecto que el entorno acústico produce en los coeficientes de la señal limpia, tanto en términos de pdf (Figura 7.1.a.1), como de incertidumbre (Figura 7.1.a.2). Por otra parte, 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-scattergram* (Figura 7.1.b.2). Si se comparan estas representaciones con las obtenidas con el método MEMLIN convencional (Figuras 5.7.b), se puede constatar como el histograma normalizado se acerca sobremanera al de la señal limpia, a la vez que se reduce sensiblemente la incertidumbre. Por último, en las Figuras 7.1.c se incluyen las gráficas obtenidas tras aplicar la modificación de la técnica PD-MEMLIN, presentándose tanto los histogramas de la señal limpia y la normalizada (Figura 7.1.c.1), como el correspondiente *log-scattergram* (Figura 7.1.c.2). Si se comparan estas representaciones con las correspondientes asociadas a la técnica PD-MEMLIN convencional (Figuras 6.6.b), se puede certificar nuevamente una mejor aproximación del histograma normalizado con respecto al de la señal limpia, así como una 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, con la señal adaptada, se podría llegar a alcanzarse resultados de RAH propios de la señal limpia, hecho este que no siempre es posible decir de otras técnicas de adaptación de vectores de características. Sin embargo, todo esto pasa por mejorar considerablemente el modelado de la probabilidad condicionada entre espacios de señal o, más concretamente, la probabilidad entre Gaussianas. A continuación se presenta la solución propuesta, que consiste en modelar mediante GMMs 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.

7.3 Modelado de la Probabilidad entre Gaussianas Basado en GMMs.

Tal y como se ha mencionado anteriormente, para mejorar el modelado de la probabilidad condicionada entre espacios de señal, y más concretamente el término asociado a la probabilidad entre Gaussianas, que hasta ahora se ha aproximado siempre mediante una expresión independiente del vector de características ruidoso, se propone hacer uso de GMMs, de modo que representen a los vectores de características degradados asociados a cada par de Gaussianas de los modelos de los entornos básicos y el espacio limpio. Sin embargo, en la definición de la GMM propuesta, y de cara a simplificar la notación, se considerará únicamente un entorno básico y un fonema, de modo que se eliminarán dichas dependencias. Obsérvese que 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 6.4 y 5.4), a cada vector

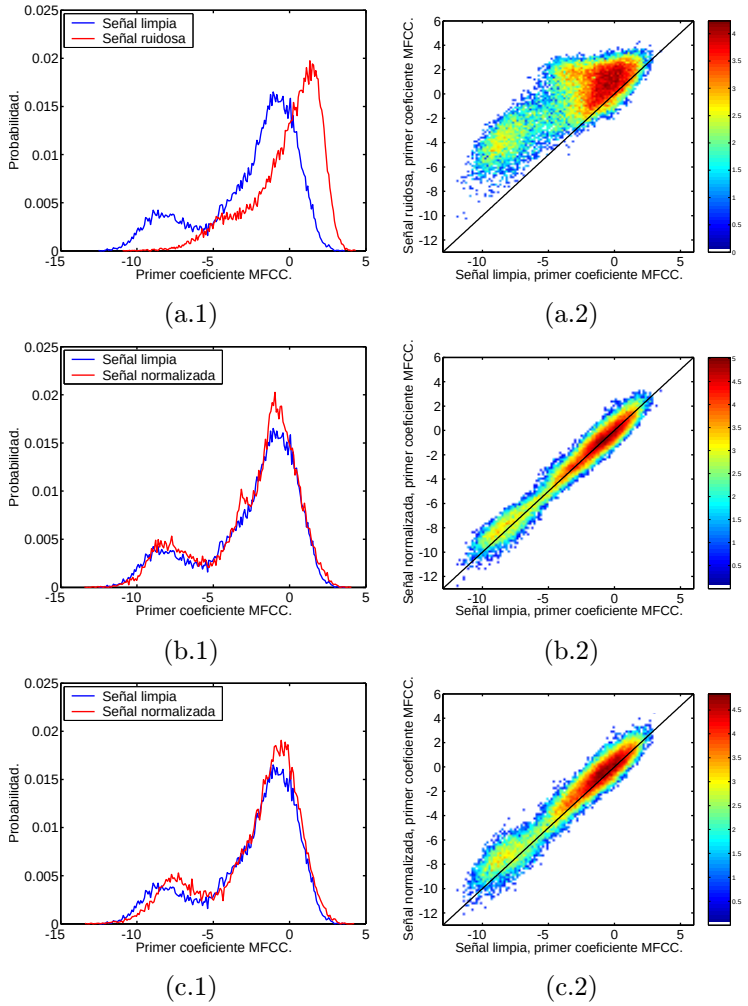


Figura 7.1: *Log-scattergrams* e histogramas del primer coeficiente MFCC de los vectores de características de voz limpia y ruidosa (a), o normalizada usando la técnica MEMLIN con 128 Gaussianas por entorno básico y señal limpia para calcular el modelado de la probabilidad entre Gaussianas (b). En la figura (c) se representa el *log-scattergram* y el histograma obtenidos tras aplicar la técnica PD-MEMLIN con 16 Gaussianas por fonema y 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-scattergrams* representa la función $x = y$.

de características del corpus de entrenamiento se le puede asociar una etiqueta como perteneciente a un entorno básico y fonema concretos.

Sea pues la GMM compuesta por C''' componentes que modela los vectores de características ruidosos asociados al par de Gaussianas de los espacios limpio y degradado s_x y s_y (se considera que los vectores acústicos 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 algoritmo EM [DLR77] tal y como se expone en el Anexo 7.6 de este mismo Capítulo.

7.4 Aplicación del Modelado de Probabilidad entre Gaussianas Basado en GMMs a las Técnicas MEMLIN y PD-MEMLIN.

Hasta el momento simplemente se ha propuesto un modelo basado en GMMs para representar los vectores de características ruidosos asociados a cada par de Gaussianas. Sin embargo, no se ha indicado como aplicar este nuevo modelado a las técnicas de adaptación de vectores de características propuestas en este trabajo. Llegado a este punto, se podrían considerar todas ellas: MEMLIN, P-MEMLIN, MEMHIN y PD-MEMLIN, sin embargo se tomarán únicamente las dos más representativas, esto es: MEMLIN y PD-MEMLIN. A continuación se presentan las extensiones de ambos algoritmos al incluir el nuevo modelado de la probabilidad entre Gaussianas. No obstante, la aplicación a las técnicas P-MEMLIN y MEMHIN es directa a partir de la solución desarrollada para el método MEMLIN.

7.4.1 Extensión para la técnica MEMLIN: MEMLIN MP.

Para extender el método MEMLIN incluyendo el nuevo Modelado de Probabilidad condicionada entre espacios de señal, MEMLIN MP, es necesario, tal y como ya se adelantó, estimar los parámetros de las correspondientes GMMs asociadas a cada entorno básico, e [BLN+06] [BML+07a]. Esto se realiza de modo independiente para cada uno de ellos, de manera que se obtiene el consiguiente modelo

$$p(\mathbf{y}_t|s_x, s^e_y, e) = \sum_{s'_y} p(\mathbf{y}_t|s_x, s^e_y, s'_y, e) p(s'_y|s_x, s^e_y, e), \tag{7.5}$$

$$p(\mathbf{y}_t|s_x, s^e_y, s'_y, e) = \mathcal{N}(\mathbf{y}_t; \mu_{s_x,s^e_y,s'_y}, \mathbf{\Sigma}_{s_x,s^e_y,s'_y}), \tag{7.6}$$

donde μ_{s_x,s^e_y,s'_y}, $\mathbf{\Sigma}_{s_x,s^e_y,s'_y}$, 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 el par de Gaussianas s_x y s^e_y, y cuyas expresiones, (H.15), (H.21) y (H.24), que se pueden consultar en el Anexo 7.6, se obtienen con la correspondiente señal estéreo del corpus de entrenamiento de cada entorno básico. Con todo ello la nueva estimación de $p(s_x|\mathbf{y}_t, e, s^e_y)$ se calcula del siguiente modo

$$p(s_x|\mathbf{y}_t, e, s^e_y) = \frac{p(\mathbf{y}_t|s_x, s^e_y, e) p(s_x|s^e_y, e)}{\sum_{s_x} p(\mathbf{y}_t|s_x, s^e_y, e) p(s_x|s^e_y, e)}. \tag{7.7}$$

Nótese que 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, a la vez que el algoritmo sigue haciendo uso de un proceso de entrenamiento plenamente no supervisado. Por otra parte, se sigue empleando el término $p(s_x|s_y^e, e)$, que se estima como (5.27) o (5.28) y que en este caso hace la función de probabilidad a priori de las distintas GMMs entrenadas, puesto que no todos los pares de componentes s_x y s_y^e son igualmente probables. Si se desease realizar la correspondiente extensión para incluir el modelado de la probabilidad condicionada 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.

7.4.2 Extensión para la técnica PD-MEMLIN: PD-MEMLIN MP.

Para extender el método PD-MEMLIN incluyendo el nuevo Modelado de Probabilidad condicionada entre espacios de señal, PD-MEMLIN MP, es necesario estimar los parámetros de las correspondientes GMMs para cada entorno básico, e, y fonema, ph [BLM+06]. Esto se lleva a cabo de modo independiente, obteniéndose el siguiente modelo

$$p(\mathbf{y}_t|s_x^{ph}, s_y^{e,ph}, e, ph) = \sum_{s'_y} p(\mathbf{y}_t|s_x^{ph}, s_y^{e,ph}, s'_y, e, ph)p(s'_y|s_x^{ph}, s_y^{e,ph}, e, ph), \tag{7.8}$$

$$p(\mathbf{y}_t|s_x^{ph}, s_y^{e,ph}, s'_y, e, ph) = \mathcal{N}(\mathbf{y}_t; \mu_{s_x^{ph}, s_y^{e,ph}, s'_y}, \mathbf{\Sigma}_{s_x^{ph}, s_y^{e,ph}, s'_y}), \tag{7.9}$$

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 el par de componentes s_x^{ph} y $s_y^{e,ph}$, y cuyas expresiones, que se evalúan con la señal estéreo del corpus de entrenamiento para cada entorno básico y fonema, (H.15), (H.21) y (H.24), se pueden consultar en el Anexo 7.6 de este mismo Capítulo. A partir de lo anterior, la nueva estimación de $p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph})$ se obtiene del siguiente modo

$$p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph}) = \frac{p(\mathbf{y}_t|s_x^{ph}, s_y^{e,ph}, e, ph)p(s_x^{ph}|s_y^{e,ph}, e, ph)}{\sum_{s_x^{ph}} p(\mathbf{y}_t|s_x^{ph}, s_y^{e,ph}, e, ph)p(s_x^{ph}|s_y^{e,ph}, e, ph)}. \tag{7.10}$$

Nuevamente se ha eliminado la aproximación $p(s_x^{ph}|\mathbf{y}_t, e, ph, s_y^{e,ph}) \simeq p(s_x^{ph}|e, ph, s_y^{e,ph})$. Por su parte, la expresión $p(s_x^{ph}|s_y^{e,ph}, e, ph)$, que se estima como (6.20) o (6.21), hace las veces de probabilidad a priori de las distintas GMMs entrenadas.

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

La experimentación realizada con las técnicas de adaptación empíricas MEMLIN MP y PD-MEMLIN MP se llevó a cabo con la base de datos *SpeechDat Car* en español. A la hora de realizar las distintas fases de entrenamiento se hará uso del corpus de entrenamiento correspondiente a cada entorno básico y fonema, esto último sólo si procede. Por otra

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

Tabla 7.2: Mejores resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando distintas técnicas de adaptación de vectores de características. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con las técnicas MEMLIN y MEMLIN con modelado de probabilidad entre Gaussianas basado en GMMs, MEMLIN MP. Junto al nombre de las diferentes técnicas aparece el número de Gaussianas con que se modelaron los correspondientes espacios, 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 completa la tabla con el WER medio, MWER, y la mejora media, MIMP.

parte, y una vez que se han adaptados los vectores acústicos degradados con los diferentes algoritmos, se aplicará el método CMN. Para toda esta experimentación se utilizó la parametrización estándar ETSI y modelos acústicos fonéticos, de modo que los resultados de referencia se pueden consultar en la Sección 4.4. Nótese igualmente que todos los parámetros que definen los experimentos coinciden con los aplicados en las Secciones 5.5 y 6.6, de manera que las tasas obtenidas son totalmente comparables. Asimismo la Figura 5.5 sigue siendo válida para explicar los diferentes pasos precisados para llevar a cabo la experimentación.

7.5.1 Resultados obtenidos con la técnica MEMLIN MP.

En la Tabla 7.2 se pueden apreciar los mejores resultados de RAH para la técnica de adaptación de vectores de características MEMLIN MP; asimismo, y aunque ya fueron introducidos en la Sección 5.5, se exponen a modo de comparación las tasas correspondientes para el método MEMLIN. En ambos casos, junto al nombre de la técnica, MEMLIN y MEMLIN MP, se incluye el número de componentes que conforman las distintas GMMs en cada caso: el primer valor, 128, se corresponde con el número de Gaussianas 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 componentes, cuyos resultados completos se pueden consultar en los Anexos 5.7 y 7.7. 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, (aunque se realizó un barrido con 1, 2 y 4 componentes, por cuestiones de coste computacional, se decidió emplear únicamente 2, desechando el resto). 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 componentes empleadas para modelar el espacio limpio será el mismo que el utilizado para representar cada entorno básico ruidoso. Asimismo se incluye en la Tabla 7.2, además del WER medio, MWER, la mejora media de WER, MIMP, en tanto por ciento, calculada a partir de la expresión (5.29).

Una vez presentados los resultados, resulta conveniente analizar mediante la prueba de hipótesis estadística *z-test* si el comportamiento de la técnica MEMLIN MP es estadísticamente diferente al del algoritmo MEMLIN para la base de datos *SpeechDat*

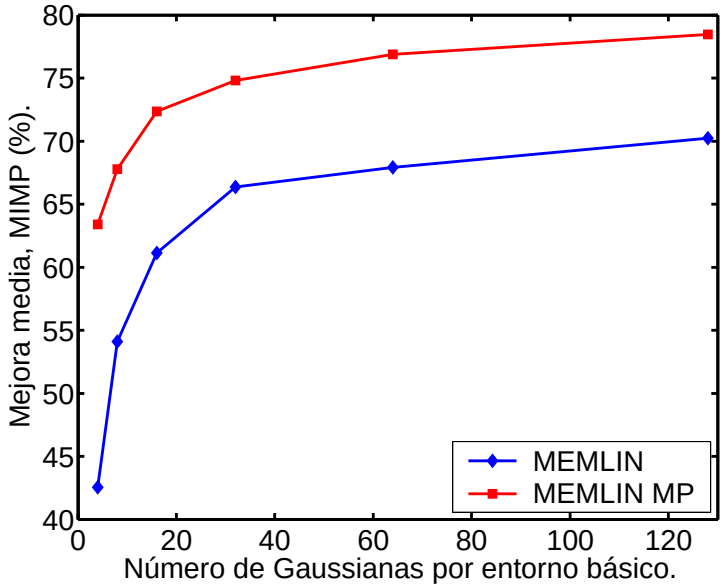


Figura 7.2: Mejora media del WER, MIMP, para las técnicas MEMLIN y MEMLIN con modelado de probabilidad entre Gaussianas basado en GMMs (MEMLIN MP), atendiendo al número de componentes con que se modela cada entorno básico. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia.

Car en español. De este modo, el valor del estadístico W, w, es $w = 2{,}8 > 1{,}96$, por lo que la mejora lograda en este caso sí se puede considerar independiente de la base de datos con un intervalo de confianza del 95 % con respecto a la alcanzada con el algoritmo MEMLIN. A pesar de ello, 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 4.3.

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 únicamente mediante dos componentes para cada s_x s_y^e. Se puede apreciar como el empleo del nuevo modelado de la probabilidad entre Gaussianas propuesto produce en todos los casos una importante mejora en términos de RAH, aunque ésta es sensiblemente mayor cuando el número de componentes es reducido; así, al representar los entornos básicos con 4 Gaussianas, la mejora media se incrementa desde 42.56 % hasta 63.39 %, mientras que si se modelan con 128 componentes, se alcanza el 78.48 % (70.22 % para la técnica MEMLIN).

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, MEMLIN MP, 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 relevante cuando el número de componentes con que se modela el espacio limpio y los entornos básicos es reducido, tal y como queda patente en la Figura 7.2.

Sin embargo, y aunque el comportamiento proporcionado por la técnica MEMLIN MP es claramente satisfactorio para cualquier número de componentes con que se representen los entornos básicos, el coste computacional en este caso es mucho mayor que el precisado para el algoritmo MEMLIN. Esto es debido a que se debe evaluar un número más elevado de *scores* de Gaussianas por entorno básico y vector de características ruidoso en el proceso de adaptación, n_G, lo que es, a la postre, el término más gravoso computacionalmente hablando del proceso de adaptación. De hecho, para el método MEMLIN MP n_G será

$$n_G = n_{s_y^e}(1 + n_{s_x} \times n_{{s'_y}^e}), \tag{7.11}$$

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 modelan los vectores de características ruidosos asociados a cada par de Gaussianas s_x y s_y^e. Obsérvese 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 componentes, $n'_{s_y^e}$, de los modelos de cada entorno ruidoso básico con mayor *score* haciendo uso de las expresiones (5.19) y (5.20). Posteriormente se elige para cada una de ellas, las componentes 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

$$n_G = n_{s_y^e} + n'_{s_y^e} \times n'_{s_x} \times n_{{s'_y}^e}, \tag{7.12}$$

En la Tabla 7.3 se muestran los resultados de RAH 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.

7.5.2 Resultados obtenidos con la técnica PD-MEMLIN MP.

A continuación se comparan los resultados obtenidos con las técnicas PD-MEMLIN y PD-MEMLIN MP. Para realizar la correspondiente adaptación se entrenaron y emplearon transformaciones para los 25 fonemas españoles más el silencio.

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

Tabla 7.3: Resultados medios obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%), MWER, utilizando la técnica de adaptación de vectores de características MEMLIN MP cuando se reduce el número de Gaussianas evaluadas en el proceso de normalización ($n'_{s_y^e}$ y n'_{s_x}). Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica MEMLIN MP. 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. Se incluye igualmente la mejora media, MIMP.

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

Tabla 7.4: Mejores resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando distintas técnicas de adaptación de vectores de características. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con las técnicas PD-MEMLIN y PD-MEMLIN con modelado de probabilidad entre Gaussianas basado en GMMs, PD-MEMLIN MP. Junto al nombre de las diferentes técnicas aparece el número de Gaussianas con que se modelaron los fonemas de los correspondientes espacios, incluyendo además para el caso del método PD-MEMLIN MP el número de componentes de las GMMs que constituyeron el modelado de la probabilidad entre Gaussianas. Se incluye 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 6.6) y PD-MEMLIN MP. En ambos casos, junto al nombre de la técnica se incluye el número de componentes que conforman las diferentes GMMs necesarias en cada caso: el primer valor, 16 en ambos casos, se corresponde con el número de Gaussianas empleadas para modelar cada fonema de los espacios limpio y de los entornos básicos ruidosos (se realizó un barrido con 2, 4, 8, 16 y 32 componentes, cuyos resultado completos se pueden consultar en los Anexos 6.11 y 7.7). 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 y entorno básico, 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, calculadas a partir de la expresión (5.29).

A la luz de los resultados presentados en la Tabla 7.4 se puede asegurar que la técnica PD-MEMLIN MP proporciona unas tasas, al menos para la mejor combinación de

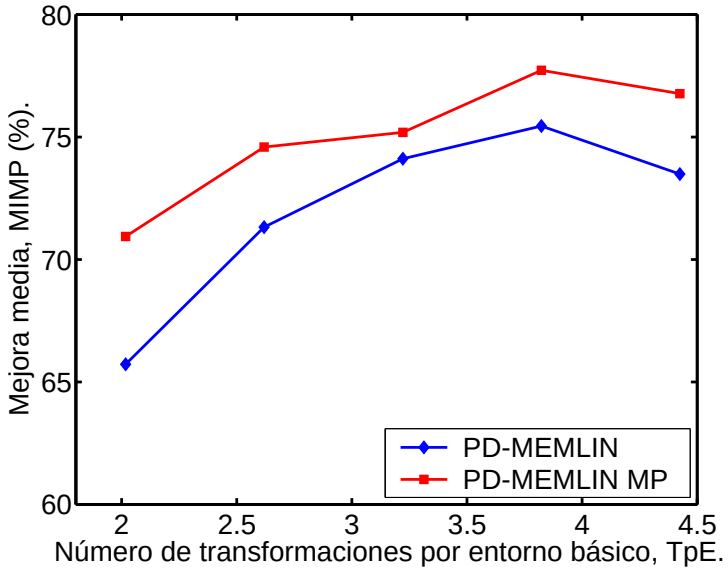


Figura 7.3: Mejora media del WER, MIMP, para las técnicas PD-MEMLIN y PD-MEMLIN con modelado de probabilidad entre Gaussianas basado en GMMs (PD-MEMLIN MP), atendiendo al número de transformaciones por entorno básico, TpE en $\log_{10}$. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia.

número de Gaussianas tratada, superiores a las alcanzadas por el algoritmo PD-MEMLIN.

Por otra parte, de cara a analizar 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 se hace uso de la prueba de hipótesis estadística *z-test*. Así, se calcula el valor del estadístico W, $w = 0{,}8 < 1{,}96$, 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 respecto a la técnica PD-MEMLIN con un intervalo de confianza del 95 %. Sin embargo, si se realiza la comparación entre los algoritmos PD-MEMLIN MP y MEMLIN, el valor del estadístico pasa a ser $w = 2{,}53 > 1{,}96$, de modo que en este caso sí hay una diferencia de comportamiento estadísticamente significativa 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 convenientemente en la Sección 4.3.

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 ella 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 compone de 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 condicionada entre espacios de señal; así, cuando se representa cada fonema

con 2 componentes, se incrementa la mejora media desde 65.72 % hasta 70.94 %, mientras que si los fonemas se modelan 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 nuevo modelado de la probabilidad condicionada entre espacios de señal es notoria para cualquier número de transformaciones por entorno básico, TpE, el coste computacional es, como sucedía con el método MEMLIN MP, y por la misma causa, mucho mayor, siendo en este caso n_G

$$n_G = n_{ph} \times n_{s_y^{e,ph}}(1 + n_{s_x^{ph}} \times n_{s_y'^{\,e,ph}}), \tag{7.13}$$

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 características 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, y haciendo uso de las mismas expresiones, se eligen las $n'_{s_y^{e,ph}}$ componentes 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, aquéllas del modelo limpio más probables, $n'_{s_x^{ph}}$, empleando (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

$$n_G = n_{ph} \times n_{s_y^{e,ph}} + n'_{ph} \times n'_{s_y^{e,ph}} \times n'_{s_x^{ph}} \times n_{s_y'^{\,e}}, \tag{7.14}$$

En la Tabla 7.5 se muestran los resultados para la técnica PD-MEMLIN PM con distintos valores de n'_{ph}, $n'_{s_y^{e,ph}}$ y $n'_{s_x^{ph}}$, siendo en todos los casos $n_{s_y'^{\,e,ph}} = 2$. 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 ruidosos. Se puede observar como, si bien la disminución del número de componentes calculadas reduce ligeramente las prestaciones del método, los resultados obtenidos siguen siendo altamente competitivos a

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

Tabla 7.5: Resultados medios obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%), MWER, utilizando la técnica de adaptación de vectores de características PD-MEMLIN MP 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 ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica PD-MEMLIN MP. Junto al nombre del método aparece el número de Gaussianas con que se modelaron los correspondientes fonemas para los distintos espacios (el limpio y los asociados a los entornos básicos ruidosos). El número de componentes de las GMMs que constituyen el modelado de la probabilidad entre Gaussianas es, en todos los casos, 2. Se incluye igualmente la mejora media, MIMP.

la vez que se reduce el coste computacional hasta en un factor 4.38.

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 primero. Adicionalmente, y por completar la comparación, se ha incluido la Figura 7.4, en la que se muestran los histogramas y *log-scattergrams* del primer coeficiente MFCC de los vectores de características de voz limpios y los correspondientes adaptados mediante los métodos MEMLIN MP (Figura 7.4.b) y PD-MEMLIN MP (Figura 7.4.c). En todos los casos se ha hecho uso del entorno básico E4 del corpus de reconocimiento de la base de datos *SpeechDat Car* en español. Para el primer caso se emplean 128 Gaussianas por entorno básico, mientras que para la técnica PD-MEMLIN MP se modelan los distintos fonemas con 16 componentes por entorno básico (las combinaciones que mejores resultados en términos de WER han proporcionado en cada caso). Se puede observar como, a nivel visual, no hay grandes diferencias con las representaciones ya incluidas para los métodos MEMLIN (Figura 5.7) o PD-MEMLIN (Figura 6.6), respectivamente. Así pues, teniendo en cuenta todos los resultados presentados en esta Sección, se puede concluir que el nuevo modelado de probabilidad condicionada entre espacios de señal 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 ello suponga una seria reducción de las tasas de reconocimiento.

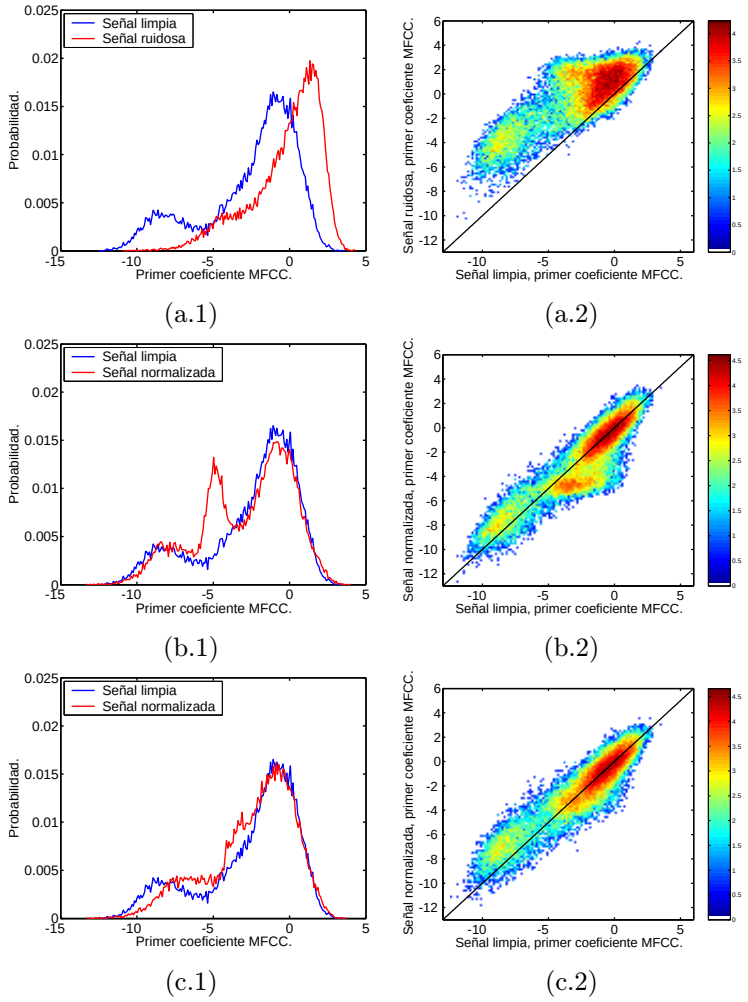


Figura 7.4: *Log-scattergrams* e histogramas del primer coeficiente MFCC de los vectores de características de voz limpia y ruidosa (a), o limpia y normalizada usando la técnica MEMLIN MP con 128 Gaussianas por entorno básico (b). En la figura (c) se representa el *log-scattergram* y el histograma obtenidos tras aplicar el método 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 entrenadas con 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-scattergrams* representa la función $x = y$.

7.6 Anexo H.

En este Anexo se incluye el desarrollo teórico necesario para estimar los parámetros que definen el nuevo modelado de la probabilidad entre Gaussianas basado en GMMs considerado en el Capítulo 7. De cara a simplificar la notación se considerará únicamente un entorno básico y un fonema, de modo que se eliminarán dichas dependencias, lo que, por otra parte, no resta generalidad alguna puesto que cada entorno y fonema se pueden tratar de modo independiente.

Sea pues la GMM que se desea estimar compuesta por C''' componentes que modela los vectores de características ruidosos asociados al par de Gaussianas de los espacios limpio y degradado s_x y s_y, (7.3) y (7.4). De cara a obtener las estimaciones de los parámetros asociados a cada componente, s'_y, que define la GMM, esto es: los vectores de medias, μ_{s_x,s_y,s'_y}, las matrices diagonales de covarianzas, $\mathbf{\Sigma}_{s_x,s_y,s'_y}$, y las probabilidades a priori, $p(s'_y|s_x,s_y)$, se hace uso del criterio ML, para lo cual es preciso aplicar el algoritmo EM [DLR77]. Por ello se define primeramente una función de verosimitud a partir de los parámetros dadas las observaciones, que en este caso se corresponden con los vectores de características ruidosos (paso E), y posteriormente se maximiza dicha función con respecto a cada uno de los tres parámetros que definen la GMM (paso M).

Sea $(\mathbf{X},\mathbf{Y}) = \{(\mathbf{x}_1,\mathbf{y}_1),...(\mathbf{x}_t,\mathbf{y}_t)...,(\mathbf{x}_T,\mathbf{y}_T)\}$ con $t \in [1,T]$ un corpus de entrenamiento compuesto por señal estéreo. Adicionalmente se asume 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, nombradas como s_x, (5.7) y (5.8). 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 ($\mathbf{W} = \{\mathbf{w}_1,...,\mathbf{w}_T\}$), y $\mathbf{z}_{y,t} \in \{0,1\}^{C'''}$, que será el segundo vector indicador y 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 acústico limpio, $\mathbf{x}_t$, se puede ver igualmente 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 componente s'_x del modelo de probabilidad entre Gaussianas que genera $\mathbf{x}_t$ y ceros en el resto de las C'' posiciones ($\mathbf{Z}_x = \{\mathbf{z}_{x,1},...,\mathbf{z}_{x,T}\}$). Con todo lo anterior, la pdf de los datos completos es

$$\begin{aligned} p(\mathbf{x},\mathbf{y},\mathbf{v},\mathbf{w},\mathbf{z}_x,\mathbf{z}_y) &\simeq p(\mathbf{v},\mathbf{w},\mathbf{z}_x)p(\mathbf{x}|\mathbf{v},\mathbf{w},\mathbf{z}_x) \\ &\times p(\mathbf{v},\mathbf{w},\mathbf{z}_y)p(\mathbf{y}|\mathbf{v},\mathbf{w},\mathbf{z}_y), \end{aligned} \tag{H.1}$$

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 (H.1) se puede expresar como (H.2), 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.

$$\begin{gathered} p(\mathbf{x},\mathbf{y},\mathbf{v},\mathbf{w},\mathbf{z}_x,\mathbf{z}_y) \simeq \\ \prod_{s_x}\prod_{s_y}\prod_{s'_x}\left(p(v_{s_x}=1,w_{s_y}=1,z_{x,s'_x}=1)p(\mathbf{x}|v_{s_x}=1,w_{s_y}=1,z_{x,s'_x}=1)\right)^{v_{s_x}w_{s_y}z_{x,s'_x}} \\ \times\prod_{s_x}\prod_{s_y}\prod_{s'_y}\left(p(v_{s_x}=1,w_{s_y}=1,z_{y,s'_y}=1)p(\mathbf{y}|v_{s_x}=1,w_{s_y}=1,z_{y,s'_y}=1)\right)^{v_{s_x}w_{s_y}z_{y,s'_y}}. \end{gathered} \tag{H.2}$$

7.6.1 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, 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 en la variable $\mathbf{\Theta}$ se incluyen todos los parámetros que se pretenden estimar ($p(\mathbf{y}_t|s_x,s_y,s'_y)$, μ_{s_x,s_y,s'_y} y $\mathbf{\Sigma}_{s_x,s_y,s'_y}$).

$$\begin{aligned} \mathcal{L}(\mathbf{\Theta}|\mathbf{X},\mathbf{Y},\mathbf{V},\mathbf{W},\mathbf{Z}_x,\mathbf{Z}_y) &= \sum_t log\left(p(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t,\mathbf{w}_t,\mathbf{z}_{x,t},\mathbf{z}_{y,t}|\mathbf{\Theta})\right) \\ &= \sum_t\sum_{s_x}\sum_{s_y}\sum_{s'_x} v_{s_x}w_{s_y}z_{x,s'_x}\Big(log\left(p(v_{s_x}=1,w_{s_y}=1,z_{x,s'_x}=1)\right) \\ &+ log(p(\mathbf{x}_t|v_{s_x}=1,w_{s_y}=1,z_{x,s'_x}=1))\Big) \\ &+ \sum_t\sum_{s_x}\sum_{s_y}\sum_{s'_y} v_{s_x}w_{s_y}z_{y,s'_y}\Big(log(p(v_{s_x}=1,w_{s_y}=1,z_{y,s'_y}=1)) \\ &+ log(p(\mathbf{y}_t|v_{s_x}=1,w_{s_y}=1,z_{y,s'_y}=1))\Big), \end{aligned} \tag{H.3}$$

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

$$\begin{aligned} p(v_{s_x}=1,w_{s_y}=1,z_{x,s'_x}=1) &\simeq p(v_{s_x}=1)p(w_{s_y}=1) \\ &\times p(z_{x,s'_x}=1|v_{s_x}=1,w_{s_y}=1) \\ &= P_{s_x}P_{s_y}P_{s_xs_ys'_x}, \end{aligned} \tag{H.4}$$

$$\begin{aligned} p(v_{s_x}=1,w_{s_y}=1,z_{y,s'_y}=1) &\simeq p(v_{s_x}=1)p(w_{s_y}=1) \\ &\times p(z_{y,s'_y}=1|v_{s_x}=1,w_{s_y}=1) \\ &= P_{s_x}P_{s_y}P_{s_xs_ys'_y}, \end{aligned} \tag{H.5}$$

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 ($P_{s_x s_y s'_y} = p(s'_y|s_x, s_y)$). 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 previa o k-ésima, $\mathbf{\Theta}^{(k)}$. 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 $E[\,]$ representa el valor esperado. Considerando esto último se puede observar que

$$\begin{aligned} Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)}) &= \sum_t \sum_{s_x} \sum_{s_y} \sum_{s'_x} (v_{s_x} w_{s_y} z_{x s'_x})^{(k)} \\ &\times \left(log(P_{s_x} P_{s_y} P_{s_x s_y s'_x}) + log(p(\mathbf{x}_t|v_{s_x} = 1, w_{s_y} = 1, z_{x s'_x} = 1)) \right) \\ &+ \sum_t \sum_{s_x} \sum_{s_y} \sum_{s'_y} (v_{s_x} w_{s_y} z_{s'_y})^{(k)} \\ &\times \left(log(P_{s_x} P_{s_y} P_{s_x s_y s'_y}) + log(p(\mathbf{y}_t|v_{s_x} = 1, w_{s_y} = 1, z_{y s'_y} = 1)) \right). \quad \text{(H.6)} \end{aligned}$$

$$\begin{aligned} (v_{s_x} w_{s_y} z_{x s'_x})^{(k)} &= E\left[v_{s_x} w_{s_y} z_{x s'_x} | \mathbf{x}_t, \mathbf{y}_t, \mathbf{\Theta}^{(k)} \right] \\ &\simeq E[v_{s_x}|\mathbf{x}_t] E[w_{s_y}|\mathbf{y}_t] E\left[z_{x s'_x} | \mathbf{x}_t, v_{s_x}, w_{s_y}, \mathbf{\Theta}^{(k)} \right] = A B^{(k)}, \quad \text{(H.7)} \end{aligned}$$

$$\begin{aligned} (v_{s_x} w_{s_y} z_{y s'_y})^{(k)} &= E\left[v_{s_x} w_{s_y} z_{y s'_y} | \mathbf{x}_t, \mathbf{y}_t, \mathbf{\Theta}^{(k)} \right] \\ &\simeq E[v_{s_x}|\mathbf{x}_t] E[w_{s_y}|\mathbf{y}_t] E\left[z_{y s'_y} | \mathbf{y}_t, v_{s_x}, w_{s_y}, \mathbf{\Theta}^{(k)} \right] = A C^{(k)}, \quad \text{(H.8)} \end{aligned}$$

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 (H.9). 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 a como se ha calculado (H.9); sin embargo, en este trabajo a la hora de estimar estas dos últimas 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 este desarrollo teórico propuesto resulta intrascendente.

$$E\left[z_{s'_y} | \mathbf{y}_t, v_{s_x}, w_{s_y}, \mathbf{\Theta}^{(k)} \right] = \frac{p(s'_y|s_x, s_y)^{(k)} \mathcal{N}(\mathbf{y}_t|\mu^{(k)}_{s_x, s_y, s'_y}, \mathbf{\Sigma}^{(k)}_{s_x, s_y, s'_y})}{\sum_{s'_y} p(s'_y|s_x, s_y)^{(k)} \mathcal{N}(\mathbf{y}_t|\mu^{(k)}_{s_x, s_y, s'_y}, \mathbf{\Sigma}^{(k)}_{s_x, s_y, s'_y})}. \quad \text{(H.9)}$$

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

7.6.2.1 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

$$\mathcal{L}(\mathbf{\Theta}, \lambda_{s_x s_y}) = Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)}) - \sum_{s_x}\sum_{s_y}\lambda_{s_x s_y}\sum_{s'_y}\left(P_{s_x s_y s'_y} - 1\right), \tag{H.10}$$

donde $\lambda_{s_x s_y}$ son los correspondientes multiplicadores de Lagrange. De este modo, a la hora de obtener las probabilidades a priori óptimas, $P_{s_x s_y s'_y}$, resultará 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

$$\left.\frac{\delta\mathcal{L}(\mathbf{\Theta}, \lambda_{s_x s_y})}{\delta P_{s_x s_y s'_y}}\right|_{\Theta=\Theta^{k+1}} = \sum_t \frac{(v_{s_x} w_{s_y} z_{y s'_y})^{(k)}}{P^{k+1}_{s_x s_y s'_y}} - \lambda_{s_x s_y} = 0, \tag{H.11}$$

$$P^{(k+1)}_{s_x s_y s'_y} = \frac{1}{\lambda_{s_x s_y}}\sum_t (v_{s_x} w_{s_y} z_{y s'_y})^{(k)}. \tag{H.12}$$

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

$$\left.\frac{\delta\mathcal{L}(\mathbf{\Theta}, \lambda_{s_x s_y})}{\delta \lambda_{s_x s_y}}\right|_{\Theta=\Theta^{k+1}} = -\sum_{s'_y} P^{(k+1)}_{s_x s_y s'_y} + 1 = 0, \tag{H.13}$$

$$\sum_{s'_y} P^{(k+1)}_{s_x s_y s'_y} = 1. \tag{H.14}$$

A partir de (H.12) y (H.14), se da con 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

$$p(s'_y|s_x, s_y)^{(k+1)} = \frac{\sum_t (v_{s_x} w_{s_y} z_{s'_y})^{(k)}}{\sum_t \sum_{s'_y} (v_{s_x} w_{s_y} z_{s'_y})^{(k)}} = \frac{\sum_t AC^{(k)}}{\sum_t \sum_{s'_y} AC^{(k)}}, \tag{H.15}$$

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

7.6.2.2 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 $\mathcal{L}(\mathbf{\Theta}) = Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$

$$\begin{aligned}\left.\frac{\delta \mathcal{Q}(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})}{\delta \mu_{s_x s_y s'_y}}\right|_{\Theta=\Theta^{k+1}} &= \mathbf{0} \\ &= \sum_t (v_{s_x} w_{s_y} z_{y s'_y})^{(k)} \\ &\times \left.\frac{\delta}{\delta \mu_{s_x s_y s'_y}}\right|_{\Theta=\Theta^{k+1}} \left[log(p(\mathbf{y}_t|v_{s_x}=1, w_{s_y}=1, z_{s'_y}=1))\right] \\ &= \sum_t (v_{s_x} w_{s_y} z_{y s'_y})^{(k)} \\ &\times \left.\frac{\delta}{\delta \mu_{s_x s_y s'_y}}\right|_{\Theta=\Theta^{k+1}} \left[-\frac{1}{2}(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}(\mathbf{y}_t - \mu_{s_x s_y s'_y})\right], \end{aligned} \quad \text{(H.16)}$$

donde se ha hecho uso de

$$p(\mathbf{y}_t|v_{s_x}=1, w_{s_y}=1, z_{y s'_y}=1) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_{s_x,s_y,s'_y}|^{1/2}} e^{-\frac{1}{2}(\mathbf{y}_t - \mu_{s_x,s_y,s'_y})^T \mathbf{\Sigma}^{-1}_{s_x,s_y,s'_y}(\mathbf{y}_t - \mu_{s_x,s_y,s'_y})}, \quad \text{(H.17)}$$

siendo 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

$$\begin{aligned}(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}(\mathbf{y}_t - \mu_{s_x s_y s'_y}) &= Tr\left[(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}(\mathbf{y}_t - \mu_{s_x s_y s'_y})\right] \\ &= -Tr\left[\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mu_{s_x s_y s'_y} \mathbf{y}_t{}^T\right] + Tr\left[\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mu_{s_x s_y s'_y} \mu_{s_x s_y s'_y}{}^T\right] \\ &+ Tr\left[\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mathbf{y}_t \mathbf{y}_t{}^T\right] - Tr\left[\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mathbf{y}_t \mu_{s_x s_y s'_y}{}^T\right]. \end{aligned} \quad \text{(H.18)}$$

$$\begin{aligned}&\left.\frac{\delta}{\delta \mu_{s_x s_y s'_y}}\right|_{\Theta=\Theta^{k+1}} \left[-\frac{1}{2}(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}(\mathbf{y}_t - \mu_{s_x s_y s'_y})\right] \\ &= -\frac{1}{2}\left(-\mathbf{\Sigma}^{-1}_{s_x s_y s'_y}\mathbf{y}_t - \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}{}^T \mathbf{y}_t + (\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} + \mathbf{\Sigma}^{-1}_{s_x s_y s'_y}{}^T)\mu_{s_x s_y s'_y}\right) \\ &= -\frac{1}{2}\left(-2\mathbf{\Sigma}^{-1}_{s_x s_y s'_y}\mathbf{y}_t + 2\mathbf{\Sigma}^{-1}_{s_x s_y s'_y}\mu_{s_x s_y s'_y}\right). \end{aligned} \quad \text{(H.19)}$$

Con todo lo anterior, e introduciendo la expresión (H.19) en (H.16), se tiene finalmente que

$$\left.\frac{\delta \mathcal{Q}(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})}{\delta \mu_{s_x s_y s'_y}}\right|_{\Theta=\Theta^{k+1}} = \mathbf{0}$$

$$= \sum_t \left(v_{s_x} w_{s_y} z_{y s'_y}\right)^{(k)} \left(-\mathbf{\Sigma}^{-1}_{s_x s_y s'_y} \mathbf{y}_t + \mathbf{\Sigma}^{-1}_{s_x s_y s'_y} {\mu_{s_x s_y s'_y}}^{(k+1)}\right), \quad \text{(H.20)}$$

$$\mu^{(k+1)}_{s_x, s_y, s'_y} = \frac{\sum_t (v_{s_x} w_{s_y} z_{s'_y})^{(k)} \mathbf{y}_t}{\sum_t (v_{s_x} w_{s_y} z_{s'_y})^{(k)}}. \quad \text{(H.21)}$$

7.6.2.3 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 $\mathcal{L}(\mathbf{\Theta}) = Q(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})$.

$$\left.\frac{\delta \mathcal{Q}(\mathbf{\Theta}|\mathbf{\Theta}^{(k)})}{\delta \mathbf{\Sigma}_{s_x s_y s'_y}}\right|_{\Theta=\Theta^{k+1}} = \mathbf{0}$$

$$= \sum_t \left(v_{s_x} w_{s_y} z_{y s'_y}\right)^{(k)} \times \left.\frac{\delta}{\delta \mathbf{\Sigma}_{s_x s_y s'_y}}\right|_{\Theta=\Theta^{k+1}} \left[log(p(\mathbf{y}_t | v_{s_x} = 1, w_{s_y} = 1, z_{s'_y} = 1))\right]$$

$$= \sum_t \left(v_{s_x} w_{s_y} z_{y s'_y}\right)^{(k)} \times \left.\frac{\delta}{\delta \mathbf{\Sigma}_{s_x s_y s'_y}}\right|_{\Theta=\Theta^{k+1}} \left[-\frac{1}{2}(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T \mathbf{\Sigma}^{-1}_{s_x s_y s'_y} (\mathbf{y}_t - \mu_{s_x s_y s'_y})\right]. \quad \text{(H.22)}$$

Mediante propiedades del cálculo matricial, y teniendo en cuenta que la matriz de covarianza es diagonal, así como la expresión (H.18), se puede observar que

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

$$= -\frac{1}{2}\left[{\mathbf{\Sigma}^{(k+1)}_{s_x s_y s'_y}}^{-1} - {\mathbf{\Sigma}^{(k+1)}_{s_x s_y s'_y}}^{-1} (\mathbf{y}_t - \mu_{s_x s_y s'_y})(\mathbf{y}_t - \mu_{s_x s_y s'_y})^T {\mathbf{\Sigma}^{(k+1)}_{s_x s_y s'_y}}^{-1}\right]. \quad \text{(H.23)}$$

Con todo lo anterior, y llevando la expresión (H.23) a (H.22), se tiene finalmente que

$$\mathbf{\Sigma}^{(k+1)}_{s_x, s_y, s'_y} = \frac{1}{\sum_t (v_{s_x} w_{s_y} z_{s'_y})^{(k)}} \times \sum_t (v_{s_x} w_{s_y} z_{s'_y})^{(k)} \left(\mathbf{y}_t - \mu^{(k)}_{s_x, s_y, s'_y}\right)\left(\mathbf{y}_t - \mu^{(k)}_{s_x, s_y, s'_y}\right)^T. \quad \text{(H.24)}$$

7.7 Anexo I.

En este Anexo se presentan los resultados en términos de WER (%) obtenidos para los diferentes entornos básicos (E1,..., E7) de la base de datos *SpeechDat Car* en español utilizando distintas técnicas de adaptación de vectores de características (MEMLIN y PD-MEMLIN con modelado de probabilidad entre Gaussianas basado en GMMs, identificados como MEMLLIN MP y PD-MEMLIN MP). Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados con señal limpia. A su vez, y junto al nombre de las distintas técnicas, se ha incluido bien el número de Gaussianas con que se modelan los correspondientes entornos básicos (4, 8, 16, 32, 64 y 128 Gaussianas), Tabla 7.6, o bien el número de componentes con que se representan los distintos fonemas para cada entorno básico (2, 4, 8, 16 y 32), Table 7.7. Para ambos casos, se emplean 2 componentes para modelar los vectores de características ruidosos asociados a cada par de Gaussianas.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF MEMLIN MP 4	2.88	8.06	5.59	6.52	9.25	8.10	12.59	7.04	63.40
CLK	HF MEMLIN MP 8	2.49	7.98	4.90	6.77	8.20	6.83	9.52	6.41	67.77
CLK	HF MEMLIN MP 16	2.21	7.55	4.48	6.39	7.63	5.24	6.80	5.75	72.37
CLK	HF MEMLIN MP 32	2.11	7.12	4.34	6.14	7.15	4.92	5.44	5.39	74.81
CLK	HF MEMLIN MP 64	2.11	6.69	4.34	6.02	6.96	4.29	3.74	5.10	76.87
CLK	HF MEMLIN MP 128	2.01	6.43	3.92	5.76	6.48	4.13	4.42	4.86	78.47

Tabla 7.6: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características MEMLIN con modelado de probabilidad entre Gaussianas basado en GMMs, MEMLIN MP. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica MEMLIN MP. Junto al nombre de la técnica aparece el número de Gaussianas con que se modeló cada entorno básico ruidoso. Adicionalmente se emplean 2 componentes para modelar los vectores de características ruidosos asociados a cada par de Gaussianas: s_x y s_y^e. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF PD-MEMLIN MP 2	2.78	8.32	5.31	5.89	7.63	5.24	5.10	5.95	70.94
CLK	HF PD-MEMLIN MP 4	2.49	7.55	5.59	5.51	6.67	3.81	5.78	5.43	74.59
CLK	HF PD-MEMLIN MP 8	2.68	7.80	5.87	5.14	6.01	4.13	4.42	5.34	75.19
CLK	HF PD-MEMLIN MP 16	1.92	7.46	5.31	5.14	5.82	3.81	4.08	4.97	77.72
CLK	HF PD-MEMLIN MP 32	2.01	7.63	4.62	5.64	6.58	3.49	4.08	5.11	76.76

Tabla 7.7: Resultados con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características PD-MEMLIN con modelado de probabilidad entre Gaussianas basado en GMMs, PD-MEMLIN MP. Se ha empleado la parametrización estándar ETSI y modelos acústicos fonéticos generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica PD-MEMLIN MP. Junto al nombre de la técnica aparece el número de Gaussianas con que se modelaron los fonemas para cada entorno básico ruidoso. Adicionalmente se emplean 2 componentes para modelar los vectores de características ruidosos asociados a cada par de Gaussianas: s_x^{ph} y $s_y^{e,ph}$. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Capítulo 8

Adaptación Conjunta de Señal y Modelos Acústicos.

8.1 Introducción.

En el Capítulo 3 se presentó una taxonomía formal de los métodos más empleados para dotar de robustez a los sistemas de RAH. En ella se distinguía principalmente entre dos grandes líneas de actuación: las técnicas de adaptación de señal hacia los modelos acústicos y los algoritmos de adaptación de modelos acústicos hacia la señal. En el primero de los casos se proyectan los vectores acústicos del 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 consideró en dicho Capítulo las ventajas e inconvenientes que ambas filosofías poseen, siendo en algunos casos complementarias. Basándose en este hecho, así como en la incapacidad de que las técnicas de adaptación 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 soluciones híbridas [NY94] [SL96].

Si se hace un repaso a los métodos presentados hasta el momento en este trabajo y se analizan aquéllos cuyos comportamientos han resultados más satisfactorios, se puede comprobar que en todos los casos (PD-MEMLIN, MEMLIN MP y PD-MEMLIN MP), se propone una transformación lineal con término de pendiente 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 *log-scattergrams* presentados en capítulos anteriores, no hace lo propio con otros efectos, como por ejemplo las rotaciones.

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. Así, a partir del método de normalización seleccionado, que en general puede ser cualquiera de los presentados previamente en este trabajo, se pretende compensar el desplazamiento de los vectores de características, mientras que, por otra parte, con la correspondiente técnica de adaptación

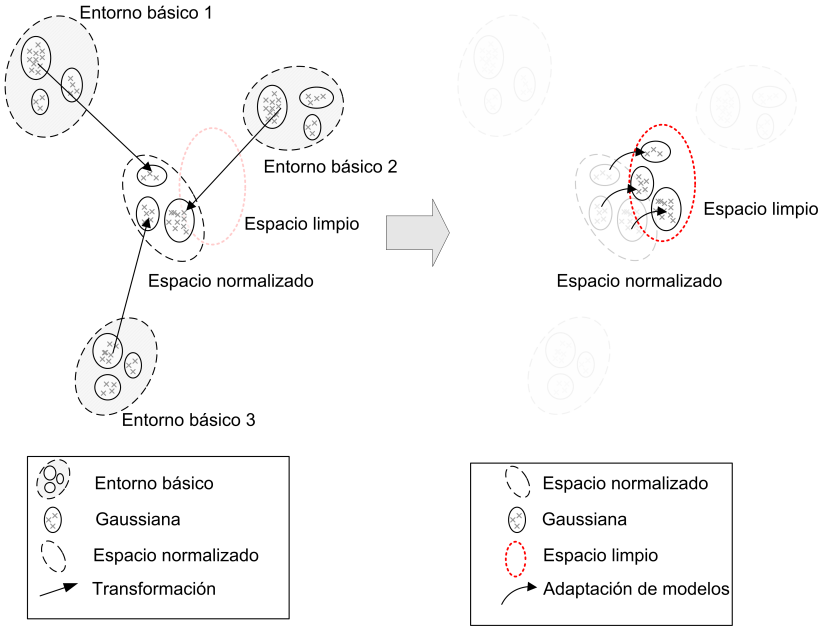


Figura 8.1: Representación gráfica de las técnicas híbridas propuestas en este Capítulo. La parte izquierda está dedicada a la adaptación de los vectores de características, cuya misión es proyectar los ruidosos desde un determinado entorno básico a un espacio normalizado que, por las limitaciones del 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.

de modelos acústicos se busca reducir de un modo estadístico aquellos desajustes de los vectores de características que el método de normalización no ha considerado previamente. Esta segunda fase, al igual que la primera, 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 de 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 proyectándolo sobre el 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 se obtienen o no de modo supervisado. Así, para la segunda opción, tras unos estudios previos [MBL$^+$07], se desarrollaron diversas técnicas basadas en matrices de rotación dependientes de GMMs (técnicas híbridas basadas en la estimación no supervisada de matrices de rotación dependientes

de GMMs) [BML+07b] [BML+07c]. 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 los nuevos modelos acústicos asociados al espacio normalizado tras compensar convenientemente el corpus de entrenamiento ruidoso de que se disponga [BML+07a].

Este Capítulo se estructura del siguiente modo: en la Sección 8.2 se presenta la filosofía y las bases teóricas que subyacen en las distintas técnicas híbridas basadas en la estimación no supervisada de matrices de rotación dependientes de GMMs. Por su parte, las técnicas híbridas consistentes en reentrenamiento supervisado se introducen en la Sección 8.3.Finalmente los resultados obtenidos con los distintos algoritmos con la base de datos *SpeechDat Car* en español se incluyen en la Sección 8.4.

8.2 Técnicas Híbridas Basadas en la Estimación no Supervisada de 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 adaptar los modelos acústicos del espacio de referencia tras considerar un modelado del espacio de señal semejante a los planteados hasta el momento para las distintas técnicas normalización ya introducidas. De hecho, y bajo ciertas circunstancias, algunos métodos de normalización de vectores de características se pueden ver como algoritmos de adaptación de modelos acústicos, como la técnica MEMLIN. En ella, el vector compensado, $\hat{\mathbf{x}}_t = \mathbf{y}_t - \mathbf{u}_t$, donde $\mathbf{u}_t$ es el vector de desplazamiento final para el instante de tiempo t, se decodifica haciendo uso de los modelos acústicos asociados al espacio limpio. Sin embargo, este procedimiento es matemáticamente idéntico a reconocer el vector ruidoso, $\mathbf{y}_t$, con modelos acústicos adaptados, entendiendo por adaptación en este caso a la modificación de todos los vectores de medias de los diferentes estados de los HMMs que componen el modelado acústico, 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 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 6.2, 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, $\mathbf{\Sigma}_{y,t}$, de los modelos acústicos asociados a los vectores de características ruidosos en el instante de tiempo t son, respectivamente y con respecto a los consiguientes parámetros

de los modelos acústicos limpios, μ_x y $\mathbf{\Sigma}_x$: $\mu_{y,t} = \mathbf{A}_t\mu_x + \mathbf{g}_t$ 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 considerada previamente, 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 $|\mathbf{\Sigma}_x|$ 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 mediante únicamente 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 en lo sucesivo como "basadas en matrices de rotación".

Obsérvese 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 métodos de normalización de vectores de características como los planteados en este trabajo, se ha incluido un bloque de entrenamiento para los mismos, "Entrenamiento normalización", 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 mediante el vector de características normalizado, esto último ya en la fase de decodificación, "Decodificador". El bloque identificado como "Normalización" incluye la técnica de compensación de vectores de características empleada en el método híbrido.

8.2.1 Técnicas híbridas a partir del cálculo de matrices de rotación dependientes de GMMs.

En esta Sección, 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 adaptación 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.

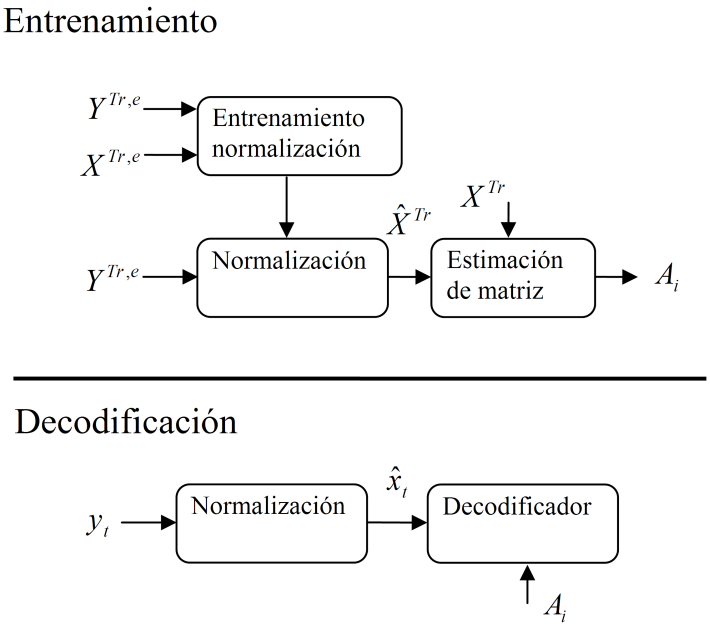


Figura 8.2: Esquema gráfico de las técnicas híbridas basadas en matrices de rotación. Se incluye tanto la fase de entrenamiento como la de decodificación. La primera de ellas está compuesta por tres bloques. El primero de ellos, "Entrenamiento normalización", se ha incluido en previsión de utilizar técnicas de adaptación de vectores de características que precisen de una fase previa de entrenamiento. Por su parte, el sistema de "Normalización" proporciona la estimación de los vectores acústicos limpios a partir de los correspondientes degradados. Finalmente el bloque "Estimación de matriz" calcula un conjunto de matrices de rotación, una de las cuales será seleccionada por cada vector de características normalizado en la fase de decodificación a partir del bloque "Decodificador".

Siguiendo un esquema similar al utilizado para múltiples de las técnicas de adaptación de vectores de características presentadas en este trabajo, 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 un única mezcla de Gaussianas

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

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

  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 de medias, 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 la siguiente función lineal multiplicativa del vector

de características limpio: $\hat{\mathbf{x}}_t \approx \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t$, 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 algún modo este término hace las veces de modelado del espacio de señal ya considerado en técnicas anteriores.

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 diferentes 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)

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

$$\begin{aligned}\mathbf{A}_{s_x,s_{\hat{x}}} &= \underset{\mathbf{A}_{s_x,s_{\hat{x}}}}{arg\,min}(\xi_{s_x,s_{\hat{x}}}) \\ &= \sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})(\hat{\mathbf{x}}_t^{Tr}(\mathbf{x}_t^{Tr})^T)\left(\sum_t p(s_x|\mathbf{x}_t^{Tr})p(s_{\hat{x}}|\hat{\mathbf{x}}_t^{Tr})(\mathbf{x}_t^{Tr}(\mathbf{x}_t^{Tr})^T)\right)^{-1}\end{aligned} \tag{8.4}$$

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 de ellos (8.6). En el Anexo 8.5 en este mismo Capítulo se encuentra 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 expandidos de modo similar a como se realiza para el método *augMented stAte space acousTic modEl*, MATE, [MLR+05] [MLJ+06], en el que se busca una serie de modelos acústicos que sean capaces de representar la variabilidad inter-locutor basándose en el algoritmo *Vocal Tract Length Normalization*, VTLN, [LR98]. Posteriormente, y una

vez obtenidos los modelos expandidos, se evalua el proceso de decodificación a partir del algoritmo de Viterbi generalizado de manera similar a las presentadas en los trabajos [VM90] [GY92].

Considerando que los modelos acústicos del espacio de 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 transformando los de la pdf del estado original q mediante 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 expandido (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 se mantienen 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), aunque pueden estimarse a partir del algoritmo EM [MLJ+06], en los experimentos que se desarrollarán en este trabajo se considerarán independientes de la matriz de rotación. De este modo, la correspondiente probabilidad de transición del estado (q',n') al (q,n) quedará como $\pi_{q',n',q,n} = p(q,n|q',n') = p(q|q')/N$. A pesar de todo ello, 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.

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

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 generalizar el algoritmo de decodificación. De cara a 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 normalizado $\hat{\mathbf{x}}_t$, y que constituye la base para ejecutar el algoritmo de Viterbi, se calculará como

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

donde $p(hat\mathbf{x}_t|n,q)$ es el *score* del vector de características $\hat{\mathbf{x}}_t$, dado el estado expandido (q,n), $(p(\hat{\mathbf{x}}_t|n,q) = \sum_{s_{q,n}} \mathcal{N}(\hat{\mathbf{x}}_t; \mathbf{A}_n\mu_{s_q}, \mathbf{A}_n\mathbf{\Sigma}_{s_q}\mathbf{A}_n^T)p(s_{q,n}))$. Se puede observar como la expresión recursiva anterior es similar a la expuesta en [MLR$^+$05], 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 [PN05], hecho este que no se hubiera producido si únicamente se normalizaran los vectores de características, generando por tanto un cierto desajuste

Adviértase que las técnicas híbridas a partir del cálculo no supervisado de matrices de rotación dependientes de GMMs poseen la importante ventaja de que son no supervisadas e independientes 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... Ya para finalizar, es importante también resaltar como el método de normalización previo necesario puede ser genérico, ya que el algoritmo final es independiente de él, aunque en la experimentación propuesta para este trabajo se han empleado únicamente los algoritmos MEMLIN y MEMLIN MP, dando lugar a sendas técnicas híbridas.

8.3 Técnicas Híbridas Basadas en Reentrenamiento Supervisado.

A la hora de conjugar las técnicas de adaptación 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, otra solución, quizás la más directa pero también la que más recursos precisa, puede aplicarse. En este caso se decodifican los vectores acústicos ruidosos compensados mediante modelos acústicos representantes del espacio normalizado 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 no es perfecta, de modo que a pesar de que se traten de proyectar los vectores de características 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 mediante los modelos acústicos de referencia. Para evitar este desajuste se propone reentrenar de modo supervisado unos nuevos modelos acústicos que representen convenientemente al espacio normalizado, haciendo uso en este caso de la señal del corpus de entrenamiento ruidoso previamente compensada. La diferencia de esta nueva solución con respecto a la mostrada en la Sección precedente consiste en que los nuevos modelos acústicos se obtienen de forma supervisada haciendo uso de las técnicas clásicas de entrenamiento: criterio ML, si el corpus del que se dispone es suficientemente amplio y la aplicación concreta lo permite, o bien técnicas de adaptación de modelos acústicos como MAP, MLLR..., en caso contrario. En este trabajo se decidió emplear la primera opción.

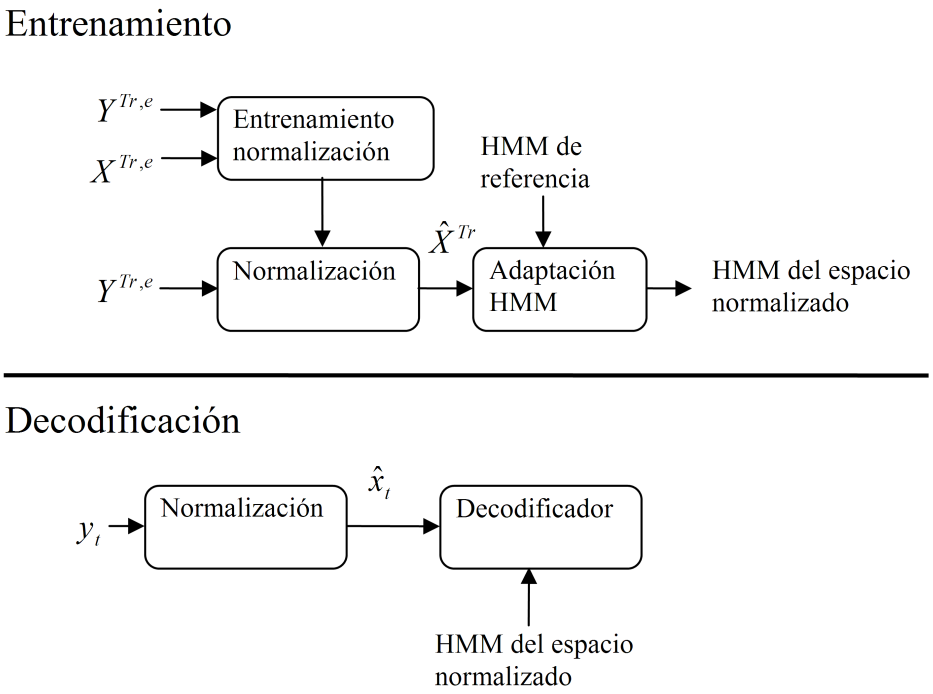


Figura 8.3: Esquema gráfico de las técnicas híbridas basadas en reentrenamiento supervisado. Se incluye tanto la fase de entrenamiento como la de decodificación. La primera de ellas está compuesta por tres bloques. El primero de ellos, "Entrenamiento normalización", se ha incluido en previsión de utilizar técnicas de normalización de vectores de características que precisen una fase previa de entrenamiento. Por su parte, el sistema de "Normalización" proporciona la estimación de los vectores de características limpios a partir de los correspondientes degradados. 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 previamente compensada. Dichos modelos son los empleados para reconocer los vectores de características normalizados en el bloque de "Decodificación".

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 basadas en reentrenamiento supervisado. Del mismo modo que en la Sección 8.2, y en previsión de emplear técnicas de adaptación de vectores de características que precisen de una fase previa de entrenamiento, se ha incluido nuevamente un bloque para la misma, "Entrenamiento normalización". 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, "Decodificador". El bloque identificado como "Normalización" incluye la técnica de compensación de vectores de características empleada en el método híbrido.

Al igual que en la sección anterior, el esquema propuesto para las técnicas híbridas basadas en reentrenamiento supervisado es independiente del método de adaptación de vectores de características aplicado, aunque en este trabajo se han empleado únicamente los algoritmos MEMLIN y MEMLIN MP.

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

La experimentación comparativa de las técnicas híbridas tratadas en las Secciones 8.2 y 8.3 se llevó a cabo con la base de datos *SpeechDat Car* en español, utilizándose los corpora de entrenamiento de los diversos entornos básicos para realizar los distintos procesos de entrenamiento, necesarios en esta ocasión tanto para las técnicas de adaptación de vectores de características seleccionadas, como para generar los nuevos modelos acústicos. Por otra parte, en todas las experimentaciones presentadas en esta Sección se aplicará en última instancia el método CMN a los vectores acústicos que se pretenda reconocer, teniendo 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 consultar, de este modo, 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 adaptación de vectores de características se aplicarán únicamente sobre los coeficientes estáticos de los mismos, tal y como hasta ahora se ha venido realizando, de modo que las correspondientes derivadas se calculan posteriormente.

8.4.1 Resultados obtenidos con 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 la 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 5.4 y 7.4, se hace uso de los vectores de características 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 y 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 8.2, 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 emplear la técnica CMN. 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 o MEMLIN MP haciendo uso de los modelos expandidos construidos a partir de los modelos acústicos limpios y las distintas matrices de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$.

En la Tabla 8.1 se pueden apreciar los mejores resultados para las técnicas SPLICE ME ("Entre." CLK, "Reco." HF SPLICE ME), MEMLIN ("Entre." CLK, "Reco." HF MEMLIN), MEMLIN MP ("Entre." CLK, "Reco." HF MEMLIN MP) y las técnicas híbridas a partir del cálculo de matrices de rotación dependientes de GMMs basadas en

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

Tabla 8.1: Mejores resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando distintas técnicas híbridas a partir del cálculo de matrices de rotación dependientes de GMMs. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabras generados con la señal limpia (CLK en la columna de "Entre."), o extendidos se extienden a partir de los anteriores haciendo uso de 16 matrices de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$, (CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con las técnicas SPLICE ME, MEMLIN o MEMLIN MP. Junto al nombre de las diferentes técnicas aparece el número de Gaussianas con que se modelaron los correspondientes espacios. 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.

los algoritmos MEMLIN y MEMLIN MP ("Entre." CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$, "Reco." HF MEMLIN o HF MEMLIN MP, respectivamente). Los tres primeros experimentos se incluyen a modo de comparación. En todos los casos, junto al nombre de la técnica de adaptación 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 cuyos resultados completos se pueden consultar en el Anexo 8.6 de este mismo Capítulo). Por otra parte, para el algoritmo MEMLIN MP, los vectores de características ruidosos se representan con dos Gaussianas para cada par de componentes, s_x y s_y^e. 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, calculándose esta última como (5.29).

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. Se puede observar que el valor del estadístico W, w, es, para el primero de los casos, $w = 7{,}91 > 1{,}96$, por lo que la mejora del algoritmo híbrido asociado a la técnica MEMLIN presentada en esta Sección se puede considerar independiente de la base de datos con un intervalo de confianza del 95 % con respecto a la mejora alcanzada con el método MEMLIN. Por otra parte, para el segundo de los casos, en el que se cotejan los resultados alcanzados 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 dos resultados se han de tratar con cautela dadas las limitaciones de la propia prueba, ya mencionadas en la Sección 4.3.

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, tal y como se puede comprobar por la experimentación presentada en la Tabla 4.4): MWER de 4.63 % ("Entre." HF, "Reco." HF) y MWER de 3.42 % ("Entre." HF†, "Reco." HF). 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 muchas menos variables 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 se transforman 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 se adaptaron los modelos acústicos limpios hacia el espacio ruidoso mediante el algoritmo MLLR no supervisado. Para ello se hizo uso de las trascripciones provinientes de decodificar la señal ruidosa con los modelos acústicos limpios. Bajo esas condiciones de experimentación la técnica MLLR obtuvo 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 MIMP 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 observada en la Tabla 8.1 es extensible a cualquier número de componentes, aunque bien es cierto que ésta es más importante 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 62.57 %, mientras que los algoritmos híbridos a partir del cálculo de matrices de rotación dependientes de GMMs basados en 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 alcanzan importantes mejoras.

8.4.2 Resultados obtenidos con técnicas híbridas basadas en reentrenamiento supervisado.

Paso previo a la presentación de los resultados de RAH obtenidos al decodificar los vectores de características normalizados aplicando modelos acústicos reentrenados de

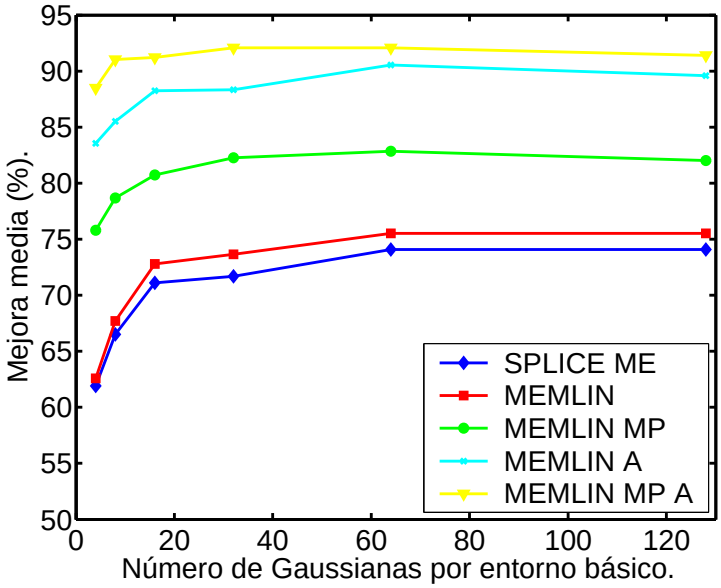


Figura 8.4: Mejora media del WER, MIMP, para las técnicas SPLICE ME, MEMLIN, MEMLIN MP, la técnica híbrida a partir del cálculo de matrices de rotación dependientes de GMMs y basada en el algoritmo MEMLIN, identificada como MEMLIN A, y, ya por último, la técnica híbrida a partir del cálculo de matrices de rotación dependientes de GMMs y basada en el algoritmo MEMLIN MP, nombrada como MEMLIN MP A. En todos los casos se representan en función del número de Gaussianas por entorno básico empleado. Se ha utilizado la parametrización estándar ETSI y modelos acústicos de palabra generados a partir de la señal limpia.

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 se estima en una fase de entrenamiento previa, si es que fuera preciso, los diversos parámetros necesarios para la técnica de normalización de vectores de características seleccionada; para lo que, en general y para las técnicas presentadas en este trabajo, se emplea un corpus de entrenamiento estéreo. Asimismo, y todavía en la misma fase de entrenamiento, se obtienen los modelos acústicos que representan al espacio normalizado haciendo uso de los previamente compensados vectores de características ruidosos del corpus de entrenamiento. 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 de los vectores acústicos los algoritmos MEMLIN y MEMLIN MP, ("Entre." HF MEMLIN, "Reco." MEMLIN y "Entre." HF MEMLIN MP, "Reco." MEMLIN MP, respectivamente). Por otra parte, y a modo de comparación, se vuelven a incluir los resultados alcanzados cuando se reconoce la señal limpia con modelos acústicos limpios ("Entre." CLK, "Reco." CLK), la señal ruidosa con modelos acústicos generados a partir de todo el corpus de entrenamiento ruidoso ("Entre." HF, "Reco." HF), y la señal ruidosa con modelos acústicos específicos para cada entorno básico ("Entre." †HF, "Reco." HF). Del mismo modo, allí donde fuera procedente, junto al nombre de la técnica

Entre.	Reco.	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

Tabla 8.2: Mejores resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando distintas técnicas híbridas basadas en reentrenamiento supervisado. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabra generados a partir de la señal limpia (CLK en la columna de "Entre."), ruidosa (HF en la columna de "Entre.", o †HF, si los modelos son dependientes del entorno básico), o tras adaptar el corpus de entrenamiento ruidoso mediante las técnicas MEMLIN o MEMLIN MP (HF MEMLIN o HF MEMLIN MP en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la limpia (CLK), la ruidosa (HF), o la ruidosa normalizada con las técnicas MEMLIN (HF MEMLIN), o MEMLIN MP (HF MEMLIN MP). Junto al nombre de las diferentes técnicas aparece el número de Gaussianas con que se modelaron los correspondientes espacios. 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.

de adaptación 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 componentes, cuyos resultados completos se incluyen en el Anexo 8.6 de este mismo Capítulo). Adicionalmente, y par al atécnica MEMLIN MP, 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). 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 como (5.29).

A pesar de que ya se pueden intuir los resultados, conviene, 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 basadas en reentrenamiento supervisado. 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 correspondiente algoritmo híbrido 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 basado en reentrenamiento supervisado, se tiene que $w = 8{,}56 > 1{,}96$, con lo que se puede considerar nuevamente que la diferencia de comportamiento entre 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 4.3.

A partir de los resultados presentados en la Tabla 8.2 se puede concluir que, teniendo en cuenta únicamente las mejores tasas medias 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 reentrenados supervisadamente (técnicas híbridas basadas en reentrenamiento supervisado) 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 cotejan los resultados alcanzados cuando las señales ruidosas se decodifican con modelos acústicos generados con todo el corpus de entrenamiento degradado ("Entre." HF, "Reco." HF) o específicamente para cada entorno básico ("Entre." † HF, "Reco." 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 degradado inicial, 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 que constituyen el método híbrido, se puede constatar que no difieren de un modo estadísticamente significativo. Así, por ejemplo, si la técnica MEMLIN se aplica con 4 Gaussianas por entorno básico 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 aparecieran las mejoras medias (MIMP) para distintos números de Gaussianas con que se modelaran los entornos básicos.

8.5 Anexo J.

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 acústicos 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 8.2.1. De este modo, el error cuadrático medio asociado a cada par de Gaussianas, $\xi_{s_x,s_{\hat{x}}}$, se define como

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

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

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

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

$$\begin{aligned}\frac{\delta\xi_{s_x,s_{\hat{x}}}}{\delta\mathbf{A}_{s_x,s_{\hat{x}}}} &= \frac{1}{T}\sum_t p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t) \\ &\times \frac{\delta}{\delta\mathbf{A}_{s_x,s_{\hat{x}}}}\left[Tra\left[\hat{\mathbf{x}}_t(\hat{\mathbf{x}}_t)^T - \hat{\mathbf{x}}_t(\mathbf{x}_t)^T(\mathbf{A}_{s_x,s_{\hat{x}}})^T\right.\right. \\ &- \left.\left.\mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t(\hat{\mathbf{x}}_t)^T + \mathbf{A}_{s_x,s_{\hat{x}}}\mathbf{x}_t(\mathbf{x}_t)^T(\mathbf{A}_{s_x,s_{\hat{x}}})^T\right]\right] = \mathbf{0}. \quad \text{(J.3)}\end{aligned}$$

O, lo que es lo mismo

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

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

$$\mathbf{A}_{s_x,s_{\hat{x}}} = \sum_t p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t)\hat{\mathbf{x}}_t(\mathbf{x}_t)^T\left(\sum_t p(s_x|\mathbf{x}_t)p(s_{\hat{x}}|\hat{\mathbf{x}}_t)\mathbf{x}_t(\mathbf{x}_t)^T\right)^{-1}, \quad \text{(J.5)}$$

que coincide con la expresión (8.4) presentada en la Sección 8.2.1

8.6 Anexo K.

En este Anexo se presentan los resultados obtenidos en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) de la base de datos *SpeechDat Car* en español utilizando tanto distintas técnicas de adaptación de vectores de características (SPLICE ME, MEMLIN y MEMLIN MP), como métodos de adaptación conjunta de señal y modelos acústicos. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabras que pueden ser limpios (CLK) o extendidos a partir de los anteriores haciendo uso de matrices de rotación dependientes de GMMs (CLK $A_{s_x,s_{\hat{x}}}$). A su vez, y junto al nombre de las distintas técnicas, se ha incluido el número de Gaussianas con que se modelan los correspondientes entornos básicos (4, 8, 16, 32, 64 y 128 Gaussianas). En el caso de aplicar la técnica MEMLIN MP, se emplean 2 componentes para representar los vectores de características ruidosos asociados a cada par de Gaussianas: s_x y s_y^e.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF SPLICE ME 4	2.96	8.15	6.01	11.90	12.86	7.62	17.69	8.75	61.90
CLK	HF SPLICE ME 8	2.48	7.20	5.17	10.28	11.34	6.67	18.70	7.80	66.50
CLK	HF SPLICE ME 16	2.57	6.95	4.76	8.77	9.82	5.71	13.61	6.86	71.10
CLK	HF SPLICE ME 32	2.57	6.86	4.34	8.77	9.34	6.03	13.61	6.73	71.70
CLK	HF SPLICE ME 64	2.29	6.17	3.92	8.40	8.77	5.56	12.93	6.25	74.08
CLK	HF SPLICE ME 128	2.48	6.52	3.92	7.64	9.06	5.08	12.93	6.25	74.08

Tabla 8.3: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características SPLICE ME. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabra generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica SPLICE ME. Junto al nombre de la técnica aparece el número de Gaussianas con que se modelaron los distintos entornos básicos. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF MEMLIN 4	2.96	7.80	6.29	11.78	12.68	7.14	17.69	8.61	62.57
CLK	HF MEMLIN 8	2.19	6.95	4.90	10.28	10.68	6.67	19.05	7.56	67.69
CLK	HF MEMLIN 16	2.38	6.78	4.76	8.65	8.77	5.56	12.59	6.51	72.80
CLK	HF MEMLIN 32	2.10	6.43	4.20	8.27	8.87	6.19	12.24	6.33	73.66
CLK	HF MEMLIN 64	2.10	6.00	3.78	7.89	8.29	5.71	11.56	5.95	75.53
CLK	HF MEMLIN 128	2.10	6.26	3.78	7.52	8.48	5.24	11.90	5.95	75.53

Tabla 8.4: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características MEMLIN. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabra generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica MEMLIN. Junto al nombre de la técnica aparece el número de Gaussianas con que se modelaron los distintos entornos básicos. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK	HF MEMLIN MP 4	2.10	5.83	3.78	7.77	8.01	5.56	12.93	5.89	75.79
CLK	HF MEMLIN MP 8	1.91	5.75	2.66	6.64	7.53	5.56	9.86	5.30	78.69
CLK	HF MEMLIN MP 16	1.91	5.40	2.80	6.77	6.58	4.60	7.82	4.88	80.73
CLK	HF MEMLIN MP 32	2.00	5.23	2.52	6.02	6.20	4.29	6.80	4.56	82.27
CLK	HF MEMLIN MP 64	2.00	4.97	2.24	6.02	6.01	4.29	6.80	4.44	82.86
CLK	HF MEMLIN MP 128	1.91	4.89	2.80	6.02	6.67	4.29	7.14	4.61	82.01

Tabla 8.5: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica de adaptación de vectores de características MEMLIN MP. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabra generados a partir de la señal limpia (CLK en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica MEMLIN MP. Junto al nombre de la técnica aparece el número de Gaussianas con que se modelaron los distintos entornos básicos. Adicionalmente se emplean 2 componentes para representar los vectores de características ruidosos asociados a cada par de Gaussianas: s_x y s_y^e. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN 4	2.29	5.40	2.24	6.02	5.24	3.81	5.10	4.30	83.55
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN 8	2.48	4.72	2.52	4.89	4.67	3.17	5.10	3.89	85.50
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN 16	2.19	4.63	2.10	3.88	4.10	2.38	3.06	3.33	88.24
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN 32	2.00	4.37	1.54	4.51	4.29	2.54	3.06	3.31	88.33
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN 64	2.19	3.95	2.10	3.26	3.24	1.90	2.38	2.86	90.54
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN 128	1.81	4.29	1.40	4.39	3.72	2.38	2.04	3.05	89.59

Tabla 8.6: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica híbrida a partir del cálculo de matrices de rotación dependientes de GMMs basada en el algoritmo de compensación MEMLIN. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabra generados tras extender los obtenidos con la señal limpia haciendo uso de 16 matrices de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$, (CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica MEMLIN. Junto a su nombre 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.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN MP 4	2.29	4.37	1.96	4.14	4.00	2.06	3.40	3.28	88.49
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN MP 8	2.10	4.20	1.96	2.38	3.34	1.59	2.72	2.75	91.04
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN MP 16	2.19	3.95	1.96	3.26	3.05	1.43	1.70	2.72	91.21
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN MP 32	2.10	3.86	1.54	2.88	2.86	1.43	1.70	2.54	92.07
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN MP 64	1.91	3.86	1.68	2.88	2.96	1.27	2.04	2.54	92.06
CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$	HF MEMLIN MP 128	2.19	4.03	1.68	3.26	2.96	1.43	1.70	2.68	91.39

Tabla 8.7: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica híbrida a partir del cálculo de matrices de rotación dependientes de GMMs basada en el algoritmo de compensación MEMLIN MP. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabra generados tras extender los obtenidos con la señal limpia haciendo uso de 16 matrices de rotación $\mathbf{A}_{s_x,s_{\hat{x}}}$, (CLK-$\mathbf{A}_{s_x,s_{\hat{x}}}$). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa (HF) normalizada con la técnica MEMLIN MP. Junto a su nombre aparece el número de Gaussianas con que se modelaron los correspondientes espacios. Por su parte, se representan 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.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
HF MEMLIN 4	HF MEMLIN 4	0.86	3.86	0.84	2.63	2.10	1.75	0.34	2.02	94.63
HF MEMLIN 8	HF MEMLIN 8	0.67	3.86	1.54	2.26	2.19	1.43	0.34	2.00	94.71
HF MEMLIN 16	HF MEMLIN 16	0.67	3.52	1.68	2.01	1.72	1.27	0.34	1.81	95.65
HF MEMLIN 32	HF MEMLIN 32	0.57	3.26	1.82	2.26	1.62	1.59	0.00	1.79	95.74
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 128	HF MEMLIN 128	0.48	3.52	1.40	2.26	1.53	1.11	0.34	1.72	96.08

Tabla 8.8: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica híbrida basada en reentrenamiento supervisado a partir del método de compensación MEMLIN. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabra generados tras adaptar con el criterio ML el corpus de entrenamiento ruidoso normalizado con la técnica MEMLIN (HF MEMLIN en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa compensada con la técnica MEMLIN (HF MEMLIN). Junto a su nombre 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.

Entre.	Reco.	E1	E2	E3	E4	E5	E6	E7	MWER (%)	MIMP (%)
HF MEMLIN MP 4	HF MEMLIN MP 4	0.95	3.69	1.68	2.01	1.72	1.43	0.34	1.91	95.14
HF MEMLIN MP 8	HF MEMLIN MP 8	0.67	3.09	1.26	2.38	1.62	1.11	0.00	1.67	96.33
HF MEMLIN MP 16	HF MEMLIN MP 16	0.86	2.92	1.54	2.51	1.33	0.95	0.00	1.65	96.42
HF MEMLIN MP 32	HF MEMLIN MP 32	0.48	3.00	1.12	2.13	1.62	1.11	0.00	1.56	96.85
HF MEMLIN MP 64	HF MEMLIN MP 64	0.95	3.17	1.26	2.13	1.62	1.11	0.00	1.70	96.16
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

Tabla 8.9: Resultados obtenidos con la base de datos *SpeechDat Car* en español en términos de WER (%) para los diferentes entornos básicos (E1,..., E7) utilizando la técnica híbrida basada en reentrenamiento supervisado a partir del método de compensación MEMLIN MP. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabra generados tras adaptar con el criterio ML el corpus de entrenamiento ruidoso normalizado con la técnica MEMLIN MP (HF MEMLIN MP en la columna de "Entre."). La columna marcada como "Reco." hace referencia a la señal empleada para reconocer, que será la ruidosa compensada con la técnica MEMLIN MP (HF MEMLIN MP). Junto a su nombre aparece el número de Gaussianas con que se modelaron los correspondientes espacios. Asimismo, se emplean 2 componentes para representar los vectores de características ruidosos asociados a cada par de Gaussianas s_x y s_y^e. Se incluye igualmente el WER medio, MWER, así como la mejora media, MIMP.

Capítulo 9

Resultados con la Base de Datos *Aurora* 2

9.1 Introducción.

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 dicho corpus se ha grabado en ambientes hostiles reales, *Aurora* 2 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, un banco de pruebas muy útil de cara a presentar resultados ante la comunidad científica.

Con vistas a realizar posteriores análisis sobre los resultados obtenidos a lo largo de la experimentación, resulta conveniente recordar brevemente ciertos aspectos del corpus *Aurora* 2. De este modo, se insiste en que se generó a partir de la base de datos de dígitos aislados y conectados en inglés *TIDigits* [Leo84], a la que se le añadió, posteriormente y de modo artificial, tanto distintos tipos de ruidos aditivos con diferentes SNRs (20dB, 15dB, 10dB, 5dB, 0dB y -5dB), como, en algunos casos, distorsión convolucional, con lo que se busca simular algunos de los escenarios más característicos 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 4.4, dos corpora distintos de entrenamiento, según si constan 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 esta solución se aproxima más al problema de robustez inicialmente considerado.

Por su parte, cabe recordar asimismo que los tres *sets* en que se haya dividido el corpus de reconocimiento responden 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 está construido a partir de distorsión convolucional y con los mismos tipos de ruidos, no así SNRs, que aparecen en el corpus de entrenamiento multi-condición, con lo que mediante los correspondientes resultados se puede conjeturar sobre 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 presentes 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-condición y otro que no lo está, aunque en ambos casos se ha aplicado una distorsión convolucional distinta de la presente en el corpus de entrenamiento multi-condición, de modo que con este *set* es factible se puede analizar el comportamiento de las distintas técnicas ante una distorsión convolucional no vista previamente y a la que, en uno de los casos, se le ha añadido un ruido aditivo tampoco observado con anterioridad.

Este Capítulo se articula en tres Secciones, tantas como técnicas seleccionadas de cara a completar la experimentación con la base de datos *Aurora* 2. Para tal fin se se eligieron los métodos quizás más representativos de entre los presentados en este trabajo, a saber, MEMLIN (Sección 9.2), MEMLIN MP (Sección 9.3) y la técnica híbrida basada en el algoritmo MEMLIN MP haciendo uso de matrices de rotación dependientes de GMMs (Sección 9.4). En todos los casos se presentan resultados de RAH haciendo uso tanto de parametrización estándar ETSI, como ETSI *advanced*.

9.2 Resultados Obtenidos con la Técnica MEMLIN.

Para llevar a cabo la experimentación sobre la base de datos *Aurora* 2, 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 este modo 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 la base de datos *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 CMN, salvo para el caso de emplear la parametrización ETSI *advanced*, que, de algún modo, ya lo lleva implícito a partir de una ecualización ciega. 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 4.4. En cuanto a la extracción de características, ésta puede realizarse bien mediante la parametrización estándar ETSI, bien a partir de la parametrización ETSI *advanced*, de modo que los resultados de referencia se pueden consultar, según los casos, en las Tablas 4.5 y 4.7.

9.2.1 Resultados obtenidos con la técnica MEMLIN y parametrización estándar ETSI

En la Tabla 9.1 se presentan, del modo típico en que se suele hacer, los correspondientes resultados alcanzados tras aplicar la técnica de compensación MEMLIN, tanto en

Aurora 2 Small Vocabulary	Clean 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	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	Restaurant	Street	Airport	Station	Average	Subway M	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%

Tabla 9.1: Exatitud por palabra, *word accuracy*, (%) y mejoras relativas (%) obtenidas para los distintos *sets* (A, B y C) de la base de datos *Aurora* 2 utilizando la técnica de adaptación de vectores de características MEMLIN. Cada entorno básico, así como el espacio limpio, se representan con 128 Gaussianas. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*.

términos de exactitud por palabra como en mejora relativa. Tanto los 24 entornos básicos como el espacio limpio se representan con 128 Gaussianas (se realizó un barrido para distintos números de componentes: 4, 8, 16, 32, 64 y 128). 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 4.5 se puede concluir que el método MEMLIN 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 *sets*. Esto es algo que no debe sorprender puesto que los tipos de ruido incluidos en dicho *test* habían sido previamente 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 acústico 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

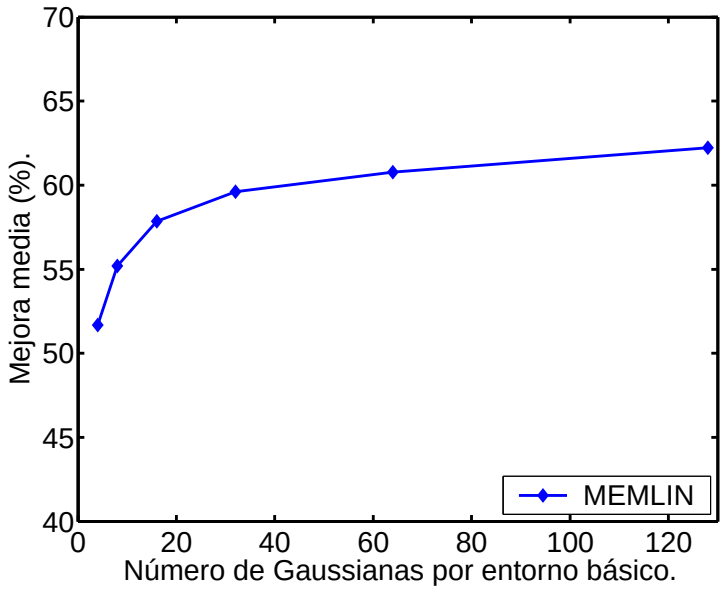


Figura 9.1: Mejoras medias de la exactitud por palabra, *word accuracy*, (%) obtenidas con la base de datos *Aurora* 2 utilizando la técnica de adaptación de vectores de características MEMLIN y empleando distinto número de componentes para modelar los entornos básicos. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*.

son sensiblemente menos satisfactorios (mejora media del 39.67 %), lo que indica 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 9.1 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 que las prestaciones mejoran conforme se incrementa el número de componentes, el rango dinámico de ésta no llega en ningún momento a los niveles que se habían alcanzado con la base de datos *SpeechDat Car* en español.

9.2.2 Resultados obtenidos con la técnica MEMLIN y parametrización ETSI *advanced*

En la Tabla 9.2 se pueden observar los resultados que la técnica de compensación MEMLIN proporciona al aplicarse sobre los coeficientes ETSI *advanced*. Se incluyen, al igual que en la subsección anterior, tanto la exactitud por palabra como la mejora relativa. Asimismo, los 24 entornos básicos y el espacio limpio se representan con 128 Gaussianas, aunque previamente se realizó un barrido para distintos números de componentes: 4, 8, 16, 32 y 64.

A la luz pues de los valores presentados en las Tablas 9.2 y 4.7 se puede concluir que, teniendo en cuenta únicamente los mejores resultados para la técnica MEMLIN, el

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restauran	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	98,97	99,17	99,29	99,13	99,11	98,97	99,17	99,29	99,13	99,02	98,88	98,95	99,10
	20 dB	98,10	98,43	98,57	98,15	98,31	98,44	97,77	98,69	98,80	98,42	97,76	98,07	97,91	98,28
	15 dB	96,94	97,53	97,88	96,77	97,28	97,16	96,83	97,42	97,56	97,24	96,50	96,53	96,52	97,11
	10 dB	94,00	94,71	95,62	94,55	94,72	94,88	93,44	95,23	95,67	94,80	92,70	92,72	92,71	94,35
	5 dB	87,29	87,34	89,40	88,29	88,08	86,53	86,31	88,38	88,50	87,43	82,75	80,80	81,78	86,56
	0 dB	70,73	65,04	71,63	70,56	69,49	67,39	65,91	70,65	70,64	68,65	58,63	54,74	56,68	66,59
	-5dB	39,42	32,73	36,02	40,97	37,28	35,36	35,59	37,98	39,48	37,10	29,37	27,29	28,33	35,42
	Average	89,41	88,61	90,62	89,66	**89,58**	88,88	88,05	90,07	90,23	**89,31**	85,67	84,57	**85,12**	**88,53**

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restauran	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Detailed relative results in terms of error reduction. Halving the error rate = +50%	Clean	19,81%	-5,96%	12,13%	4,19%	7,54%	19,81%	-5,96%	12,13%	4,19%	7,54%	-18,37%	-22,91%	-20,64%	1,90%
	20 dB	41,52%	83,39%	50,99%	51,34%	56,81%	84,14%	46,05%	86,95%	76,91%	73,51%	66,22%	60,30%	63,26%	64,78%
	15 dB	63,83%	91,11%	81,51%	67,57%	76,01%	88,85%	73,27%	90,12%	87,01%	84,81%	74,98%	68,19%	71,58%	78,64%
	10 dB	75,50%	89,91%	87,98%	80,33%	83,43%	89,36%	80,67%	90,60%	90,33%	87,74%	74,00%	70,85%	72,42%	82,95%
	5 dB	75,87%	83,57%	84,69%	80,77%	81,23%	81,60%	78,52%	84,58%	84,67%	82,35%	65,06%	61,16%	63,11%	78,05%
	0 dB	62,27%	62,99%	68,23%	65,67%	64,79%	64,89%	58,75%	67,21%	67,56%	64,60%	45,18%	40,73%	42,95%	60,35%
	-5dB	32,19%	32,65%	31,33%	36,63%	33,20%	34,74%	29,52%	34,53%	35,52%	33,58%	18,91%	18,13%	18,52%	30,41%
	Average	63,80%	82,20%	74,68%	69,14%	72,45%	81,77%	67,45%	83,89%	81,30%	78,60%	65,09%	60,25%	62,67%	**72,96%**

Tabla 9.2: Exactitud por palabra, *word accuracy* (%), y mejoras relativas (%) obtenidas para los distintos *sets* (A, B y C) de la base de datos *Aurora* 2 utilizando la técnica de adaptación de vectores de características MEMLIN, modelando cada uno de los entornos básicos con 128 Gaussianas. Se ha empleado la parametrización ETSI *advanced* y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*.

método proporciona una cierta mejora añadida (72.96 %) a la ya importante obtenida por la propia parametrización ETSI *advanced* (67.41 %). Obsérvese asimismo como el *set* C sigue proporcionando las peores tasas de RAH, mientras que para los *sets* A y B se alcanzan valores de reconocimiento muy similares.

Dada la parametrización ETSI *advanced*, y 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, se presenta la Figura 9.2. En ella se muestra la mejora media alcanzada cuando se barre el número de componentes: 4, 8, 16, 32, 64 y 128. Igualmente, y a modo de comparación, se ha incluido la mejora media obtenida al emplear únicamente la parametrización ETSI *advanced*. Se puede apreciar como, para todos los casos, el aplicar el método MEMLIN aporta una cierta mejora de reconocimiento, a pesar de los ya importantes resultados alcanzados por el método de extracción de características seleccionado.

A partir de los resultados presentandos en las secciones presente y 5.5 se puede constatar, y ésta es una de las más importantes conclusiones, que el método de adaptación de vectores acústicos MEMLIN proporciona una cierta mejora, mayor o menor según los casos, para bases de datos con características claramente diferentes, como son *SpeechDat Car* y *Aurora* 2). Asimismo, la mejora proporcionada por la técnica MEMLIN se reproduce tanto si se aplica sobre la parametrización estándar ETSI, como si se hace sobre los coeficientes ETSI *advanced*. Obsérvese que todo lo anterior certifica la robustez del método MEMLIN ante distintos parámetros de experimentación, convirtiéndola en una

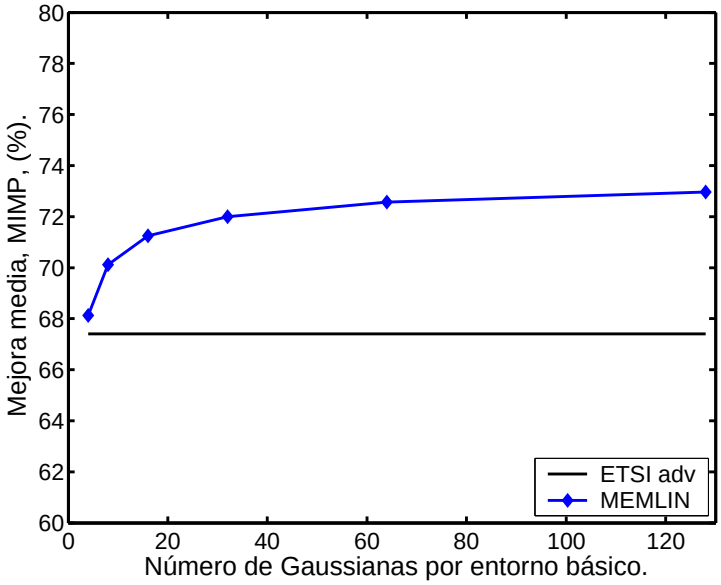


Figura 9.2: Mejoras medias de la exatitud por palabra, *word accuracy*, obtenidas para los distintos *sets* (A, B y C) de la base de datos *Aurora*2 utilizando la técnica de adaptación de vectores de características MEMLIN empleando distinto número de componentes para modelar los entornos básicos. Se ha empleado la parametrización ETSI *advanced* y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*. Igualmente, y a modo de comparación, se ha incluido la mejora media obtenida al emplear la parametrización ETSI *advanced*.

solución apta para en un número elevado de circunstancias y situaciones.

9.3 Resultados Obtenidos con la Técnica MEMLIN MP.

La experimentación sobre la base de datos *Aurora* 2 para la técnica MEMLIN MP se rige por los mismos parámetros considerados en la Sección 9.2: selección de corpus de entrenamiento, métodos de extracción de características (estándar ETSI y ETSI *advanced*) y modelados de lenguaje y acústico. De esta manera, los resultados de referencia se pueden consultar nuevamente, y según los casos, en las Tablas 4.5 y 4.7.

9.3.1 Resultados obtenidos con la técnica MEMLIN MP y parametrización estándar ETSI

En la Tabla 9.3 se presentan los resultados alcanzados tras aplicar la técnica de compensación MEMLIN MP sobre los coeficientes estándar ETSI, tanto en términos de exactitud por palabra como en mejora relativa. Los 24 entornos básicos así como el espacio limpio se representan con 128 Gaussianas, aunque, como en la mayoría de los experimentos, se realizó un barrido previo con distintos números de componentes: 4, 8, 16, 32, 64 y 128. Por su parte, los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y^e, se representan en todos los casos con 2 componentes.

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restauran	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	99,20	98,91	99,25	99,41	99,20	99,20	98,91	99,25	99,41	99,20	99,02	99,06	99,04	99,16
	20 dB	98,25	98,22	98,54	98,27	98,32	98,37	97,95	98,45	98,55	98,33	97,48	97,43	97,46	98,15
	15 dB	97,39	97,31	97,59	96,86	97,29	97,00	96,75	96,34	97,32	96,85	94,91	94,67	94,79	96,61
	10 dB	94,45	93,47	93,60	94,12	93,91	93,25	91,34	92,59	93,77	92,74	87,05	87,60	87,33	92,13
	5 dB	86,84	80,62	81,96	84,86	83,57	81,05	76,05	80,25	81,26	79,66	63,61	72,05	67,83	78,86
	0 dB	65,79	52,05	55,27	64,85	59,49	55,41	51,78	58,80	55,24	55,31	31,91	46,22	39,07	53,73
	-5dB	35,09	27,65	27,32	34,55	31,15	28,62	27,35	29,75	28,84	28,64	14,94	24,10	19,52	27,82
	Average	88,55	84,33	85,39	87,79	**86,52**	85,02	82,77	85,29	85,23	**84,58**	74,99	79,59	**77,29**	**83,90**

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restauran	Street	Airport	Station	Average	Subway M	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%	-18,37%	-2,98%	-10,68%	9,51%
	20 dB	46,17%	81,15%	49,95%	54,56%	57,96%	83,50%	50,44%	84,60%	72,21%	72,69%	62,04%	47,24%	54,64%	63,19%
	15 dB	69,23%	90,34%	78,91%	68,49%	76,74%	88,23%	72,53%	86,02%	85,74%	83,13%	63,58%	51,12%	57,35%	75,42%
	10 dB	77,34%	87,53%	82,45%	78,80%	81,53%	85,98%	74,50%	85,39%	86,10%	82,99%	53,84%	50,35%	52,10%	76,23%
	5 dB	75,02%	74,86%	73,95%	75,15%	74,74%	74,12%	62,42%	73,81%	75,03%	71,35%	26,30%	43,45%	34,88%	65,41%
	0 dB	55,89%	49,25%	49,91%	59,01%	53,51%	52,00%	41,66%	53,97%	50,54%	49,54%	9,78%	29,57%	19,68%	45,16%
	-5dB	27,35%	27,56%	21,99%	29,73%	26,66%	27,93%	20,49%	25,83%	24,19%	24,61%	2,34%	14,53%	8,44%	22,20%
	Average	64,73%	76,62%	67,03%	67,20%	68,90%	76,76%	60,31%	76,76%	73,92%	71,94%	43,11%	44,35%	43,73%	65,08%

Tabla 9.3: Exactitud por palabra, *word accuracy* (%), y mejoras relativas (%) obtenidas para los distintos *sets* (A, B y C) de la base de datos *Aurora* 2 utilizando la técnica de normalización de vectores de características MEMLIN MP, modelando cada uno de los entornos básicos con 128 Gaussianas. A su vez, los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y^e, se representan con 2 componentes. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*.

A la luz de los valores presentados en las Tablas 9.3 y 4.5 se puede concluir que, teniendo en cuenta únicamente los mejores resultados para la técnica MEMLIN MP, el método se comporta satisfactoriamente, obteniendo una mejora media final del 65.08 % y mejorando consistentemente todos los valores medios alcanzados con el método MEMLIN para las distintas SNRs cuando este último hace uso de GMMs compuestas por 128 Gaussianas (ver Tabla 9.1). Sin embargo los resultados muestran nuevamente un ligero déficit del método ante el *set* C, mientras que las tasas de RAH logradas para los *sets* A y B se mantienen similares entre sí, lo que sustenta la idea de la robustez del algoritmo MEMLIN MP ante ruidos no vistos. Con todo ello se puede concluir que el comportamiento y limitaciones observados para la técnica MEMLIN (Subsección 9.2.1) siguen dándose en este caso en menor medida.

A la hora de observar como se comporta el algoritmo MEMLIN MP cuando se incrementa el número de Gaussianas con que se modelan los distintos entornos básicos, en la Figura 9.3 se muestra la mejora media al realizar el siguiente barrido del número de componentes: 4, 8, 16, 32, 64 y 128. Apréciese como para todos los casos la técnica MEMLIN MP proporciona mejores resultados que el algoritmo MEMLIN, incluido asimismo a modo de comparación. A pesar de todo, en este caso la mejora no llega a ser tan importante como la que se alcanzaba con la base de datos *SpeechDat Car* en español.

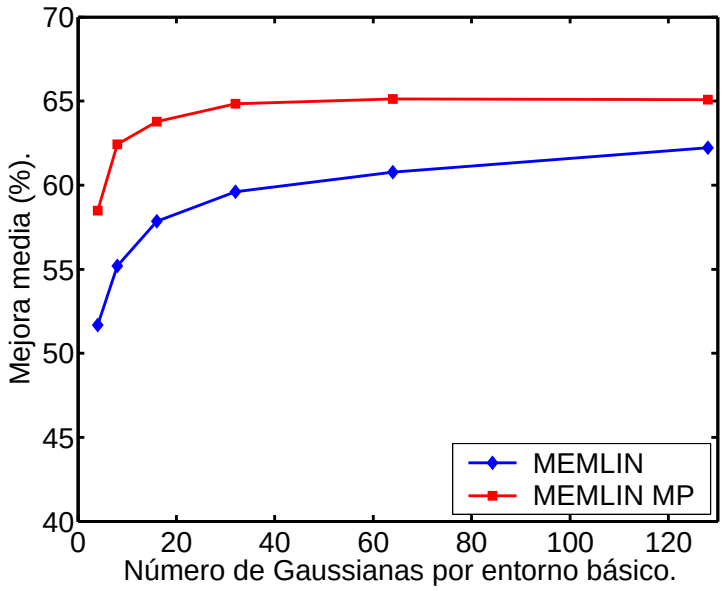


Figura 9.3: Mejoras medias de la exactitud por palabra, *word accuracy* (%), obtenidas con la base de datos *Aurora* 2 utilizando la técnica de adaptación de vectores de características MEMLIN MP y empleando distinto número de componentes para modelar los entornos básicos. A su vez, los vectores de características ruidosos asociados a cada par de Gaussianas s_x y s_y^e se representan con 2 componentes. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*. A modo de comparación se han incluido los resultados alcanzados con la técnica MEMLIN.

9.3.2 Resultados obtenidos con la técnica MEMLIN MP y parametrización ETSI *advanced*

En la Tabla 9.4 se presentan los resultados que la técnica de compensación MEMLIN MP proporciona cuando se aplica sobre los coeficientes ETSI *advanced*, incluyendo tanto la exactitud por palabra como la mejora relativa. Al igual que en subsecciones anteriores, los 24 entornos básicos y el espacio limpio se representan con 128 Gaussianas, aunque previamente se realizó un barrido con distintos números de componentes (4, 8, 16, 32, 64 y 128), cuyas mejoras medias se pueden observar, de un modo resumido, en la Figura 9.4. Por su parte, los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y^e, se representan en todos los casos con 2 componentes.

A la luz pues de los valores presentados en las Tablas 9.4 y 4.7 se puede concluir que, teniendo en cuenta únicamente los mejores resultados para la técnica MEMLIN MP, el método proporciona una mejora añadida (75.46 %) a la ya obtenida por la parametrización ETSI *advanced* (67.41 %). Cabe destacar asimismo como el comportamiento del algoritmo, atendiendo a los distintos *sets*, resulta muy similar al que poseía la técnica MEMLIN bajo las mismas condiciones de experimentación, esto es, el *set* C sigue proporcionando las peores tasas de RAH, mientras que para los *sets* A y B se alcanzan valores de reconocimiento muy similares.

A la hora de observar como se comporta el algoritmo MEMLIN MP cuando se incrementa el número de Gaussianas con que se modelan los distintos entornos básicos,

Aurora 2 Small Vocabulary	Clean 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	99,14	99,00	99,17	99,35	99,17	99,14	99,00	99,17	99,35	99,17	99,05	98,94	99,00	99,13
	20 dB	98,37	98,46	98,72	98,27	98,46	98,56	97,83	98,81	98,92	98,53	98,10	98,16	98,13	98,42
	15 dB	97,34	97,59	98,15	97,20	97,57	97,55	97,13	97,71	97,66	97,51	97,18	96,77	96,98	97,43
	10 dB	94,68	95,17	96,30	95,26	95,35	95,07	94,07	95,23	96,22	95,15	93,72	93,21	93,46	94,89
	5 dB	88,88	88,13	91,30	88,98	89,32	86,99	87,75	88,87	89,68	88,32	85,47	82,68	84,08	87,87
	0 dB	72,80	66,11	75,58	72,56	71,76	67,93	68,96	72,47	72,82	70,55	63,33	59,15	61,24	69,17
	-5dB	42,33	34,40	40,61	43,74	40,27	36,85	37,34	40,49	43,17	39,46	32,95	30,05	31,50	38,19
	Average	90,41	89,09	92,01	90,45	**90,49**	89,22	89,15	90,62	91,06	**90,01**	87,56	85,99	**86,78**	**89,56**

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Detailed relative results in terms of error reduction. Halving the error rate = +50%	Clean	22,58%	-2,85%	12,13%	12,50%	11,09%	22,58%	-2,85%	12,13%	12,50%	11,09%	-14,67%	-16,26%	-15,47%	5,78%
	20 dB	49,99%	83,72%	56,09%	54,57%	61,09%	85,38%	47,50%	88,14%	79,27%	75,07%	71,30%	62,16%	66,73%	67,81%
	15 dB	68,55%	91,34%	83,85%	71,90%	78,91%	90,40%	75,82%	91,26%	87,52%	86,25%	79,82%	70,43%	75,13%	81,09%
	10 dB	78,24%	90,78%	89,87%	82,88%	85,44%	89,75%	82,53%	90,60%	91,56%	88,61%	77,62%	72,80%	75,21%	84,66%
	5 dB	78,89%	84,60%	87,44%	81,92%	83,21%	82,22%	80,78%	85,24%	86,25%	83,62%	70,58%	64,95%	67,76%	80,29%
	0 dB	64,93%	64,12%	72,65%	68,00%	67,42%	65,47%	62,44%	69,24%	69,97%	66,78%	51,41%	46,50%	48,96%	63,47%
	-5dB	35,45%	34,32%	36,26%	39,60%	36,41%	36,24%	31,43%	37,17%	39,45%	36,07%	23,02%	21,24%	22,13%	33,42%
	Average	68,12%	82,91%	77,98%	71,85%	75,22%	82,64%	69,82%	84,90%	82,91%	80,07%	70,14%	63,37%	66,76%	75,46%

Tabla 9.4: Exactitud por palabra, *word accuracy* (%), y mejoras relativas (%) obtenidas para los distintos *sets* (A, B y C) de la base de datos *Aurora* 2 utilizando la técnica de normalización de vectores de características MEMLIN MP, modelando cada uno de los entornos básicos con 128 Gaussianas. A su vez, los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y^e, se representan con 2 componentes. Se ha empleado la parametrización ETSI *advanced* y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*.

en la Figura 9.4 se muestra la mejora media alcanzada cuando se barre el número de componentes: 4, 8, 16, 32, 64 y 128. Igualmente, y a modo de comparación, se han incluido las mejoras medias obtenidas al emplear la parametrización ETSI *advanced* y el método MEMLIN. Obsérvese como el método MEMLIN MP aporta una mejora consistente para todos los casos con respecto a los resultados alcanzados con el algoritmo MEMLIN bajo las mismas condiciones de experimentación.

Al igual que para el caso del método MEMLIN, y a partir de los resultados incluidos en la presente Sección y en 5.5, se puede concluir que la técnica MEMLIN MP proporciona una mejora consistente para las bases de datos *SpeechDat Car* y *Aurora* 2, tanto si se aplica sobre parametrización estándar ETSI o ETSI *advanced*, lo que permite asegurar la robustez del algoritmo MEMLIN MP ante múltiples condiciones de experimentación.

9.4 Resultados Obtenidos con la Técnica Híbrida a Partir del Cálculo de Matrices de Rotación Dependientes de GMMs y MEMLIN MP.

La experimentación realizada en esta Sección sobre la base de datos *Aurora* 2 con la técnica híbrida MEMLIN MP a partir del cálculo de matrices de rotación dependientes de GMMs se rige por los mismos parámetros considerados en las Secciones anteriores.

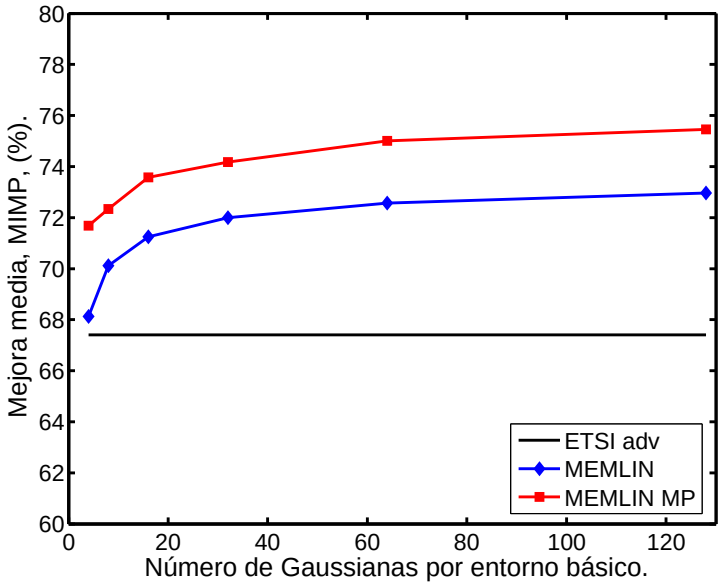


Figura 9.4: Mejoras medias de la exatitud por palabra, *word accuracy* (%), obtenidas con la base de datos *Aurora* 2 utilizando las técnicas de adaptación de vectores de características MEMLIN MP y MEMLIN tras emplear distinto número de componentes para modelar los entornos básicos. A su vez, los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y^e, se representan con 2 componentes para el caso del método MEMLIN MP. Se ha empleado la parametrización ETSI *advanced* y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*. Igualmente, y a modo de comparación, se ha incluido la mejora media obtenida al emplear la parametrización ETSI *advanced*.

De esta manera, los resultados de referencia se pueden consultar nuevamente, y según los casos, en las Tablas 4.5 y 4.7.

9.4.1 Resultados obtenidos con la técnica híbrida MEMLIN MP a partir del cálculo de matrices de rotación dependientes de GMMs y parametrización estándar ETSI

En la Tabla 9.5 se presentan las tasas de exactitud por palabra y mejora relativa alcanzadas tras aplicar la técnica híbrida MEMLIN MP a partir del cálculo de matrices de rotación dependientes de GMMs (MEMLIN MP A). La parametrización empleada en este caso es estándar ETSI y los 24 entornos básicos, así como el espacio limpio, se representan con 128 Gaussianas. Del mismo modo que en casos precedentes, se realizó un barrido con distintos números de componentes: 4, 8, 16, 32, 64 y 128. Para finalizar con los parámetros que rigen la experimentación, cabe destacar que los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y^e, se representan con 2 componentes. Asimismo, se han estimado 16 matrices de rotación entre el espacio normalizado y el limpio, a las que se ha añadido la matriz identidad.

A la luz de los valores presentados en las Tablas 9.5 y 4.5 se puede concluir que, teniendo en cuenta únicamente los mejores resultados para la técnica MEMLIN MP A, el método se comporta satisfactoriamente, obteniendo una mejora media final del

Aurora 2 Small Vocabulary	Clean 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	98,86	98,73	98,87	98,98	98,86	98,86	98,73	98,87	98,98	98,86	98,56	98,67	98,61	98,81
	20 dB	97,94	97,74	98,33	98,15	98,04	97,77	97,50	97,69	98,37	97,83	97,42	96,92	97,17	97,78
	15 dB	97,24	96,90	97,02	97,04	97,05	95,99	96,39	96,24	96,84	96,37	95,89	95,00	95,45	96,46
	10 dB	94,34	93,10	93,66	94,77	93,97	92,14	93,11	93,03	93,23	92,88	89,50	90,27	89,89	92,71
	5 dB	87,74	83,01	82,73	86,29	84,94	81,82	83,70	84,37	84,99	83,72	67,47	76,06	71,76	81,82
	0 dB	68,06	60,41	59,05	68,45	63,99	61,88	62,41	68,55	65,56	64,60	29,38	48,29	38,84	59,20
	-5dB	33,12	33,91	27,63	35,08	32,44	35,78	32,70	41,76	35,76	36,50	12,77	24,66	18,72	31,32
	Average	89,07	86,23	86,16	88,94	**87,60**	85,92	86,62	87,98	87,80	**87,08**	75,93	81,31	**78,62**	85,60

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Detailed relative results in terms of error reduction. Halving the error rate = +50%	Clean	-2,28%	-30,86%	-19,26%	-37,34%	-22,43%	-2,28%	-30,86%	-19,26%	-37,34%	-22,43%	-73,81%	-46,12%	-59,96%	-29,94%
	20 dB	36,76%	76,06%	42,80%	51,31%	51,74%	77,34%	39,56%	77,00%	68,67%	65,64%	61,12%	36,76%	48,94%	56,74%
	15 dB	67,41%	88,85%	73,97%	70,34%	75,14%	84,27%	69,57%	85,63%	83,17%	80,66%	70,60%	54,20%	62,40%	74,80%
	10 dB	76,87%	86,82%	82,62%	81,12%	81,86%	83,66%	79,70%	86,27%	84,89%	83,63%	62,58%	61,03%	61,81%	78,56%
	5 dB	76,72%	77,96%	75,07%	77,49%	76,81%	75,17%	74,41%	79,27%	80,00%	77,21%	34,11%	51,56%	42,83%	70,18%
	0 dB	58,82%	58,09%	54,13%	63,21%	58,56%	58,96%	54,52%	64,86%	61,95%	60,07%	6,42%	32,29%	19,35%	51,33%
	-5dB	25,15%	33,83%	22,33%	30,31%	27,90%	35,16%	26,35%	38,52%	31,55%	32,89%	-0,15%	15,16%	7,51%	25,82%
	Average	63,32%	77,56%	65,72%	68,69%	68,82%	75,88%	63,55%	78,61%	75,74%	73,44%	46,97%	47,17%	47,07%	66,32%

Tabla 9.5: Exatitud por palabra, *word accuracy*, (%) y mejoras relativas (%) obtenidas para los distintos *sets* (A, B y C) de la base de datos *Aurora* 2 utilizando la técnica híbrida MEMLIN MP a partir del cálculo de matrices de rotación dependientes de GMMs. Cada uno de los entornos básicos se modela con 128 Gaussianas y la señal ruidosa asociada a cada par de Gaussianas se representa con 2 componentes. Asimismo se utilizan 16 matrices de rotación más la identidad. Se ha empleado la parametrización estándar ETSI y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*.

66.32 %, y ello a pesar de que sigue habiendo una cierta limitación ante el *set* C, compuesto por distorsiones convolucionales no observadas en la fase de entrenamiento y cuyos resultados están lejos de los alcanzados por los *sets* A y B, que están constituidos a su vez por ruidos aditivos, tanto vistos como no vistos con anterioridad, respectivamente.

En la Figura 9.5 se muestra la mejora media al realizar un barrido del número de componentes con que se modela cada entorno básico. En este caso se presentan los resultados al emplear 4, 8, 16, 32, 64 y 128 Gaussianas. De este modo se pretende observar el comporta del algoritmo MEMLIN MP A al modificar el número de Gaussianas con que se modelan los distintos entornos básicos. Apréciese como para todos los casos la técnica MEMLIN MP A proporciona mejores resultados que los algoritmos MEMLIN y MEMLIN MP. Como viendo siendo habitual, también se puede observar que la mejora obtenida no llega a ser tan elevada como la que se alcanzaba con la base de datos *SpeechDat Car* en español, aunque bien es cierto que las mejores tasas de reconocimiento se dan cuando la SNR se haya comprendida entre 15dB y 5dB, que es el margen entre el que se encuentra principalmente la base de datos *SpeechDat Car* en español.

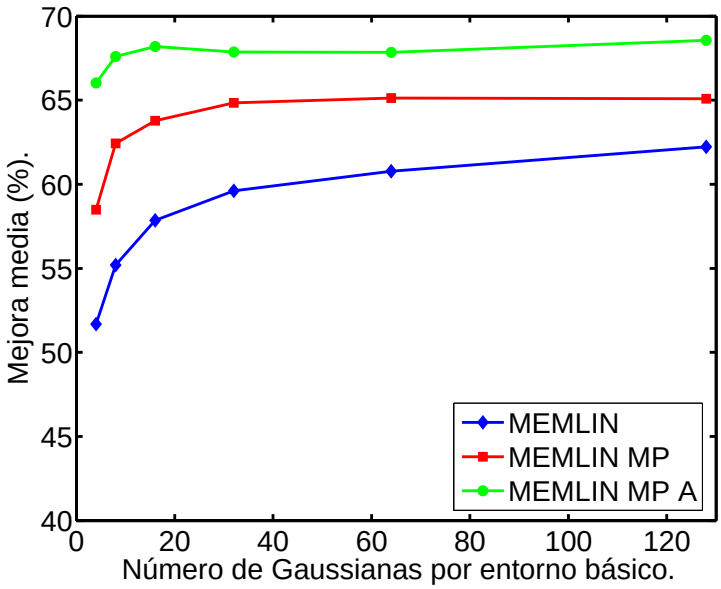


Figura 9.5: Mejoras medias de la exatitud por palabra, *word accuracy* (%), obtenidas con la base de datos *Aurora* 2 utilizando las técnicas MEMLIN, MEMLIN MP y la técnica híbrida a partir del cálculo de matrices de rotación dependientes de GMMs y basada en el algoritmo MEMLIN MP, identificada como MEMLIN MP A. En todos los casos se representan en función del número de Gaussianas por entorno básico empleado. Se ha utilizado la parametrización estándar ETSI y modelos acústicos de palabras generados a partir de la señal limpia *clean training*. A su vez, los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y^e, se representan con 2 componentes para el caso de los métodos MEMLIN MP y MEMLIN MP A, utilizándose además en este último caso 16 matrices de rotación más la identidad.

9.4.2 Resultados obtenidos con la técnica híbrida MEMLIN MP a partir del cálculo de matrices de rotación dependientes de GMMs y parametrización estándar ETSI *advanced*

En la Tabla 9.6 se presentan los resultados que la técnica de compensación MEMLIN MP A proporciona cuando se aplica sobre los coeficientes ETSI *advanced*, incluyendo tanto la exactitud por palabra como la mejora relativa. Del mismo modo que se ha venido realizando, los 24 entornos básicos y el espacio limpio se representan con 128 Gaussianas, aunque en un paso previo se realizó un barrido con distintos números de componentes (4, 8, 16, 32, 64 y 128), y cuyas mejoras medias se pueden observar, de un modo resumido, en la Figura 9.6. Por su parte, los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y^e, se representan en todos los casos con 2 componentes, mientras que se han estimado 16 matrices de rotación entre el espacio normalizado y el limpio, a las que se ha añadido la matriz identidad.

A la luz pues de los valores presentados en las Tablas 9.6 y 4.7 se puede concluir que, teniendo en cuenta únicamente los mejores resultados para la técnica MEMLIN MP, el método proporciona una mejora añadida (75.16 %) a la ya obtenida por la parametrización ETSI *advanced* (67.41 %). Cabe destacar, por otra parte, como el comportamiento del método ante los distintos *sets* es similar al que ya se había podido apreciar para las técnicas MEMLIN y MEMLIN MP, siendo el *set* C el que produce una mejora menos

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restauran	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	99,17	98,85	99,25	99,38	99,17	99,17	98,85	99,25	99,38	99,17	98,96	99,03	98,99	99,13
	20 dB	98,31	98,61	98,66	98,40	98,49	98,16	97,98	98,66	98,89	98,42	98,13	97,86	97,99	98,36
	15 dB	97,33	97,47	98,06	97,05	97,48	97,13	97,16	97,68	97,53	97,38	97,12	96,89	97,01	97,34
	10 dB	94,84	94,69	96,18	95,10	95,20	94,62	94,26	94,69	95,91	94,87	94,07	93,09	93,58	94,75
	5 dB	89,05	87,16	91,22	89,22	89,16	86,58	87,35	88,13	89,65	87,93	86,54	82,80	84,67	87,77
	0 dB	73,40	64,56	76,00	72,89	71,71	66,52	68,90	71,40	72,59	69,85	65,03	59,10	62,06	69,04
	-5dB	41,81	32,39	40,62	42,67	39,37	35,60	37,70	39,14	42,69	38,78	33,66	29,91	31,78	37,62
	Average	90,59	88,50	92,02	90,53	90,41	88,60	89,13	90,11	90,91	89,69	88,18	85,95	87,06	89,45

Aurora 2 Small Vocabulary	Clean training, multicondition testing													
		A					B					C			
		Subway	Babble	Car	Exhibition	Average	Restauran	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Detailed relative results in terms of error reduction. Halving the error rate = +50%	Clean	25,34%	-18,32%	21,54%	16,66%	11,31%	25,34%	-18,32%	21,54%	16,66%	11,31%	-25,77%	-6,27%	-16,02%	5,84%
	20 dB	48,09%	85,31%	54,05%	57,83%	61,32%	81,35%	51,12%	86,65%	78,68%	74,45%	71,75%	55,97%	63,86%	67,08%
	15 dB	68,52%	90,91%	83,07%	70,36%	78,22%	88,74%	76,06%	91,14%	86,86%	85,70%	79,39%	71,52%	75,45%	80,66%
	10 dB	78,92%	89,86%	89,54%	82,33%	85,16%	88,82%	83,08%	89,53%	90,88%	88,08%	78,85%	72,34%	75,60%	84,41%
	5 dB	79,21%	83,35%	87,32%	82,31%	83,05%	81,67%	80,15%	84,26%	86,21%	83,07%	72,73%	65,20%	68,96%	80,24%
	0 dB	65,71%	62,49%	73,12%	68,39%	67,43%	63,96%	62,37%	68,05%	69,71%	66,02%	53,66%	46,43%	50,05%	63,39%
	-5dB	34,87%	32,31%	36,27%	38,45%	35,47%	34,99%	31,82%	35,75%	38,94%	35,37%	23,83%	21,08%	22,45%	32,83%
	Average	68,09%	82,38%	77,42%	72,25%	75,03%	80,91%	70,56%	83,93%	82,47%	79,46%	71,27%	62,29%	66,78%	75,16%

Tabla 9.6: Exatitud por palabra, *word accuracy*, (%) y mejoras relativas (%) obtenidas para los distintos *sets* (A, B y C) de la base de datos *Aurora* 2 utilizando la técnica híbrida MEMLIN MP a partir del cálculo de matrices de rotación dependientes de GMMs. Cada uno de los entornos básicos se modela con 128 Gaussianas y la señal ruidosa asociada a cada par de Gaussianas se representa con 2 componentes. Asimismo se utilizan 16 matrices de rotación más la identidad. Se ha empleado la parametrización ETSI *advanced* y modelos acústicos de palabras generados a partir de la señal limpia, *clean training*.

satisfactoria. Por su parte, los *sets* A y B alcanan valores de reconocimiento muy próximos.

Ya para finalizar, y de cara a observar como se comporta el algoritmo MEMLIN MP A cuando se incrementa el número de Gaussianas con que se modelan los distintos entornos básicos, en la Figura 9.4 se muestra la mejora media obtenida tras realizar el siguiente barrido del número de componentes: 4, 8, 16, 32, 64 y 128. A modo de comparación, se han incluido igualmente las mejoras medias alcanzadas al emplear la parametrización ETSI *advanced* y los métodos MEMLIN y MEMLIN MP. Obsérvese como el comportamiento del método MEMLIN MP A es muy similar al de la técnica MEMLIN MP, incluso peor en algunas situaciones, de modo que se puede concluir que, para este caso concreto, la introducción de las matrices de rotación no aporta mejora alguna. A pesar de ello, y a partir de los resultados incluidos en la presente Sección y en 5.5, sí se puede afirmar que la técnica MEMLIN MP A proporciona una mejora consistente para las bases de datos *SpeechDat Car* y *Aurora* 2, tanto si se aplica sobre parametrización estándar ETSI o ETSI *advanced*, lo que permite asegurar la robustez del algoritmo ante múltiples condiciones de experimentación.

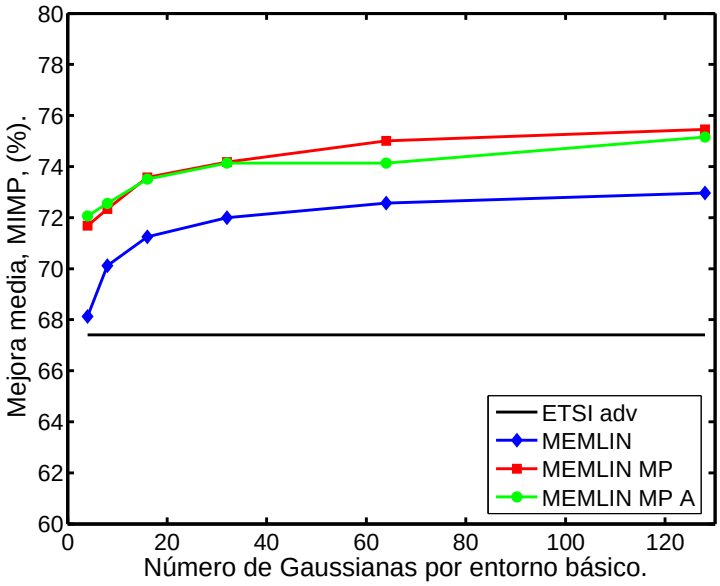


Figura 9.6: Mejoras medias de la exatitud por palabra, *word accuracy* (%), obtenidas con la base de datos *Aurora* 2 utilizando las técnicas MEMLIN, MEMLIN MP y la técnica híbrida a partir del cálculo de matrices de rotación dependientes de GMMs y basada en el algoritmo MEMLIN MP, identificada como MEMLIN MP A. En todos los casos se representan en función del número de Gaussianas por entorno básico empleado. Se ha utilizado la parametrización ETSI *advanced* y modelos acústicos de palabras generados a partir de la señal limpia *clean training*. A su vez, los vectores de características ruidosos asociados a cada par de Gaussianas, s_x y s_y^e, se representan con 2 componentes para el caso de los métodos MEMLIN MP y MEMLIN MP A, utilizándose además en este último caso 16 matrices de rotación más la identidad.

Capítulo 10

Resultados con la Base de Datos *Hiwire*

10.1 Introducción.

Del mismo modo que sucede con la base de datos *Aurora* 2, el hecho de que la parte ruidosa del corpus *Hiwire* se genere tras añadir artificialmente ruido aditivo supone, de cara a experimentaciones de RAH, una importante rémora puesto que se desestima el efecto Lombard.

Con la intención de realizar posteriormente un análisis sobre los resultados obtenidos, puede resultar interesante recordar algunos de los parámetros del corpus *Hiwire* ya introducidos en el Capítulo 4. Así pues, se insiste en que está compuesta por 8100 alocuciones en inglés pronunciadas por locutores no nativos (franceses, griegos, italianos y españoles), siendo la tarea en cuestión comandos de aviación. Dichas grabaciones se realizaron en estudio y, posteriormente, se añadió el ruido registrado en la cabina de un avión a lo largo de un vuelo ordinario. Se distinguieron tres condiciones acústicas según el volumen del mismo ruido, a saber: bajo o *low*, medio o *medium* y alto o *high*, que se corresponden aproximada y respectivamente con 10dB, 5dB y -5dB de SNR.

Aunque la experimentación definida en la Sección 4.4 incluye dos modos: *Robust Non-Native*, NNA, y *Non-Native Adaptation*, NNA, en este trabajo únicamente se presentan los resultados obtenidos para el segundo de ellos, puesto que es el que mejor se adapta a la problemática de robustez considerada en la presente tesis doctoral. Asimismo, hay que considerar que en este caso no se realizó, como ha sucedido en experimentaciones anteriores, un barrido de los distintos parámetros que definen las diferentes técnicas, sino que directamente se trabajó sobre aquéllos que, hasta la fecha, habían proporcionado los mejores resultados. De la misma manera, únicamente se empleó la técnica híbrida basada en reentrenamiento supervisado y MEMLIN MP. La razón de esta reducción de la experimentación reside en que el corpus *Hiwire* se presentó como banco de pruebas para una competición de robustez en una de las sesiones especiales del congreso *Interspeech* 2007, con la consecuente escasez de tiempo para presentar los correspondintes resultados [BMS+07]. Por ello, en la siguiente y última Sección del Capítulo (10.2), se incluye el sistema presentado, así como los resultados obtenidos con los parámetros óptimos que la experiencia previa había determinado.

10.2 Resultados Obtenidos con la Técnica Híbrida a Partir de Reentrenamiento Supervisado y MEMLIN MP.

Dado que, tal y como se ha adelantado en la Sección 10.1, los resultados extraídos con la base de datos *Hiwire* estaban destinados a una competición, se desarrolló un único sistema con aquellos parámetros que habían proporcionado los mejores resultados hasta el momento, obviando por tanto cualquier tipo de barrido de los mismos [BMS+07]. Por ello, se propuso una técnica híbrida basada en reentrenamiento supervisado, donde el correspondiente algoritmo de adaptación de vectores de características fuera MEMLIN MP. En la Figura 10.1 se incluyen los distintos bloques de que se compone el sistema finalmente empleado. Obsérvese que formalmente es el mismo que se propuso en la Sección 8.3, con la salvedad de que en este caso la adaptación de los modelos acústicos se realiza a partir de la técnica MLLR. De esta manera, los distintos parámetros necesarios para evaluar la técnica MEMLIN MP se obtienen en "Entrenamiento MEMLIN MP", y son utilizados en "Normalización MEMLIN MP" para dar lugar a la señal de entrenamiento normalizada, que será empleada posteriormente para calcular los correspondientes modelos acústicos mediante el algoritmo MLLR, "Adaptación MLLR". Por su parte, y ya en la fase de decodificación, los vectores de características normalizados se decodifican con los susodichos modelos acústicos adaptados, "Decodificación". Hay que tener en cuenta, y ésta es una diferencia importante con respecto a las experimentaciones realizadas con anterioridad, que las adaptaciones, ya sean de los vectores de características o de los modelos acústicos, son dependientes de cada condición acústica y locutor, lo que hipotéticamente le confiere una mayor robustez al sistema final propuesto.

La extracción de características previa a la evaluación del algoritmo MEMLIN MP incluye, como viene siendo habitual, 12 coeficientes estáticos más la energía, en este caso provinientes de la parametrización ETSI estándar. Por su parte, y una vez normalizadas, las tramas se completarán, de cara a la decodificación, con 3 derivaradas calculadas a partir de 9 vectores acústicos estáticos. En cuanto a la fase de entrenamiento, indicar que se entrenarán distintos parámetros para cada locutor y condición acústica haciendo uso de 50 alocuciones. Así, y en cuanto a la técnica MEMLIN MP, conviene comentar que se emplean GMMs de 128 componentes para modelar el espacio limpio y cada uno de los ruidosos, mientras que la GMM que representa la señal degradada asociada a cada par de componentes se construye, como viene siendo habitual, con dos Gaussianas. La adaptación de los modelos acústicos se realiza a partir de la técnica MLLR empleando las susodichas 50 frases ruidosas previamente compensadas y un árbol de 32 clases de regresión, de modo que para cada una de ellas la transformación de los modelos acústicos sea la misma; como modelos acústicos iniciales se toman los entrenados con la base de datos TIMIT ya presentados en la Sección 4.4.

Con todo lo anterior, y tras hacer uso del sistema previamente comentado, en la Tabla 10.1 se presentan las tasas de error por palabra, WER, obtenidas para las distintas nacionalidades de los locutores y condiciones acústicas, donde "MWER" es el WER medio. A la luz de dichos resultados, y comparándolos con los de referencia, ya presentados en la Tabla 4.10, se puede observar claramente una importante mejora

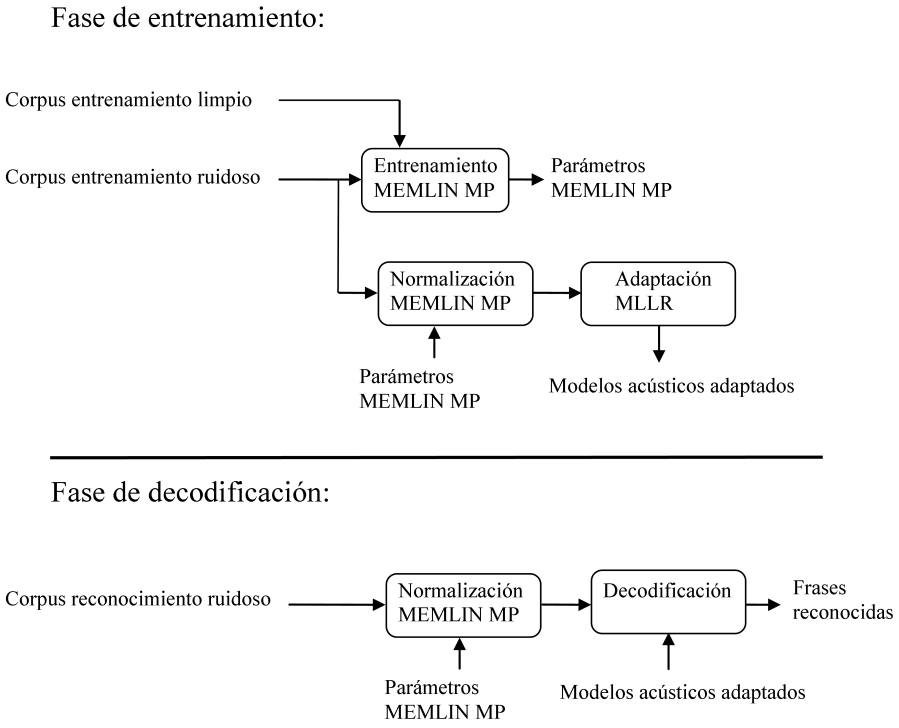


Figura 10.1: Esquema gráfico de la técnica híbrida basada en reentrenamiento supervisado con la que se realizaron los experimentos con el corpus *Hiwire*. Se incluye tanto la fase de entrenamiento como la de decodificación. La primera de ellas está compuesta por tres bloques. El primero de ellos, "Entrenamiento MEMLIN MP", obtiene los distintos parámetros necesarios para la correspondiente técnica de normalización. Por su parte, el sistema de "Normalización MEMLIN MP" proporciona la estimación de los vectores de características limpios a partir de los degradados. El bloque "Adaptación MLLR" calcula los nuevos modelos acústicos asociados al espacio normalizado a partir de los limpios y de la señal del corpus de entrenamiento degradado previamente compensada. Dichos modelos son los empleados para reconocer los vectores de características normalizados en el bloque identificado como "Decodificación"

debido a la inclusión de la técnica MEMLIN MP (recuérdese que las tasas incluidas en la Tabla 4.10 se habían obtenido tras aplicar el método MLLR). Sin embargo, el comportamiento no es del todo homogéneo para los tres niveles de ruido, de manera que, si bien se obtiene una muy importante mejora para los dos más benignos (77.19 % y 75.38 %, respectivamente), no ocurre lo mismo con el tercero, siendo en este caso la mejora relativa sensiblemente menor (42.98 %). Aun así, se puede concluir que el sistema propuesto proporciona una importante ganancia, prueba de lo cual es que obtuvo los mejores resultados en la competición.

Además de las conclusiones directas que se hayan podido extraer tras los resultados presentados, cabe destacar que se puede empezar a intuir asimismo que las técnicas incluidas en este trabajo también pueden ser válidas ante tareas más complejas que los dígitos, aunque para poderlo asegurar sin nigún tipo de dudas haya que trabajar previamente con corpora de gran vocabulario, algo que se considerará como una futura línea de investigación. Del mismo modo, la utilización de las técnicas presentadas en esta tesis doctoral para aplicaciones de adaptación al locutor también se vislumbra otro posible trabajo venidero.

RNN	Francés	Griego	Italiano	Español	MWER
Bajo	6.65	6.74	6.61	5.70	6.54
Medio	14.76	15.60	11.19	10.40	13.52
Alto	57.96	60.71	45.95	40.62	53.45

Tabla 10.1: Resultados en términos de WER (%) obtenidos para el modo *Non-Native Adaptation*, NNA, de la base de datos *Hiwire*, para los distintos niveles de ruido (bajo, medio y alto). Se utiliza una técnica híbrida basada en reentrenamiento supervisado en la que se combinan los algoritmos MEMLIN MP y MLLR. Se ha empleado la parametrización ETSI estándar y modelos acústicos fonéticos para cada locutor y condición acústica.

Capítulo 11

Conclusiones y Líneas Futuras de Trabajo.

11.1 Introducción.

Durante los algo más de cuatro años que se han necesitado para completar este trabajo, a la vez que se iban cumpliendo 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 presentadas en esta tesis, así como las líneas de actuación futuras sobre las que seguir desarrollando soluciones de cara a obtener cada vez métodos de adaptación más robusto ante cualquier tipo de efecto producido por el entorno acústico.

Por todo lo anterior, el presente Capítulo se haya dividido en dos grandes unidades. Primeramente, y en la Sección 11.2, 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 11.3 se hace hincapié en aquellas debilidades de las técnicas propuestas que, aun observadas y constatadas durante el desarrollo de las mismas, no se trataron, dejando así abiertas las puertas para futuras investigaciones.

11.2 Conclusiones.

El que las prestaciones de los sistemas de RAH decaen ante los efectos producidos por los 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, lo que, sin duda, ayudará al desarrollo posterior de técnicas de robustez más eficaces. En este sentido en el Capítulo 4 y sucesivos se han presentado tanto los resultados de RAH, como distintos histogramas y *log-scattergrams* obtenidos a partir de las señales ruidosas de las bases de datos *SpeechDat Car* en español, *Aurora* 2 y *Hiwire*, de manera que, a partir de todo lo anterior, se ha podido constatar que el ruido

propio del entorno acústico produce serias alteraciones tanto en la media como en la varianza de los coeficientes MFCC de los vectores de características; y esto incluso para aquellos entornos acústicos más benévolos y controlados, como por ejemplo en los que incluyen únicamente una ligera distorsión convolucional. Así, se ha podido comprobar como el ruido aditivo principalmente introduce una alteración de las varianzas de los coeficientes MFCC de los vectores acústicos, mientras que la distorsión convolucional se manifiesta principal, pero no únicamente, en un desplazamiento de la media de dichos parámetros. Por su parte, los entornos reales suelen incluir ambos efectos de una manera conjunta, además de introducir ciertas modificaciones en la señal de voz por el simple hecho de encontrarse el locutor en un ambiente hostil (efecto Lombard). También se ha observado que la aleatoriedad del ruido introduce incertidumbre en los coeficientes de los vectores acústicos, de manera que, para una realización limpia dada, se pueden tener varias ruidosas y viceversa, lo que supone el mayor reto para las técnicas de adaptación de vectores de características ya que éstas se suelen sustentar en la aplicación de una función de transformación dependiente de las tramas ruidosas. Así pues, todas estas alteraciones propias del entorno acústico producen, consecuentemente, una cierta degradación en el comportamiento del sistema de RAH que puede llegar a hacer inviable su uso en multitud de situaciones.

Una vez analizado el problema que se pretende compensar, y tras estudiar las distintas soluciones que a lo largo del tiempo ha ido desarrollando la comunidad científica, en el Capítulo 5 se estudiaron las técnicas de adaptación de vectores de características empíricas más empleadas en la actualidad: CMS, RATZ y SPLICE, presentándolas mediante un desarrollo teórico conjunto, de manera que se pudo comprobar que la única diferencia entre ellas, desde un punto de vista matemático, reside en la aplicación de ciertas aproximaciones. Así pues, y dejando a un lado el método CMS por ser el más simple de los tres, el algoritmo RATZ propone modelar el espacio limpio mediante una GMM, para estimar posteriormente, y a modo de transformación de compensación, un vector de desplazamiento para cada Gaussiana. Por su parte, el método SPLICE representa el espacio degradado mediante una GMM, obteniendo igualmente un vector de desplazamiento para cada componente. En ambos casos el vector acústico normalizado se calcula a partir del degradado haciendo uso de todos los vectores de desplazamiento y del criterio MMSE. Estas transformaciones dependientes de las distintas componentes no son, como se ha podido demostrar, todo lo específicas que se podría desear, ya que si se considera la GMM que representa el espacio ruidoso como un modelo de generación, se ha podido observar que los vectores de características producidos por una Gaussiana del espacio ruidoso tienen asociados una serie de tramas limpias que, en general, y debido a la aleatoriedad del ruido del entorno acústico, no se encuentran concentradas en una región específica del espacio limpio, sino que se distribuyen en mayor o menor medida por todo él. La misma conclusión se extrae si se modela el espacio limpio mediante una GMM y se supone un determinado entorno acústico ruidoso real. A partir de ello se puede concluir 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. Para compensar este hecho se desarrolló el algoritmo MEMLIN, que modela mediante sendas GMMs tanto el espacio limpio como el ruidoso. Así, en esta ocasión los vectores de transformación se asocian a cada par de Gaussianas,

reduciendo consecuentemente el rango de proyección de los vectores acústicos degradados a nivel de componente y no de espacio, como sucedía con las técnicas RATZ y SPLICE. Los resultados obtenidos con la base de datos *SpeechDat Car* en español muestran el correcto funcionamiento de la técnica propuesta, obteniéndose una mejora media de 70.22 % con modelos acústicos de fonemas, superior que las alcanzadas con las técnicas IRATZ (61.84 %) y SPLICE ME (65.39 %), siendo ambas las versiones multi-entorno de los métodos RATZ y SPLICE.

Tras un análisis en profundidad de la técnica MEMLIN, se detectaron principalmente dos aproximaciones que afectan en gran medida al comportamiento final de la técnica: por una parte la elección del modelo simplificado del espacio de señal, que presupone en esta ocasión 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 modificando la media de los vectores acústicos mediante un vector de desplazamiento. La segunda aproximación consiste en presuponer que el modelo de la probabilidad condicionada entre espacios de señal es independiente del vector de características ruidoso. Esto hace que la probabilidad de una determinada Gaussiana de la GMM del espacio limpio dada otra del modelo de la GMM del espacio ruidoso sea en todo momento la misma, independientemente del vector acústico degradado, lo que no deja de ser una aproximación. Por todo ello se desarrollaron diversas soluciones para tratar de compensar ambas limitaciones. Así, la primera de ellas se estudió de un modo directo en el Capítulo 6, mientras que en el Capítulo 7 se generó un modelo de la probabilidad condicionada entre espacios de señal más realista mediante una solución basada en GMMs.

El considerar que el modelo del espacio de señal para cada par de Gaussianas es lineal con término dependiente unitario presupone que el entorno acústico únicamente afecta, para dicho par de componentes, a las medias de los vectores de características y no así a las correspondientes varianzas, lo que, en general, es aproximadamente cierto para distorsión convolucional, pero no así para ruido aditivo. Es por ello por lo que se propuso una serie de nuevos modelos del espacio de señal en los que las transformaciones asociadas a cada par de componentes fueran algo más complejas. De esta manera, y en aras de adecuarse en mayor medida a la realidad, se propuso tanto un modelo lineal con término dependiente no unitario, lo que dio lugar a la técnica P-MEMLIN, como un esquema no lineal basado en ecualización de histograma, lo que generó el algoritmo MEMHIN. Asimismo, y ya para concluir las modificaciones sobre el modelado del espacio de señal, 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, denominada PD-MEMLIN.

A partir de la experimentación llevada a cabo se pudo observar que, si bien las técnicas P-MEMLIN y MEMHIN apenas si aportan mejora alguna sobre la base de datos *SpeechDat Car* en español con respecto al algoritmo MEMLIN cuando se modelan los entornos básicos y el espacio limpio con un número elevado de Gaussianas (70.47 % y 70.22 % de MIMP, respectivamente), ambas sí poseen un mejor comportamiento cuando el número de componentes es reducido, debido al importante papel que en dicha situación juega la normalización de varianza propuesta por ambos métodos. Así, por ejemplo, si se

aplica la técnica MEMLIN modelando cada entorno básico únicamente 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 superiores, 51.53 % y 49.84 %, respectivamente; esto es debido a que, en dichos casos, el espacio representado por cada Gaussiana es más heterogéneo 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 componentes. De esta manera, un modelo más complejo de $\mathbf{x}$ proporciona interesantes mejoras. A su vez también se pudo comprobar que, no sólo el número de componentes con que se modelan los entornos básicos es importante a la hora de justificar el uso de técnicas como P-MEMLIN o MEMHIN, sino que también lo es la naturaleza del ruido. Se ha podido demostrar que ambos algoritmos proporcionan una interesante mejora con respecto al método MEMLIN ante ruido aditivo ya que en ese caso se produce, como se pudo comprobar en la Sección 5.2, una seria degradación de la varianza de los coeficientes de los vectores de características, lo que se adecúa mejor a la transformación propuesta por los algoritmos P-MEMLIN y MEMHIN, tal y como se pudo constatar mediante la consiguiente experimentación presentada en la Sección 6.6. Así pues, se puede concluir que, aunque para la base de datos *SpeechDat Car* en español las técnicas P-MEMLIN y MEMHIN no aportan importantes mejoras con respecto al algoritmo MEMLIN, no hay que desechar estas dos técnicas, puesto que se ha constatado que, bajo ciertas condiciones de experimentación, sí pueden aportar una interesante mejora.

Por su parte, la técnica PD-MEMLIN, como último método que modifica el modelo del espacio de señal, proporciona, independientemente del número de transformaciones por entorno básico que se empleen y para la base de datos *SpeechDat Car* en español, una significativa mejora relativa con respecto al algoritmo MEMLIN (75.44 % de MIMP para el mejor de los casos). Asimimso se pudo comprobar que este algoritmo soluciona, al forzar el número de componentes con que se modela cada fonema, uno de los problemas del método MEMLIN. Éste no es otro que la proyección de gran cantidad de vectores de características ruidosos hacia el silencio del espacio limpio. Obsérvese que mediante la técnica PD-MEMLIN se reduce 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 que posteriormente se emplearán a la hora de la decodificación.

Una vez tratadas las limitaciones que el modelado del espacio de señal considerado para la técnica MEMLIN posee, y tras proporcionar, como se ha visto, tres nuevas soluciones cuyas mejoras, más o menos restrictivas, quedaron patentes, se estudió, como segunda gran línea de actuación sobre la técnica MEMLIN, el modelado de la probabilidad condicionada entre espacios de señal, término este de gran importancia puesto que determina, a nivel de Gaussiana, el entorno de proyección del vector de características ruidoso sobre el espacio limpio y, por tanto, el nivel de incertidumbre en el que se puede mover el vector de características normalizado, que consecuentemente estará en función de las varianzas de las Gaussianas que modelan el espacio limpio. Si bien el método MEMLIN considera la probabilidad entre Gaussianas independiente del vector acústico degradado, obteniendo aún así un comportamiento satisfactorio, pruebas preliminares determinaron que es en dicho término donde radica la verdadera capacidad de normalización de la técnica MEMLIN, pudiéndose alcanzar, en un caso hipotético, una compensación casi

perfecta. Por todo ello, para compensar las limitaciones mostradas por el modelado de la probabilidad entre Gaussianas considerado en las técnicas presentadas hasta el momento, se propuso representar los vectores de características ruidosos asociados a cada par de Gaussianas mediante una GMM. De esta manera se elimina la independencia temporal considerada anteriormente, creando una solución mucho más dinámica que proporciona un mejor comportamiento tanto si se aplica al método MEMLIN, como al PD-MEMLIN, dando lugar a los algoritmos MEMLIN MP y PD-MEMLIN MP (78.48 % y 77.72 % de mejora media con la base de datos *SpeechDat Car* en español, respectivamente). Por otra parte, cabe destacar que las mejoras con respecto a las correspondientes versiones que no aplican el modelado entre Gaussianas basado en GMMs es mayor en aquellos casos en los que se representan los entornos básicos y el espacio limpio con un número reducido de Gaussianas, lo que es debido a que el espacio de proyección de los vectores de características asociados a una determinada componente es mucho más heterogéneo en ese caso. No obstante, y a pesar de los satisfactorios comportamientos de las técnicas MEMLIN MP y PD-MEMLIN MP, también es cierto que en ambos métodos se requiere de un mayor coste computacional derivado del incremento de *scores* que se han de evaluar. Para reducirlo en la medida de lo posible se decidió considerar únicamente aquellos pares de Gaussianas más probables, haciendo que la probabilidad entre Gaussianas para el resto fuera nula. Esto permite una elevada reducción en el número de *scores* evaluadas sin que las tasas de RAH se vieran seriamente comprometidas.

Desde un primer momento se tuvo consciencia de las limitaciones que las técnicas de adaptación de vectores de características poseen debido a que emplean funciones de transformación que, en general, no pueden compensar perfectamente la incertidumbre entre los vectores acústicos introducida por la aleatoriedad propia del ruido, al contrario de lo que sucede con los métodos de adaptación de modelos acústicos, que sí pueden tratarla de mejor modo. Por ello se propuso en el Capítulo 8, la combinación de las técnicas de adaptación más características que se habían propuesto hasta ese momento en el trabajo con métodos de adaptación de modelos acústicos, dando lugar a una serie de métodos híbridos que pueden ser supervisados o no supervisados, según si se emplean las trascripciones de las señales del corpus de entrenamiento, o no. En cualquier caso, la razón que subyace en ambos tipos de técnicas es que la proyección de los vectores acústicos degradados que proporcionan las distintas técnicas de normalización no es perfecta, de modo que surge un nuevo espacio normalizado que no queda perfectamente representado por los modelos acústicos del espacio limpio.

En las técnicas híbridas no supervisadas presentadas, a cada vector de características normalizado (se presentaron soluciones con los métodos MEMLIN y MEMLIN MP) se le asoció una matriz de rotación en el proceso de decodificación de entre un conjunto de posibles matrices estimadas previamente en una fase de entrenamiento. Dicha matriz de rotación, seleccionada mediante el criterio ML y con la que se modificarán los vectores de medias y las matrices de covarianzas de los modelos acústicos de referencia, estará asociada a un determinado par de Gaussianas, entendiendo por par la unión de una componente del modelo del espacio limpio y otra del modelo con que se representa el espacio normalizado. De esta manera se ha podido constatar que las correspondientes soluciones híbridas, que en el fondo se pueden ver igualmente como unos métodos de adaptación de modelos acústicos no supervisados, alcanzan unas importantes mejoras cuando se aplican sobre la base de datos *SpeechDat Car* en español y modelado acústico

de palabras: 90.54 % y 92.07 % cuando la técnica de normalización correspondiente es MEMLIN o MEMLIN MP, respectivamente (MEMLIN y MEMLIN MP bajo esas mismas condiciones obtienen unas mejoras medias de 75.79 % y 83.89 %). Asimismo, resulta interesante observar como en este caso los resultados son prácticamente independientes del número de Gaussianas con que se modelan los distintos entornos básicos, lo que supone una importante mejora de cara a obtener un comportamiento satisfactorio sin tener que recurrir a un elevado coste computacional. Por su parte, la comparación con respecto a técnicas de adaptación de modelos acústicos no supervisadas también es igualmente satisfactoria. Así, el método MLLR no supervisado obtiene una mejora media de 78.77 %, aún lejos de las obtenidas con las técnicas híbridas propuestas. De esta manera, y a partir de las distintas experimentaciones realizadas, se puede constatar como, si bien el desplazamiento de los vectores de características es, en la mayoría de los casos, el efecto más importante que introduce el entorno acústico, no es el único, de modo que el compensar convenientemente otros efectos, como las rotaciones o ciertas dependencias entre coeficientes de los vectores de características, también puede permitir alcanzar una importante mejora.

Si se dispone de la trascripción de la señal de entrenamiento es posible definir técnicas híbridas supervisadas. De este modo es posible entrenar directamente modelos acústicos del espacio normalizado a partir de los vectores de características ruidosos compensados y de los modelos acústicos limpios. Para ello se puede hacer uso del algoritmo ML, si se dispone de suficiente señal, como se considera en este trabajo, o, en caso contrario, de métodos de adaptación de modelos acústicos clásicos, como MLLR, MAP... Así pues, se puede constatar que, si se aplica esta técnica híbrida a la base de datos *SpeechDat Car* en español, 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 MP. Las mejoras medias alcanzadas, sensiblemente más satisfactorias que las logradas cuando se reentrenan los modelos acústicos con la señal ruidosa (81.93 % de mejora media), son debidas a que tras compensar la señal ruidosa, el espacio generado es mucho más compacto y homogéneo, haciendo que el entrenamiento con el criterio ML sea más satisfactorio que si se realizara directamente sobre el entorno ruidoso, siempre mucho más heterogéneo. Asimismo es importante constatar como los resultados presentados con este tipo de técnicas híbridas son prácticamente insensibles al número de Gaussianas con que se modelan los distintos entornos básicos, lo que permite lograr interesantes resultados a partir de una rápida normalización de los vectores de características.

Llegados a este punto, dos son las grandes dudas que podían planear sobre las técnicas de adaptación de vectores de características empíricas propuestas. En primer lugar, el tener que disponer de un corpus de entrenamiento estéreo para obtener los distintos parámetros que definen las diversas técnicas, lo que no siempre es posible; y en segundo lugar, la posible falta de robustez de los algoritmos a la hora de normalizar señal perteneciente a entornos acústicos no considerados en la fase de entrenamiento. Para dar respuesta a la primera duda se definió una fase de entrenamiento "ciega", tal y como se presentó en la Sección 6.5, mientras que para la segunda observación se incluyó una exhaustiva experimentación con la base de datos *Aurora* 2 en el Capítulo 9.

La mayoría de las técnicas de adaptación de vectores de características empíricas precisan, al menos en su desarrollo más directo, de una fase de entrenamiento con señal estéreo y esto, en algunas ocasiones, no es posible debido a la naturaleza de la propia aplicación o la correspondiente base de datos. Para evitar dicha limitación se desarrolló una fase de entrenamiento "ciega" para la técnica PD-MEMLIN en la que sólo es necesaria la señal degradada. Dicha fase de entrenamiento "ciega" se sustenta en la aplicación iterativa de señal pseudo-estéreo obtenida tras normalizar los vectores acústicos degradados con la técnica de compensación KPD-MEMLIN utilizando en cada caso aquellos parámetros estimados en la iteración previa. Por su parte, y en la primer estadio de la fase de entrenamiento, también se hace uso del algoritmo EM, así como de la distancia de Kullback-Liebler modificada a tal efecto. Los resultados obtenidos con la base de datos *SpeechDat Car* en español muestran que la técnica PD-MEMLIN con fase de entrenamiento "ciega" es capaz de proporcionar unos resultados similares, e incluso en algún caso superiores, a los logrados por el método MEMLIN (72.40 % de MIMP para el mejor de los casos), 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. Así pues, y aunque las mejoras obtenidas con esta técnica aún quedan algo lejos del mejor resultado alcanzado por la técnica PD-MEMLIN con fase de entrenamiento con señal estéreo, se puede considerar que con el desarrollo presentado se cierra una de las grandes limitaciones que las técnicas de adaptación de vectores acústicos empíricas posee.

Otra de las grandes dudas que sobrevuela siempre sobre las técnicas de adaptación de vectores acústicos empíricas es, dado que el comportamiento de las mismas se basa principalmente en la fase de entrenamiento, hasta qué punto son robustas ante entornos acústicos no observados. Para analizar este hecho, así como para comprobar el comportamiento de los algoritmos presentados con una base de datos utilizada normalmente como estándar de comparación, se presentaron en el Capítulo 9 los resultados alcanzados con la base de datos *Aurora* 2 tras aplicar los métodos MEMLIN, MEMLIN MP y la técnicas híbridas no supervisadas basada en matrices de rotación a partir del algoritmo MEMLIN MP. En todos los casos, tanto si se aplicaron sobre la parametrización estándar ETSI o ETSI *advanced*, se obtuvieron importantes mejoras: 66.32 % y 75.46 %, respectivamente en el mejor de los casos. Sin embargo, el comportamiento de las diferentes técnicas, que es satisfactorio para los *sets* A y B, no lo es tanto para el *set* C, por lo que se puede concluir que los métodos con los que se experimentó son robustos ante ruido aditivo no observado en la fase de entrenamiento (*set* B), mientras que son algo más sensibles ante distorsión convolucional no vista en la fase de entrenamiento (*set* C). Asimismo es interesante observar como el comportamiento constatado con la base de datos *Aurora* 2 no llega al nivel del alcanzado con el corpus *SpaachDat Car* en español, aunque sí es cierto que ante SNR comprendidas entre 5dB y 15dB, que coincide con las que se dan en la base de datos *SpeechDat Car*, sí se consiguen importantes mejoras.

Por otra parte, y a partir de los resultados obtenidos en el Capítulo 10, se pudo observar que las técnicas de adaptación de vectores de características propuestas en este trabajo, más concretamente MEMLIN MP, proporcionan un más que satisfactorio comportamiento ante tareas sensiblemente más complejas que los dígitos, abriendo de este modo dos caminos no explorados hasta el momento: el empleo de los métodos presentados en esta tesis en tareas de gran vocabulario, lo que terminaría de cerrar

definitivamente la experimentación, y la adaptación conjunta al locutor y entorno acústico, cuyos prometedores resultados incluidos en el Capítulo 10 permiten aventurar que, para un caso genérico, dicha adaptación conjunta podría aportar una mejora añadida.

Ya por último, hay que destacar el hecho de que la mayoría de las técnicas presentadas son no supervisadas, lo que resulta especialmente interesante por cuanto independiza la fase de entrenamiento no sólo de una hipotética trascripción, sino también a la tarea de reconocimiento en cuestión, lo que, en muchas aplicaciones reales, podría permitir la realización de la fase de entrenamiento de un modo sencillo, cómodo y barato.

11.3 Líneas Futuras de Trabajo.

A lo largo del desarrollo de la presente tesis, varios han sido los proyectos que, por falta de tiempo o prioridad, no han sido completados. Sin embargo no han caído en el olvido y, dado que algunos actualmente se siguen considerando interesantes, constituyen las líneas futuras de trabajo, aunque bien es cierto que ya se está trabajando en alguno de dichos proyectos.

A partir de los resultados obtenidos al aplicar el método KPD-MEMLIN y tras comprobar en trabajos preliminares el potencial de la técnica PD-MEMLIN en tareas de verificación e identificación de locutor, es directo plantear la posibilidad de emplear el algoritmo KPD-MEMLIN para verificación e identificación de locutor de modo supervisado. En ese caso, la trascripción está disponible y se podría compensar el ruido de un hipotético entorno hostil casi por completo, al menos a nivel de reconocimiento. Dado que la tarea que se está planteando es algo diferente, sería interesante estudiar cómo emplear la técnica KPD-MEMLIN para mantener las características específicas de cada locutor tras la adaptación. En ese sentido quizás sería interesante proyectar los vectores de características desde el espacio ruidoso correpondiente a uno limpio dependiente del locutor, dejando a un lado el genérico considerado hasta la fecha y que tan buenos resultados ha proporcionado. Si al final se utilizara dicho espacio de proyección posiblemente sería interesante aplicar técnicas de *clustering* de locutores.

La fusión de información puede proporcionar un nuevo e interesante punto de vista para obtener un más preciso modelo de probabilidad condicionada entre espacios de señal, término este que en todas las técnicas de adaptación de vectores de características propuestas posee una importancia capital. Así, se está estudiando la posibilidad de emplear la información visual de los labios para determinar la probabilidad condicionada entre Gaussianas. Dado que existe una clara correlación entre la posición de los labios y la generación de sonidos, no parece descabellado utilizar los vectores visuales para estimar, previa fase de entrenamiento, la probabilidad entre Gaussianas. De esta manera se podría independizar el cálculo de dicho término de cualquier tipo de señal con ruido acústico, lo que debería proporcionar mejores resultados de RAH.

No sólo el modelado de probabilidad condicionada entre espacios de señal ocupa una línea de trabajo futura, sino que también lo puede ser el modelo del espacio de señal. De este modo, y a pesar del buen resultado que para algunas de las técnicas

presentadas ha proporcionado el criterio MMSE a la hora de estimar el vector de transformación, se ha considerado oportuno explorar nuevas maneras de calcular dicho vector. Así, en el primer horizonte se ha planteado el estudio de técnicas de entrenamiento discriminativo para dar con nuevos y más robustos modelos del espacio de señal.

A partir de los excelentes resultados logrados con la técnica KPD-MEMLIN, se puede concluir que el término definido por la probabilidad a posteriori del fonema correspondiente dado el entorno y el vector de características ruidoso es fundamental. Teniendo en cuenta que, de cara a estimar dicha probabilidad, se antoja insuficiente utilizar únicamente las GMMs entrenadas para cada entorno básico, se ha considerado oportuno estudiar nuevos métodos a tal efecto. A partir de todo lo anterior se propone incluir más información, como el modelo de lenguaje, o vectores acústicos anteriores, que podrían incluirse, por ejemplo, en un algoritmo de Viterbi.

Durante todo el trabajo se han presentado técnicas para reducir el efecto del ruido. Sin embargo, éstas también se pueden emplear para otros fines, como por ejemplo para extensión de banda frecuencial. Es conocido que, en situaciones ordinarias, las señales obtenidas a una mayor frecuencia de muestreo proporcionan mejores resultados de RAH que aquellas que se graban a menor frecuencia. Sin embargo, en algunas ocasiones esta segunda situación es inevitable, por lo que se podría esperar que, si se aprendiera cómo proyectar los correspondientes vectores de características hacia el espacio construido por la señal obtenida a una mayor frecuencia de muestreo, se podría reducir el efecto del submuestreo mediante una adaptación. Para ello, y en la versión más sencilla, bastaría con modificar los espacios de señal, de modo que el origen se compondría por los vectores de características submuestreados, y el destino estaría formado por los vectores acústicos obtenidos a una alta frecuencia de muestreo.

Otra de las líneas futuras que se han planteado es el desarrollo de una fase de entrenamiento “ciega” para el algoritmo MEMLIN. Se podría suponer que esto es innecesario puesto que ya se posee una para el método PD-MEMLIN que, teóricamente, es aplicable para la técnica MEMLIN dado que ésta se puede ver como una simplificación del método PD-MEMLIN en el que se considera un único fonema. Sin embargo, existen razonables dudas para pensar que el buen funcionamiento de la fase de entrenamiento “ciega” presentada se deba al empleo del algoritmo KPD-MEMLIN que, en el caso que se pretende tratar, no sería posible aplicar. De todos modos, y como paso previo a cualquier estudio, sería interesante aplicar de modo directo la fase de entrenamiento “ciega” ya desarrollada al algoritmo MEMLIN.

11.4 Indicios de Calidad.

Durante el tiempo empleado en la realización de la presente tesis doctoral, varios han sido los méritos alcanzados. Algunos de ellos están directamente implicados en el trabajo presentado, mientras que otros se encuentran simplemente relacionados. En esta Sección se hace una breve reseña de todos ellos.

11.4.1 Publicaciones en Congresos Nacionales.

- E.Lleida, E.Masgrau, A.Ortega, A.Miguel, L.Buera "Reconocimiento automático del habla en vehículos, resultados con SpeechDat-Car", *II Jornadas sobre Tecnologías del Habla*. 2002.

- O. Saz, L. Buera, E. Lleida, A. Miguel, A. Ortega "Algoritmos de compensación de características cepstrales para reconocimiento automático del habla robusto", *III Jornadas sobre Tecnologías del Habla*. 2004.

- L. Buera, A. Miguel, E. Lleida, A. Ortega, O. Saz "Avances en la normalización cepstral con señal estéreo para el reconocimiento robusto de voz en el entorno del vehículo", *III Jornadas sobre Tecnologías del Habla*. 2004.

- A. Miguel, R. Rose, E. Lleida, L. Buera, A. Ortega, O. Saz. "Decodificador eficiente para normalización del tracto vocal en reconocimiento automático del habla en tiempo real", *III Jornadas sobre Tecnologías del Habla*. 2004.

- A. Ortega, F. Sukno, E. Lleida, A. Frangi, A. Miguel, L. Buera y E. Zacur. "Base de datos audiovisual y multicanal en castellano para reconocimiento automático del habla multimodal en el automóvil", *III Jornadas sobre Tecnologías del Habla*. 2004.

- L. Buera, E. Lleida, A. Miguel, A. Ortega, O. Saz "Time-dependent cross-probability model for feature vector normalization", *IV Jornadas sobre Tecnologías del Habla*. 2006.

- A. Uria, A. Ortega, M. I. Torres, A. Miguel, V. Guijarrubia, L. Buera, J. Garmendia, E. Lleida, O. Aizpuru, A. Varona, E. Alonso, O. Saz "A virtual butler controlled by speech", *IV Jornadas sobre Tecnologías del Habla*. 2006.

- L. Buera, E. Lleida, J. D. Rosas, J. Villalba, A. Miguel, A. Ortega, O. Saz "Verificación e identificación de locutor con normalización de vectores de características en entornos acústicos adversos", *III jornadas de reconocimiento biométrico de personas*. 2006.

- A. Miguel, F. Sukno, J. J. Gracia, T. Carmona, J. E. García, L. Buera, C. Orrite, A. Frangi, E. Lleida "Aportaciones de la lectura de labios a la seguridad de los sistemas biométricos", *III jornadas de reconocimiento biométrico de personas*. 2006.

11.4.2 Publicaciones en Congresos Internacionales.

- L. Buera, E. Lleida, A. Miguel, A. Ortega "Multi-environments model based linear normalization for speech recognition in car conditions", *ICASSP*. 2004.

- A. Ortega, F. Sukno, E. Lleida, A. Frangi, A. Miguel, L. Buera y E. Zacur "AV@CAR: a spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition", *LREC*. 2004.

- L. Buera, E. Lleida, A. Miguel, A. Ortega "Multi-environments model based linear normalization for robust speech recognition", *9-th International conference "speech and computer", SPECOM*. 2004.

- L. Buera, E. Lleida , J. D. Rosas, J. Villalba, A. Miguel, A. Ortega, O. Saz "Speaker verification and identification using phoneme dependent multi-environment models based linear normalization in adverse and dynamic acoustic environments", *Summer school for advanced studies on biometrics for secure authentication and system integration.* 2005.

- L. Buera, E. Lleida , A. Miguel, A. Ortega "Robust speech recognition in cars using phoneme dependent multi-environment linear normalization", *EUROSPEECH.* 2005.

- A. Ortega , E. Lleida , E. Masgrau, L. Buera, A. Miguel "Acoustic feedback cancellation in speech reinforcement system for vehicles", *EUROSPEECH.* 2005.

- A. Miguel, E. Lleida , R. Rose, L. Buera, A. Ortega "Augmented state space acoustic decoding for modelling local variability in speech", *EUROSPEECH.* 2005.

- L. Buera, E. Lleida , A. Miguel, A. Ortega "Multi-environment linear normalization for robust speech analysis in cars", *Biennial on DSP for in-vehicle and mobile systems.* 2005.

- A. Ortega, E. Lleida , E. Masgrau, L. Buera, A. Miguel "Acoustic echo reduction in a two-channel speech reinforcement system for vehicles", *Biennial on DSP for in-vehicle and mobile systems.* 2005.

- L.Buera, E. Lleida , A. Miguel, A.Ortega "Recent advances in pd-memlin for speech recognition in car conditions", *ASRU.* 2005.

- A. Ortega, E. Lleida, E. Masgrau, L. Buera, A. Miguel "Stability control in a two-channel speech reinforcement system for vehicles", *ICASSP.* 2006.

- L. Buera, E. Lleida, J. A. Nolazco, A. Miguel, A. Ortega "Time-dependent cross-probability model for multi-environment model based linear normalization", *ICSLP.* 2006.

- O. Saz, A. Miguel, E. Lleida, A. Ortega, L. Buera "Study of time and frequency variability in pathological speech and error reduction methods for automatic speech recognition", *ICSLP.* 2006.

- A. Miguel, E. Lleida, A. Juan, L. Buera, A. Ortega, O. Saz "Local transformation models for speech recognition", *ICSLP.* 2006.

- A. Miguel, L. Buera, E. Lleida, A. Ortega, O. Saz "On-Line Feature and Acoustic Model Space Compensation for Robust Speech Recognition in Car Environment", *IEEE Intelligent Vehicles Symposium.* 2007.

- I. Hernández, P. García, J. Nolazco, L. Buera, E. Lleida "Robust Automatic Speech Recognition Using PD-MEEMLIN", *3rd Iberian Conference on Pattern Recognition and Image Analysis.* 2007.

- L. Buera, A. Miguel, E. Lleida, A. Ortega, O. Saz "Cross-Probability Model base don GMM for Feature Vector Normalization in Car Environments", *Bienal on DSP for in-Vehicle and Mobile Systems.* 2007.

- L. Buera, A. Miguel, E. Lleida, A. Ortega, O. Saz "Cross-Probability Model base don GMM for Feature Vector Normalization in Car Environments", *Bienal on DSP for in-Vehicle and Mobile Systems*. 2007.
- L. Buera, A. Miguel, O. Saz, E. Lleida, A. Ortega "Evaluation of the Combined Use of MEMLIN and MLLR on the Non-native Adaptation Task of Hiwire Project Database", *INTERSPEECH*. 2007.
- L. Buera, A. Miguel, O. Saz, E. Lleida, A. Ortega "On the Jointly Unsupervised Feature Vector Normalization and Acoustic Model Compensation for Robust Speech Recognition", *INTERSPEECH*. 2007.
- L. Buera, A. Miguel, O. Saz, E. Lleida, A. Ortega "Robust Speech Recognition with on-line Unsupervised Acoustic Feature", *ASRU*. 2007.

11.4.3 Publicaciones en Revistas Internacionales.

- L. Buera, E. Lleida, A. Miguel, A. Ortega y O. Saz "Cepstral Vector Normalization Based on Stereo Data for Robust Speech Recognition", *IEEE Trans. On Audio Speech and Language Processing*. 2007.

11.4.4 Capítulos de Libro.

- A. Ortega, E. Lleida, E. Masgrau, L. Buera y A. Miguel "Acoustic echo reduction in a two-channel reinforcement system for vehicles", *DSP for in-Vehicle and Mobile Systems vol. II. Abut, Hansen, Takeda (Eds.) New York. Springer (ISBN-10: 0-387-33503-X)*. 2006.

11.4.5 Proyectos en los que se ha Participado.

- Sistema de diálogo para el acceso a la información mediante habla espontánea en diferentes entornos (CICYT TIC2002-04103-C03-01). 2003-2005.
- BISECURE, Biometrics for Secure Authentication (IST-2002-507634). 2004-2007.
- EDECÁN: Tecnologías de adaptación al contexto acústico en sistemas de diálogo multidominio (TIN2005-08660-C04-01). 2005-2008.
- Sistema integral de comunicaciones para vehículos (PROFIT CIT-370100-2005-4). 2005-2006.

11.4.6 Estancias en el Extranjero.

- Instituto Tecnológico y de Estudios Superiores de Monterrey, ITESM. Campus Monterrey. México. Septiembre-diciembre 2005.
- Microsoft Research. Redmond. USA. Junio-agosto 2006.
- Microsoft Research. Redmond. USA. Junio-agosto 2007.

11.4.7 Patentes.

- James G. Droppo, Alejandro Acero, Luis Buera "A Pitch Model for Noise Estimation". Microsoft Research. 2007.

11.4.8 Otros Méritos y Proyectos.

- Se participó en la evaluación de la base de datos *Hiwire*, presentándose los mejores resultados en la tarea *Non-native adaptation task*.
- Se está trabajando en la evaluación NIST 2008, en la que se espera aplicar las técnicas de robustez presentadas en esta tesis doctoral.
- Se está preparando un capítulo para el libro de futura publicación *"Processing of In-Vehicle Signals: Data Collection and Driver Bahavior"*.

Bibliografía

[Ace90] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition.* PhD thesis, ECE Department, Carnegie-Mellon University, Pittsburgh, USA, Sep 1990.

[ADKZ00] 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.

[AH95] A. Acero and X. Huang. Augmented cepstral normalization for robust speech recognition. In *Proc. IEEE Workshop on Automatic Speech Recognition*, Snowbird, UT, Dec 1995.

[AKC94] 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.

[AS90] A. Acero and R. M. Stern. Environmental robustness in automatic speech recognition. In *Proc. of ICASSP.*, pages 849–852, 1990.

[Ata83] B.-S. Atal. Efficient coding of lpc parameters by temporal decomposition. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 81–84, 1983.

[Bak75] J. K. Baker. *Stochastic Modelling for Automatic Speech Understanding*, pages 512–542. Academic Press, New York, NJ, USA, 1975.

[Bau72] 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.

[BCDM88] 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.

[BDHM72] H. Bourlard, S. Dupont, H. Hermansky, and N. Morgan. Towards sub-band-based speech recognition. In Oved Shisha, editor, *proc. of European Signal Processing Conference*, pages 1579–1582, Trieste, Italy, 1972. Academic Press.

[Bea92] V. L. Beattie. *Hidden Markov Model State-Based Noise Compensation.* PhD thesis, Churchill College, Cambridge University, Cambridge, UK, 1992.

[Bel57] R. E. Bellman. *Dynamic Programming*. Princeton University Press., Princeton, NJ, 1957.

[Bel97] J. R. Bellegarda. Statistical techniques for robust asr: Review and perspectives. In *in Proc. Eurospeech*, pages 33–36, 1997.

[BFS99] 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.

[BGH00] 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.

[BHM96] H. Bourlard, H. Hermansky, and N. Morgan. Towards increasing speech recognition error rates. *Speech Communication*, 18(3):205–231, 1996.

[BJM83] 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.

[BK65] R. Bellman and R. Kabala. Dynamic programming and modern control theory. *Academic Press Inc.*, 1965.

[BK03] 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.

[BLM+06] 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.

[BLM+07] 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 Language and Audio Processing*, 15:1098–1113, March 2007.

[BLMO04a] 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.

[BLMO04b] 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.

[BLMO05a] 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.

[BLMO05b] 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.

[BLMO05c] 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.

[BLN+06] 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.

[BLO+04] 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.

[BLR+05] 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.

[BLR+06] 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.

[BML+07a] L. Buera, A. Miguel, E. Lleida, A. Ortega, and O. Saz. Cross-probability model based on gmm for feature vector normalization in car environments. In *Biennial on DSP for in-Vehicle and Mobile Systems*, Istanbul, Spain, Jun. 2007.

[BML+07b] L. Buera, A. Miguel, E. Lleida, O. Saz, and A. Ortega. On the jointly unsupervised feature vector normalization and acoustic model compensation for robust speech recognition. In *Interspeech*, Antwerp, Belgium, Aug. 2007.

[BML+07c] L. Buera, A. Miguel, E. Lleida, O. Saz, and A. Ortega. Robust speech recognition with on-line unsupervised acoustic feature compensation. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, Kioto, Japan, Dec. 2007.

[BMS+07] L. Buera, A. Miguel, O. Saz, E. Lleida, and A. Ortega. Evaluation of the combined use of memlin and mllr on the non-native adaptation task of hiwire project database. In *Interspeech*, Antwerp, Belgium, Aug. 2007.

[Boc93] 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.

[Bol79] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on ASSP*, 27:113–120, April 1979.

[BPS+92] 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.

[CC91] 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.

[CDEB91] H. Cerf-Danon and M. El-Béze. Three different probabilistic language models: Comparison and combination. In *Proc. ICASSP*, pages 297–300, 1991.

[CGJ+01] 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.

[CHA+95] 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.

[CRBJ00] 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.

[CSL99] 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.

[DAPH00] 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.

[DDA01] J. Droppo, L. Deng, and A. Acero. Evaluation of the splice algorithm on the aurora2 database. In *in Proc. Eurospeech*, volume 1, Sept. 2001.

[DDA02] J. Droppo, L. Deng, and A. Acero. Uncertainty decoding with splice for noise robust speech recognition. In *Proc. ICASSP*, Florida, USA, May 2002.

[DDA03] 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.

[Del90] P. Deleglise. Décomposition temporelle : une technique cinématique de segmentation et de décodage acoustico-phonétique. In *évaluations. XVIIIème JEP*, pages 347–352, Montréal, Canada, 1990.

[DH73] R. Duda and P. Hart. *Pattern Classification and Scene Analysis.* J. Wiley and sons, 1973.

[DHS00] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* Wiley-Interscience, New York, 2000.

[Dig92] V. V. Digalakis. *Segment-based stochastic models of spectral dynamics for continuous speech recognition.* PhD thesis, Boston University, Boston, MA, USA, 1992.

[DLR77] 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.

[dlTFH01] 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.

[dlTPS+05] 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.

[DM80] 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.

[DMGA05] 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.

[ECY95] 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.

[EM85] 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.

[ETS00] 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.

[ETS02] ETSI. Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. Technical report, ETSI ES 202 050 version 1.1.1, Oct. 2002.

[FL94] A. L. N. Fred and J. M. N. Leitao. Improving sentence recognition in stochastic context-free grammars. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 9–12, 1994.

[Fri97] 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.

[Fuk90] K. Fukunaga. *Statistical pattern Recognition.* Academic Press, 1990.

[Fur86] 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.

[Gal95] M. J. F. Gales. *Model-Based Techniques for Noise Robust Speech Recognition.* PhD thesis, Cambridge University, Cambridge, UK, 1995.

[Gal97a] M. J. F. Gales. Transformation smoothing for speaker and environmental adaptation. In *Proc. of EUROSPEECH*, pages 2067–2070, 1997.

[Gal97b] M.J.F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. Cued/finfeng/tr291, Cambridge University, 1997.

[GB00] Z. Ghahramani and M.J. Beal. *Variational Inference for Bayesian Mixtures of Factor Analysers*, pages 449–455. MIT Press, 2000.

[GC89] 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.

[Gha02] Z. Ghahramani. *An introduction to hidden Markov models and Bayesian networks*, pages 9–42. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2002.

[Ghi92] O. Ghitza. *Auditory nerve representation as a basis for speech processing*, pages 453–486. Dekker, 1992.

[Ghi94] 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.

[GJ97] Z. Ghahramani and M. I. Jordan. Factorial hidden markov models. *Machine Learning*, 29(2-3):245–273, 1997.

[GJ99] P. Gelin and J-C. Junqua. Techniques for robust speech recognition in the car environment. In *Proc. of EUROSPEECH*, pages 2483–2486, 1999.

[GK00] 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.

[GL94] 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.

[Gla03] J. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17(2-3):137–152, 2003.

[GLF+93] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and Victor Zue. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, 1993.

[GN96] H. Gish and K. Ng. Parametric trajectory models for speech recognition. In *Proc. ICSLP*, pages 466–469, 1996.

[Gol94] W. D. Goldenthal. *Statistical Trajectory Models for Phonetic Recognition*. PhD thesis, Massachusetts Institute of Technology, MA, USA, 1994.

[Gon95] Y. Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 3(16):261–291, 1995.

[Gra00] 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.

[GRW97] A. L. Gorin, G. Riccardi, and J. H. Wright. How may i help you? *Speech Communication*, 23(1-2):113–127, 1997.

[GSC99] 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.

[GW87] R. C. González and P. Wintz. *Digital image processing*. Addison Wesley, 1987.

[GY92] M. J. F. Gales and S. J. Young. An improved approach to the hidden markov model decomposition of speech and noise. In *Proc. of ICASSP*, pages 233–236, 1992.

[GY93] M. J. F. Gales and S. J. Young. Hmm recognition in noise using parallel model combination. In *Proc. of EUROSPEECH*, pages 837–840, 1993.

[Hag00] A. Hagen. *Robust speech recognition based on multi-stream processing*. PhD thesis, Département d'informatique,EPFL, Lausanne, Switzerland, January 2000.

[HAH01] 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.

[HB00] A. Hagen and H. Bourlard. Using multiple time scales in the framework of multi-stream speech recognition. In *ICSLP*, 2000.

[Her90] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of Acoustic Society of America*, 87(4):1738–1792, 1990.

[Her98] 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.

[HGN+07] I. Hernández, P. García, J. Nolazco, 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.

[HH94] 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.

[HHG97] J. N. Holmes, W. J. Holmes, and P. N. Garner. Using formant frequencies in speech recognition. In *Proc. of EUROSPEECH*, pages 2083–2086, 1997.

[HL97] 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.

[HM94] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[HMBK91] 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.

[HMH93] 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.

[HN94] J. Herrando and C. Nadeu. 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.

[Hos01] H. Hoshino. Noise-robust speech recognition in a car environment based on the acoustic features of car interior noise. Technical report, 2001.

[HP00] 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.

[HS94] N. Hanai and R. M. Stern. Robust speech recognition in the automobile. In *in Proc. ICSLP*, pages 1339–1342, 1994.

[HTP96] 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.

[ITU96] ITU. Transmission performance characteristics of pulse code modulation channels. Technical report, Nov. 1996.

[JBM75] 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.

[Jel69] F. Jelinek. A fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development*, 13:675–685, Nov. 1969.

[Jel76] F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, 1976.

[Jel91] F. Jelinek. *Self-Organized Language Modelling for Speech Recognition*, chapter 6.1, pages 450–506. Morgan Kaufmann, San Mateo, CA, 1991.

[JH96] J.-C. Junqua and J. P. Haton. *Robustness in Automatic Speech Recognition.* Kluwer Academic Publishers, 1996.

[JHDL95] 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.

[JJ94] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Comput.*, 6(2):181–214, 1994.

[JJT02] 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.

[Jos02] L. Josifovski. *Robust Automatic Speech Recognition Missing and Unreliable Data.* PhD thesis, Department of Computer Science, University of Sheffield, UK, 2002.

[JRW87] 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.

[JW89] 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.

[Kai90] J.F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 381–384, Albuquerque, USA, 1990.

[Kat87] 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.

[KdM90] 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.

[KL51] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–87, 1951.

[Kle02] M. Kleinschmidt. *Robust Speech Recognition Based on Spectro-temporal Processing.* PhD thesis, University of Oldenburg, Germany, 2002.

[KMG98] B. E. D. Kingbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1-3):117–132, 1998.

[KNST94] T. Kuhn, H. Niermann, and E. G. Schukat-Talamazzini. Ergodic hidden markov models and polygrams for language modeling. In *Acoustics, Speech, and Signal Processing, ICASSP*, 1994.

[KPNN00] R. Kuhn, E. Perronnin, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, 2000.

[KSN00] S. Kanthak, K. Schütz, and H. Ney. 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.

[KSS00] 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.

[KU03] J. Kybic and M. Unser. Fast parametric elastic image registration. *IEEE Transactions on Image Processing*, 11(12):1427–1442, 2003.

[KUK98] 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.

[KWR97] C. Kamm, M. Walker, and L. Rabiner. The role of speech processing in human-computer intelligent communication. *Speech Communication*, 4(23):263–278, 1997.

[LB92] 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.

[LC98] J. S. Lin and R. Chen. Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.

[LD93] R. G. Leonard and G. Doddington. Tidigits speech corpus. Technical report, Texas Instruments, Inc,, 1993.

[Lea79] W. A. Lea. *Trends in Speech Recognition.* Ed. Lawrence Erlbaum, 1979.

[Leh75] E. L. Lehmann. *Nonparametrics.* Holden Day, 1975.

[Leo84] R. G. Leonard. A database for speaker independent digit recognition. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 9, pages 328–331, 1984.

[LHH+89] 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.

[Lle90] E. Lleida. *Compresión y Selección de Información en Reconocimiento Automático del Habla*. PhD thesis, Universidad Politécnica de Cataluña, UPC, 1990.

[LMO+02] 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.

[LMP01] 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.

[Lom11] E. Lombard. Le signe de l'elevation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37:101–119, 1911.

[LR98] L. Lee and R. Rose. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing*, 1(6):49–60, 1998.

[LSA94] F. H. Liu, R. M. Stern, and A. Acero. Environment normalization for robust speech recognition using direct cepstral comparison. In *Proc. ICASSP*, 1994.

[LW95] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous-density hidden markov models. *Computer Speech and Language*, 9:171–185, 1995.

[MBB01] 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.

[MBL+07] A. Miguel, L. Buera, E. Lleida, A. Ortega, and O. Saz. On-line feature and acoustic model space compensation for robust speech recognition in car environment. In *IEEE Intelligent Vehicles Symposium*, Istanbul, Turkey, Jun. 2007.

[MC95] 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.

[McN47] I. McNemar. Note on the sampling error of the diference between correlated proportions or percentages. In *Psychometrika*, volume 12, pages 153–157, 1947.

[MFM97] N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proc. of EUROSPEECH*, pages 2079–2082, 1997.

[MH82] M. Morgenthaler and C. Hansen. Use of attributed grammars in speech signal processing. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 1311–1313, 1982.

[MHB99] 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.

[MHJ+99] S. Martin, C. Hamacher, Liermann J, F. Wesssel, and H. Ney. 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.

[MLD+00] 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.

[MLJ+06] A. Miguel, E. Lleida, A. Juan, L. Buera, A. Ortega, and O. Saz. Local transformation models for speech recognition. In *ICSLP*, Pittsburgh, USA, 2006.

[MLR+05] 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.

[MMJ93] 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.

[Mol03] S. Molau. *Normalization in the Acoustic Feature Space for Improved Speech Recognition.* PhD thesis, University of Aachen, Germany, Feb 2003.

[Moo90] R. Moore. *Speech Processing.* McGraw Hill, 1990.

[Mor96] P. Moreno. *Speech recognition in noisy environments.* PhD thesis, ECE Department, Carnegie-Mellon University, Apr. 1996.

[Mor97] R. De Mori. Recent advances in feature extraction and acoustic modeling for automatic speech recognition. Technical report, Laboratoire Informatique d'Avignon, 1997.

[MOS01] 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.

[MOSS02] 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.

[MS97] 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.

[NE91] H. Ney 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.

[NEK94] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependencies in language modelling. *Computer Speech and Language*, 2(8):1–38, 1994.

[Ney90] H. Ney. Stochastic grammars and pattern recognition. In *Proc. of NATO ASI*, pages 319–344, 1990.

[Ney93] H. Ney. Architecture and search strategies for large-vocabulary continuous-speech recognition. In *Proc. of NATO ASI*, pages 59–84, 1993.

[NHUTO92] H. Ney, 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.

[NMNP87] H. Ney, D. Mergel, A. Noll, 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.

[NN96] 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.

[NW94] L. Neumeyer and M. Weintraub. Probabilistic optimal filtering for robust speech recognition. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 417–420, 1994.

[NW95] L. Neumeyer 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.

[NY94] 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.

[ODK96] M. Ostendorf, V. Digilakis, and O. Kimball. From hmms to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):360–378, 1996.

[ON95] S. Ortmanns and H. Ney. 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.

[OS75] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing.* Prentice-Hall, Inc., 1975.

[PB92] D. Paul and J. Baker. The design for the wall street journal-based csr corpus. In *DARPA Speech and natural language Workshop*, pages 357–362, Feb. 1992.

[PFF90] 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.

[Pit05] M. Pitz. *Investigations on Linear Transformations for Speaker Adaptation and Normalization.* PhD thesis, University of Aachen, 2005.

[PN05] M. Pitz and H. Ney. Vocal tract normalization equals linear transformation in cepstral space. *IEEE Transactions on Speech and Audio Processing*, 13(5):930–944, 2005.

[PNG+03] G. Potamianos, C. Net, 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.

[PPN+03] P. Pujol, S. Pol, C. Nadeu, 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.

[PT00] 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.

[PTG+92] 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.

[PW80] 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.

[Rab88] L. R. Rabiner. *A Tutorial on HMM and selected Applications in Speech Recognition*, chapter 6.1, pages 267–295. Morgan Kaufmann, 1988.

[RFS01] 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.

[RGMS96] 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.

[RJ93] 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.

[RV91] O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, 8:11–38, 1991.

[SA91] 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.

[SBdlTR01] 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.

[SC90] 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.

[SEP+07] J. C. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P-A. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos. The hiwire database, a noisy and non-native english speech corpus for cockpit communications. In *online at hppt://www.hiwire.org/*, April 2007.

[SG91] 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.

[SH00] 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.

[Shi85] S. M. Shieber. An introduction to unification-based approaches to grammar. *CSLI Lecture Notes: Center for the Study of Language and Information*, 1985.

[SKA+00] H. Shimodaira, Y. Kato, T. Akae, M. Nakai, 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.

[SL87] 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.

[SL96] 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.

[SP54] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *J. Acoustical Society America*, 26:212–215, 1954.

[SRM97] 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.

[SSNS02] H. Shimodaira, N. Sakai, M. Nakai, 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.

[STN94] V. Steinbiss, B. H. Tran, and H. Ney. Improvements in beam search. In *International Conference on Spoken Language Processing, ICSLP*, volume IV, pages 2143–2146, Yokohama, Japan, Sept. 1994.

[Sun95] 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.

[UIE94] 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.

[US98] 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.

[vdHBC+99] 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.

[VGCJ99] 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.

[Vin71] T. K. Vintsyuk. Elementwise recognition of continuous speech composed of words from a specified dictionary. *Cybernetics*, 7:133–143, March 1971.

[Vit67] 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.

[vK92] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier Science Publishers, 1992.

[VL98] A. Viikki and K. Laurila. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25, 1998.

[VM90] A. P. Varga and R. K. Moore. Hidden markov model decomposition of speech and noise. In *Acoustics, Speech, and Signal Processing, ICASSP*, pages 845–848, 1990.

[Wak77] 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.

[WBB00] 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.

[Web03] K. Weber. *HMM Mixtures (HMM2) for Robust Speech Recognition.* PhD thesis, Swiss Federal Institute of Technology Lausanne (EPFL), 2003.

[Wel67] 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.

[Wit01] T. Wittkop. *Two-channel noise reduction algorithms motivated by models of binaural interaction.* PhD thesis, University of Oldenburg, Germany, 2001.

[WW91] E. N. 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.

[WY93] 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.

[YEG+05] 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.

[YSvVH00] 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.

[YZH02] 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.

[ZG02] 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.

[ZSM95] 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.

[Zue97] V. Zue. Conversational interfaces: Advances and challenges. In *Proc. of EUROSPEECH*, pages 9–18, 1997.