Entrenamiento Discriminativo de Modelos Ocultos de Markov de Unidad Subléxica para su Aplicación a Sistemas de Reconocimiento Automático del Habla Continua

Autor: Albino Nogueiras Rodríguez
Director: Dr. José B. Marino Acebal

Departament de Teoria del Senyal i Comunicacions
Universitat Politècnica de Catalunya
Otoño de 1999

Prólogo

El reconocimiento automático del habla constituye uno de los mayores retos tecnológicos de la actualidad. De hecho, el afán por utilizar el habla como medio de comunicación con todo tipo de aparatos, cosas y animales ha acompañado al hombre desde tiempo inmemorial. A pesar de ello, no ha sido hasta hace relativamente poco, gracias al mejor conocimiento de los mecanismos de producción y comprensión del habla, el desarrollo de las técnicas de procesado de señal y, sobre todo, el advenimiento de ordenadores electrónicos de prestaciones cada vez mayores, que se ha materializado la posibilidad de realizar este tipo de comunicación.

Dos aspectos son fundamentales en los sistemas actuales de reconocimiento del habla: la caracterización de los sonidos a ser reconocidos, o modelado acústico; y la articulación de estos sonidos en los significados correspondientes, modelado del lenguaje. Ambos aspectos están intimamente relacionados, aunque suelen tratarse de manera independiente. Así, sólo el incremento de la precisión en el modelado acústico ha permitido la utilización de modelos del lenguaje cada vez más complicados, posibilitando el reconocimiento de tareas, a su vez, cada vez más complejas.

Puede decirse que los primeros sistemas de reconocimiento del habla, en los que sólo se reconocían palabras aisladas, únicamente implicaban la fase de modelado acústico, careciendo de un modelado del lenguaje propiamente dicho. A partir de ese momento, y en buena medida gracias a los progresos experimentados por el modelado acústico, el modelado del lenguaje ha ido ganando importancia, hasta constituir hoy en día uno de los aspectos más tratados en la literatura especializada [64, 72, 77, 101, 109, 10, 46]. En la actualidad, los sistemas más ambiciosos de reconocimiento del habla se basan en la utilización de modelos de unidad subléxica [39, 59, 54]. Las unidades subléxicas son segmentos acústicos sin significado propio pero tales que, por concatenación de las unidades correspondientes, permiten la construcción del modelo acústico de cualquier palabra o frase. Aunque también se utilizan unidades subléxicas cuya propia definición depende de la tarea a reconocer —por ejemplo, los semidígitos— es habitual utilizar como unidad subléxica alguna unidad fonética definible en el idioma a reconocer —fonemas, semisílabas, etc.—.

Enlazados mediante modelos del lenguaje apropiados, los modelos de unidad subléxica permiten el reconocimiento de cualquier tarea de palabras aisladas o conectadas, así como otras alternativas más ambiciosas como son los sistemas de reconocimiento del habla espontánea, de dictado automático o de diálogo. Además, los sistemas basados en la utilización de unidades subléxicas permiten que la construcción de sistemas de reconocimiento se realice a partir de material acústico de entrenamiento extraido de una base de datos de propósito general, independiente del locutor y de la tarea. La característica de que un mismo conjunto de modelos acústicos, entrenados a partir de una base de datos

independiente tanto del locutor como de la tarea a reconocer, permita el reconocimiento de cualquier tarea, constituye el rasgo más significativo del tipo de sistema que será abordado en esta tesis: los sistemas de reconocimiento de grandes vocabularios en habla continua [92].

A pesar del creciente interés del modelado del lenguaje, el acústico todavía mantiene su interés, dado que el aumento en la precisión de la caracterización acústica es uno de los mecanismos más efectivos en la mejora del reconocimiento. Por otro lado, las técnicas actuales de modelado acústico están aún muy lejos de permitir el reconocimiento de cualquier tarea con una tasa de error suficientemente baja. Si se considera como objetivo último del modelado acústico la correcta decodificación en sus sonidos constitutivos de cualquier frase, con independencia de su significado, el camino a recorrer es aún muy largo: la tasa de fonemas correctamente reconocidos en frases castellanas se sitúa en torno al 70%, el 60% en el caso del inglés.

Entre las distintas propuestas realizadas para el modelado acústico destaca una que ha recibido especial atención en los últimos años: el entrenamiento discriminativo de sistemas de reconocimiento basados en modelos ocultos de Markov [3, 18, 41, 84, 99]. Frente a los planteamientos clásicos de estimación o entrenamiento de los modelos acústicos —que centran su esfuerzo en el correcto modelado de las señales a reconocer—, las técnicas de entrenamiento discriminativo pretenden minimizar directamente la tasa de error del sistema de reconocimiento. Una de las claves del éxito de los métodos de entrenamiento discriminativo radica en su capacidad de tratar —siempre con bastante éxito— las distintas fases del reconocimiento: extracción de características [16, 95], cuantificación vectorial [45, 30, 34], y entrenamiento de los modelos de Markov [3, 2, 18]. No obstante, los sistemas de entrenamiento discriminativo, y especialmente los que abordan el entrenamiento de los modelos de Markov, presentan la limitación de estar, inicialmente, planteados para su aplicación utilizando bases de datos dependientes de la tarea a reconocer. Así, aunque existen distintas propuestas de entrenamiento de los modelos de unidad subléxica utilizando bases de datos de propósito general, algunas de estas propuestas no tratan realmente del reconocimiento de tareas *auténticas* de reconocimiento del habla continua, sino sólo de tareas *sintéticas*, como la clasificación de fonemas [79, 98], o la decodificación acústico fonética [53, 40]. Por otro lado, en otros casos, la tarea de reconocimiento abordada sí es una autentica tarea de reconocimiento del habla continua, pero la base de datos empleada en el entrenamiento está formada por frases provinientes de la misma tarea a reconocer [59, 100, 102], con lo que la independencia de la tarea de los modelos acústicos obtenidos queda en entredicho. Sólo en unos pocos trabajos [53, 80] —y, en cierto sentido, debido a la extensión de la tarea considerada, [113]—, el objetivo del entrenamiento es la obtención de modelos acústicos de unidad subléxica independientes de la tarea de habla continua a reconocer.

La dificultad existente en la aplicación de entrenamiento discriminativo a los sistemas de reconocimiento basados en unidades subléxicas independientes de la tarea radica en la propia naturaleza de las dos metodologías que se pretende unificar: mientras el entrenamiento de unidades subléxicas para el reconocimiento del habla continua es inherentemente independiente de cualquier tarea de reconocimiento; el entrenamiento discriminativo está siempre orientado a la minimización del número de errores cometido en el reconocimiento de una tarea concreta. La solución inmediata a esta incompatibilidad consiste en el empleo de una tarea *sintética* definible a partir de material acústico independiente de la tarea, y que represente de algún modo la capacidad del sistema de

reconocimiento de diferenciar las distintas unidades subléxicas. Por ejemplo, se puede considerar la minimización del número de errores cometido en clasificación de fonemas o decodificación acústico fonética [79, 98, 53, 40]. No obstante, y aunque las prestaciones del sistema en el reconocimiento de la tarea sintética aumentan considerablemente, esta mejoría no siempre se traslada al reconocimiento de tareas de habla continua reales [81]. En este contexto, resulta significativo que Chou, Juang y Lee mencionen en 1992, al final de las conclusiones de [18]:

> "...We demonstrated de effectiveness of the proposed algorithm (*segmental GPD training*) in isolated word and connected digit recognition applications. Further research and experimentats on sub-word based systems are in progress."

Sin embargo, y aunque el resultado de estos progresos tarda cuatro años en aparecer [53], la solución adoptada es tan sencilla como minimizar la tasa de error de frase en decodificación acústico fonética, utilizando exactamente el mismo sistema de entrenamiento discriminativo que el empleado en el caso de los modelos acústicos dependientes de la tarea[1].

En esta tesis se propone una metodología, el entrenamiento de mínima confusibilidad sobre segmentos acústicos de longitud limitada que, partiendo de las líneas maestras de los trabajos de Chou y Lee, permite aplicar entrenamiento discriminativo —utilizando bases de datos independientes de la tarea a reconocer— a modelos de Markov de unidad subléxica para su utilización en el reconocimiento del habla continua. Aunque el objetivo último es la consecución de modelos independientes de la tarea, el desarrollo teórico que conduce a esta propuesta se basa en la adaptación a tareas concretas a partir del conocimiento del lenguaje de la tarea a reconocer y el material de entrenamiento disponible en una base de datos de propósito general. A partir del esquema de adaptación a la tarea, y particularizando para un lenguaje general, se obtiene un esquema de entrenamiento discriminativo de modelos de unidad subléxica independientes de la tarea que ha permitido mejorar sensiblemente las tasas de reconocimiento en distintas tareas de habla continua tanto en inglés como en castellano [80], y utilizando como unidad subléxica tanto fonemas independientes del contexto como semifonemas dependientes del mismo [82].

Objetivos y Restricciones de la Tesis

El propósito fundamental de esta tesis es el entrenamiento discriminativo de unidades subléxicas para su aplicación a tareas de reconocimiento del habla continua. Se pretende obtener las mismas mejoras conseguidas por el entrenamiento discriminativo utilizando bases de datos dependientes de la tarea, pero con bases de datos de propósito general.

Se ha considerado que, para que las propuestas sean de verdadero interés práctico, los sistemas desarrollados deben cumplir las restricciones siguientes:

1. El sistema de referencia debe ser de altas prestaciones.

2. La evaluación de la experimentación debe referirse a una tarea real de reconocimiento del habla continua.

[1]Cabe señalar, respecto a este trabajo, que las frases de entrenamiento son muy cortas, entre dos y cuatro palabras en inglés. Este detalle de la experimentación resulta en una alta coincidencia con una de las propuestas originales de esta tesis: la utilización de segmentos acústicos de longitud limitada.

3. Solo se pueden utilizar bases de datos de entrenamiento independientes de la tarea.

La primera de las restricciones se ha impuesto para garantizar que la mejoría obtenida gracias al entrenamiento discriminativo no puede ser alcanzada con un simple cambio de los parámetros del sistema. Por ejemplo, es bien sabido que los modelos acústicos que combinan múltiples informaciones —espectro más sus primera y segunda derivadas, etc.— funcionan sensiblemente mejor que los que sólo utilizan el espectro. Un procedimiento que permita mejorar las prestaciones de este último tipo de sistema sólo es realmente útil si también es capaz de mejorar las del primero. Esta restricción ha llevado a considerar, como objetivo último de la tesis, la optimización de los sistemas que mejores prestaciones aportan en modelado acústico para el reconocimiento del habla continua: las unidades subléxicas dependientes del contexto, que son tratadas en el apartado 2.4.

También se ha descartado la utilización de marcos experimentales *sintéticos* que sólo pueden, a lo sumo, reflejar indirectamente las prestaciones del sistema en el reconocimiento de tareas reales de habla continua —como los mencionados reconocimiento de fonemas aislados y decodificación acústico fonética—. Por el contrario, se ha optado por ilustrar los métodos propuestos con los resultados obtenidos en una tarea concreta: el reconocimiento de las cadenas de dígitos en inglés de Texas Instruments [60], en adelante TIDIGITS. Puede objetarse que se trata de una tarea muy poco representativa de la problemática de los sistemas de reconocimiento de grandes vocabularios en habla continua, ya que el vocabulario a reconocer es de muy reducido tamaño y el modelo del lenguaje extremadamente sencillo. No obstante, presenta la ventaja de tratarse de una tarea muy extendida, lo cual facilita la reproducibilidad de los experimentos y la comparación de los resultados con otros trabajos. Por otro lado, en paralelo con esta línea experimental, y de manera independiente, se han realizado varias series de experimentos utilizando la base de datos SpeechDat en castellano [75] diversas tareas de reconocimiento del habla continua —fechas, horas, palabras ricas fonéticamente, etc.—. Los resultados más significativos de estos experimentos se detallan en el apéndice B y confirman, a grandes rasgos, los obtenidos en TIDIGITS.

Finalmente, y aunque la tarea del reconocimiento de cadenas de dígitos es acometida usualmente utilizando modelos acústicos entrenados a partir de bases de datos específicas, sólo se ha considerado la posibilidad de efectuar el entrenamiento con material proviniente de bases de datos de propósito general. En concreto se ha optado por el uso de otra base de datos igualmente habitual: TIMIT, formada por frases balanceadas fonéticamente [51].

Estructura de la Tesis

En el capítulo de introducción son presentados los detalles generales de la experimentación utilizada a lo largo de la tesis. A continuación, se hace un repaso de los fundamentos del problema abordado, esto es: tanto el reconocimiento del habla y el habla continua, como el entrenamiento discriminativo. Este repaso no pretende ser exhaustivo. Por el contrario, se ha intentado tratar en mayor profuncidad aquellos aspectos que mayor repercusión tienen en el resto de la tesis, pasando sólo superficialmente por otros aspectos que ya han sido ampliamente tratados con anterioridad —los modelos de Markov o el entrenamiento de máxima verosimilitud—.

El capítulo 2 es el verdadero núcleo de esta tesis. En él se presenta la solución propuesta para la aplicación de entrenamiento discriminativo al entrenamiento de modelos

de unidad subléxica para el reconocimiento del habla continua. Se parte de las propuestas precedentes de entrenamiento discriminativo de modelos de unidad subléxica. A partir de las limitaciones encontradas en ellos, se propone una metodología alternativa: el entrenamiento de mínima confusibilidad aplicado a segmentos acústicos de longitud limitada. Con esta metodología se puede acometer tanto la adaptación a tareas concretas de reconocimiento del habla continua como el entrenamiento discriminativo independiente de la tarea, utilizando en ambos casos bases de datos de propósito general.

En el capítulo 3 se explica el algoritmo de optimización empleado en la minimización de la función de coste: el algoritmo de búsqueda adaptativa de gradiente. Buena parte del éxito de la experimentación presentada en esta tesis se basa tanto en la independencia frente a la elección del paso de aprendizaje del algoritmo, como a su capacidad de realizar de manera automática el escalado de las variables optimizadas, acelerando así la convergencia del proceso y las prestaciones del sistema reestimado.

Finalmente, se presentan las conclusiones de esta tesis y los apéndices. En el primer apéndice se hace referencia a algunos aspectos de la realización práctica de la experimentación. En concreto, a algunos ajustes realizados en el algoritmo de reestimación, y a las diferencias entre los distintos marcos experimentales que ilustran la tesis. En el segundo se presentan los resultados experimentales obtenidos utilizando las mismas metodologías aquí presentadas, pero llevadas a cabo con material de la base de datos SpeechDat en castellano. Estos experimentos han sido realizados por personas distintas al autor de esta tesis, aunque siempre con su colaboración y asesoramiento. Finalmente, se reproducen las fórmulas más importantes empleadas en la realización de los experimentos presentados en esta tesis.

Índice General

Capítulo 1

Introducción

Esta introducción cumple dos funciones distintas: en primer lugar se presentan los detalles del marco experimental empleado en la tesis; a continuación se repasan los fundamentos de la problemática asociada al entrenamiento discriminativo de unidades subléxicas para su aplicación al reconocimiento del habla continua. En concreto, se repasan los fundamentos del reconocimiento del habla utilizando modelos ocultos de Markov, el reconocimiento del habla continua con modelos de unidad subléxica, y el entrenamiento discriminativo de los modelos acústicos.

1.1 Marco Experimental

En esta tesis se muestran, principalmente, resultados de reconocimiento en dos tareas distintas de reconocimiento en inglés americano: el de las cadenas de dígitos y la decodificación acústico fonética (**DAF**[1]). El reconocimiento de las cadenas de dígitos es una tarea que puede ser acometida utilizando tanto sistemas dependientes de la tarea, como independientes de la misma. Por tanto, permite realizar una comparación *a tres bandas* con el resultado del objetivo de esta tesis, el entrenamiento discriminativo de modelos de unidad subléxica para tareas de reconocimiento del habla continua. Por otro lado, la decodificación acústico fonética —esto es, el reconocimiento de fonemas sin restricciones léxicas— proporciona una medida de la calidad alcanzada en el modelado acústico de las unidades subléxicas, pudiéndose considerar que el objetivo último de éste es la correcta decodificación del contenido fonético de cualquier elocución a reconocer, con independencia de su significado.

1.1.1 Tarea a reconocer y bases de datos empleadas

1.1.1.1 Reconocimiento de TIDIGITS usando modelos de Markov entrenados con TIMIT

El reconocimiento de las cadenas de dígitos se refiere siempre a las 4.312 cadenas de entre uno y siete dígitos que componen la parte masculina del corpus de reconocimiento (**test**) de la base de datos TIDIGITS [60]. Las cadenas son pronunciadas por 56 locutores extraidos de 21 zonas geográficas distintas de los EE.UU. Las señales están muestreadas a 20KHz y cuantificadas a 16 bits. Se distinguen once dígitos —zero, one, two, three, four, five, six,

[1]En la página 117 hay un glosario de todas las abreviaturas utilizadas a lo largo de esta tesis.

seven, eight, nine y mboxoh—, así como una unidad de silencio y otra para la oclusión glotal que, en ocasiones, precede a las palabras comenzadas por vocal. Ni el silencio ni la oclusión glotal son tenidos en cuenta a la hora de evaluar los resultados del reconocimiento. Es decir, no se consideran como error las inserciones o borrados de estas dos unidades, así como tampoco las confusiones entre ellas. Tampoco se considera como acierto su correcto reconocimiento.

El reconocimiento de las cadenas de dígitos se realiza utilizando modelos entrenados con la parte masculina de TIMIT [51]. TIMIT es una base de datos formada por 6300 frases, pronunciadas por 630 locutores y seleccionadas de manera que cada unidad subléxica aparezca en el máximo número posible de contextos. Las frases están muestreadas a 16KHz y cuantificadas con 16 bits. Se ha optado por realizar únicamente experimentos con locutores masculinos debido al escasez de los femeninos en TIMIT. No obstante, en todo caso, se trata de experimentos de reconocimiento independientes del locutor. En el caso de utilizarse modelos entrenados con TIMIT, sólo la parte de entrenamiento de esta base es empleada en todos los casos excepto en los experimentos con unidades subléxicas dependientes del contexto, en los cuales fue necesario incorporar también el material de reconocimiento para evitar sobreadaptación de los modelos acústicos. En un par de ejemplos de esta introducción, tablas 1.4 y 1.5, se han utilizado también modelos entrenados con la parte masculina del corpus de entrenamiento de TIDIGITS, **train**.

1.1.1.2 Decodificación acústico fonética de TIMIT

Para los experimentos de decodificación acústico fonética, la base de entrenamiento utilizada se limita a la parte masculina del corpus **train** de TIMIT (2608 frases), y se reconoce la parte masculina del corpus **test** (896 frases). En ambos casos se han omitido las frases **sa1** y **sa2** ya que son comunes a todos los locutores. Por tanto, se trata de un experimento dependiente del sexo, aunque independiente del texto y locutor. El apartado siguiente detalla la transcripción fonética utilizada. No se consideran como errores de decodificación las inserciones u omisiones de la oclusión glotal /q/. Tampoco lo son las confusiones entre los siguientes pares de unidades: /ax/ con /ah/, /ix/ con /ih/, /el/ con /l/, /en/ con /n/ y /dh/ con /d/.

1.1.1.3 Presentación de los resultados del reconocimiento

La tabla 1.1 muestra los resultados obtenidos con modelos de fonema entrenados según el criterio de máxima verosimilitud, tanto en DAF como en el reconocimiento de las cadenas de dígitos de TIDIGITS. El significado de las distintas columnas es el siguiente (esta información también aparece en el glosario la página 117):

Borr Tanto por ciento de unidades acústicas omitidas, o borradas, en la cadena reconocida.

Clas Tanto por ciento de unidades acústicas reconocidas correctamente conociendo con anterioridad los límites temporales de cada una —clasificación de unidades subléxicas, o reconocimiento de dígitos aislados—.

Corr Tanto por ciento de frases reconocidas correctamente.

Error Suma de las tasas de sustitución (Sust), inserción (Inse) y borrado (Borr).

Acierto Tanto por ciento de unidades acústicas correctamente reconocidas.

Tarea	Error	Sust	Inse	Borr	Acierto	Corr	Clas
DAF	39,9	21,2	12,3	6,4	72,4	0,0	61,1
Dígitos	2,10	1,02	0,55	0,53	98,44	94,1	99,7

Tabla 1.1: *Resultados en DAF y el reconocimiento de TIDIGITS empleando modelos de fonema entrenados con TIMIT según el criterio de máxima verosimilitud.*

Inse Tanto por ciento de unidades acústicas insertadas en la cadena reconocida.

A la vista de los resultados de la tabla 1.1 son remarcables los siguientes aspectos:

1. Ninguna de las frases de TIMIT es decodificada correctamente por entero (Corr=0).
2. Casi todos los dígitos son reconocidos correctamente cuando se conocen sus limites temporales (Clas=99,7).
3. El número de fonemas reconocidos correctamente en DAF (Acierto=72,4) supera al obtenido en clasificación de fonemas (Clas=61,1). Existen varias posibles explicaciones a este comportamiento aparentemente contradictorio:
   - En DAF, la utilización de una gramática estocástica permite reducir considerablemente el número de errores. En clasificación de fonemas aislados esta tipo de gramática es de imposible utilización.
   - Los límites temporales de las unidades, los proporcionados en la propia TIMIT, difieren de manera importante de los que se obtendrían utilizando los modelos de Markov en su determinación —y que son, de alguna manera, los empleados en DAF—.
   - Muchas realizaciones de fonema son de longitud muy corta (inferior a tres tramas). En tanto que su reconocimiento en DAF es posible —ya que siempre se puede *robar* tramas a las unidades circundantes—, en clasificación de fonemas no alcanzan el número mínimo de tramas para que el modelo correspondiente sea capaz de generarlas, y el reconocimiento correcto resulta imposible.

1.1.2 Transcripción fonética adoptada

El conjunto de fonemas utilizado en la transcripción en unidades subléxicas, tanto de la base de datos de propósito general TIMIT, como de la de cadenas de dígitos TIDIGITS, es prácticamente idéntico al ARPABET [110]. La única diferencia consiste en la asimilación en una única unidad de algunos pares de fonemas tales que, siendo muy parecidos entre sí, uno de ellos aparece muy poco en TIMIT —/em/ con /m/, /axr/ con /ax/ y /nx/ con /ng/—. Este conjunto de fonemas difiere del proporcionado por TIMIT en el tratamiento dado a las consonantes oclusivas sordas —/p/, /t/ y /k/—. En TIMIT, éstas son representadas mediante dos unidades distintas: una encargada de modelar la oclusión inicial, seguida de otra que modela el sonido impulsivo final. Este modelado presenta el problema de que cualquiera de las dos porciones que forman el fonema puede estar ausente. Así, el fonema /p/ se transcribe generalmente como /pcl/ /p/, pero también puede aparecer como /pcl/ o /p/ sólo. El principal problema de esta multiplicidad en las transcripciones es la necesidad de tenerla

Dígito	Transcripción
zero	z ih r ow
one	w ah n
two	t uw
three	th r iy
four	f ow r
five	f ay v
six	s ih k s
seven	s eh v ax n
eight	ey t
nine	n ay n
oh	ow

Tabla 1.2: *Transcripción en signos del ARPABET de los dígitos en inglés, usada en el reconocimiento de TIDIGITS mediante modelos acústicos de unidad subléxica.*

en cuenta no sólo en el entrenamiento, sino también en el reconocimiento, incrementando notoriamente la complejidad de éste ultimo. Por otro lado, las oclusiones iniciales son esencialmente idénticas a silencios de corta duración, en principio indistinguibles entre sí y con el propio silencio entre palabras. Es, incluso, habitual agrupar todas las oclusiones en una única unidad y, además, no considerar como error las confusiones entre esta unidad y el silencio [57, 54]. No obstante, esta asimilación de unidades distintas provoca una pérdida importante de información, ya que lleva, si la unidad impulsiva está ausente, a no distinguir entre las oclusivas y el silencio. Por estos dos motivos, el incremento de complejidad del sistema de reconocimiento y la pérdida de información fonética— se ha decidido adaptar la transcripción proporcionada en TIMIT de manera que el tratamiento de las oclusivas sordas sea el habitual del ARPABET. Una consecuencia de esta decisión es que la tasa de error de fonemas estimada en decodificación acústico fonética está sobrevalorada frente al caso en que cualquier oclusión puede ser confundida con cualquier otra. Este hecho no tiene mayor repercusión que la de impedir establecer comparaciones directas con otros sistemas cuyos resultados han sido publicados con anterioridad [57, 44, 24, 40].

Utilizando este conjunto de unidades, la transcripción de los dígitos en inglés utilizada para su reconocimiento es la indicada en la tabla 1.2. Se han tenido en cuenta dos situaciones distintas en la cual es habitual la fusión en un único sonido del final de un dígito con el inicial del siguiente. Son los casos en los que six es seguido por un dígito iniciado por /s/ —six o seven—, permitiéndose la omisión de la /s/ final de six; y los casos en que nine es precedido por un dígito acabado en /n/ —one, seven y nine—, permitiéndose la omisión de la /n/ inicial de nine.

1.1.3 Sistema de entrenamiento y reconocimiento

El sistema de entrenamiento y reconocimiento empleado es el propio del Grupo de Procesado de Señal de la UPC, RAMSES [70, 8, 11]. De hecho, una parte importante de los módulos de que se compone RAMSES ha sido desarrollada o actualizada para llevar a cabo la confección de esta tesis. RAMSES se basa en el uso de modelos semicontinuos de Markov [42]. Este tipo de modelado no es tan popular como los modelos continuos de

Markov [92], estando muy generalizada la opinión de que sus resultados son peores que los proporcionados por estos últimos. No obstante, no abundan las comparaciones directas entre ambos tipos de modelado y, en los casos en que sí se han realizado, el resultado de la comparación no siempre ha sido favorable a los modelos continuos [36, 23, 112].

Aunque no se disponen de comparaciones directas entre RAMSES y otros sistemas de reconocimiento, lo habitual del reconocimiento de TIDIGITS permite realizar comparaciones con otros sistemas cuyos resultados han sido publicados con anterioridad. Por ejemplo, en [18], el sistema utilizado consiste de modelos continuos de Markov de diez estados para cada dígito —no se distingue entre hombre y mujer—. La parametrización de la señal es prácticamente idéntica a la usada en RAMSES, y cada estado de los modelos está caracterizado con 64 distribuciones gaussianas. La tasa de cadenas erróneas utilizando este sistema de reconocimiento es del 1,3%, mencionándose en el artículo que este resultado representa 'top performance' en esta tarea. Utilizando RAMSES con modelos de dígito, pero configuraciones mucho más sencillas —cuantificación vectorial a 256 símbolos, lo cual implica unas mil gaussianas en total, frente a las casi diez mil de [18]—, el resultado alcanzado es de un 1,6% de cadenas erróneas [1]. Así mismo, y utilizando modelos semicontinuos semejantes a los utilizados en RAMSES, en [84] se reporta una tasa de error de sólo el 1,0%. En este caso, los modelos son de unidad subléxica dependientes de la aplicación. Usando este mismo tipo de modelado acústico, RAMSES alcanza también un 1,0%.

Puede concluirse, por tanto, que el sistema de referencia empleado, RAMSES, proporciona en entrenamiento de máxima verosimilitud prestaciones semejantes a los mejores resultados publicados hasta la fecha utilizando este mismo criterio en el entrenamiento de modelos continuos o semicontinuos de Markov, sin que la utilización de estos últimos afecte en demasía a la premisa fundamental de mejorar —utilizando entrenamiento discriminativo— los mejores resultados posibles utilizando entrenamiento de máxima verosimilitud.

1.1.4 Parametrización de la señal y modelado acústico

TIDIGITS y TIMIT están muestreadas a frecuencias diferentes. Es por tanto necesario modificar la frecuencia de muestreo de una o ambas bases para poder trabajar con las dos simultáneamente. Dado que la transformación de frecuencias requiere del uso de filtro interpolador y/o diezmador, si sólo se transforma una de las bases aparecería una discrepancia adicional entre ellas. Para evitarlo se ha optado por interpolar ambas bases hasta 80KHz —interpolando por cuatro TIDIGITIS y por cinco mboxTIMIT—, diezmando a continuación por cinco en ambos casos para dejar el resultado a 16KHz. De este modo el filtro diezmador utilizado es el mismo en los dos casos.

Una vez realizada la transformación de frecuencias a 16KHz, la señal es dividida en segmentos —en adelante, *tramas*— de 30ms tomados cada 10ms. Cada una de las tramas es caracterizada mediante cuatro vectores de caraterísticas:

Espectro: 12 coeficientes cepstrales calculados a partir de la salida de un banco de 24 filtros espaciados uniformemente en la escala Mel. La media del fichero es restada a cada vector de coeficientes para ecualizar las dos bases de datos. (En RAMSES: mMFCC).

ΔEspectro: el resultado de aplicar el operador Δ —equivalente al cálculo de la pendiente

Información	Tamaño del Codebook	Número de Centroides
mMFCC	256	6
d1MFCC	256	6
d2MFCC	256	6
d1E_d2E	128	4

Tabla 1.3: *Parámetros de la cuantificación vectorial utilizada en los experimentos de reconocimiento.*

de la recta de regresión— sobre los vectores de coeficientes cepstrales del espectro. (d1MFCC).

ΔΔEspectro: el resultado de aplicar el operador Δ al ΔEspectro. (d2MFCC).

Energía: la concatenación del resultado de aplicar los operadores Δ y ΔΔ a la sucesión de valores de la energía en cada trama. (d1E_d2E).

Los vectores de caraterísticas son cuantificados vectorialmente con *codebook*'s cuyos agrupamientos se modelan mediante distribuciones gaussianas con matriz de covarianza diagonal. La tabla 1.3 muestra los parámetros de esta cuantificación —tamaño del codebook y número de centroides utilizados en cuantificación semicontinua—.

Habitualmente, cada tipo de información de la parametrización —espectro, energía, etc.— contribuye a la función de distribución de probabilidad de manera independiente [29]. Así, siendo $\mathcal{P}(x(t), \lambda_i^k)$ la probabilidad de que la trama $x(t)$ sea producida por el estado i conociendo únicamente la información f, la probabilidad conjunta para todas las informaciones es:

$$\mathcal{P}(x(t), \lambda_i) = \prod_f \mathcal{P}(x(t), \lambda_i^f) \tag{1.1}$$

Diversos trabajos [84, 35], proponen una alternativa a este planteamiento que conduce a mejoras importantes en las prestaciones del sistema de reconocimiento. La idea consiste en ponderar de manera desigual cada una de las informaciones, utilizando unos *pesos* ó *exponentes*, γ_f, estimados mediante entrenamiento de máxima verosimilitud [35] o discriminativo[84]:

$$\mathcal{P}(x(t), \lambda_i) = \prod_f \mathcal{P}(x(t), \lambda_i^f)^{\gamma_f} \tag{1.2}$$

En los experimentos de referencia, en los cuales se utilizará entrenamiento de máxima verosimilitud, estos pesos valen siempre uno, equivaliendo el producto ponderado de 1.2 al producto habitual de 1.1. En los experimentos de entrenamiento discriminativo, uno de los parámetros reestimados es el valor de los pesos dados a cada información en cada estado de cada unidad.

Los modelos de Markov empleados son de cuatro estados para cada fonema —dos estados por modelo, en el caso de usarse semifonemas—. Sólo se puede entrar al modelo por el primero de los estados, y salir por el último. Desde cualquier estado del modelo sólo es posible acceder al mismo estado o a uno de los que lo siguen —topología *izquierda-derecha*— permitiéndose, como máximo, evitar un estado en cada salto. En el caso de

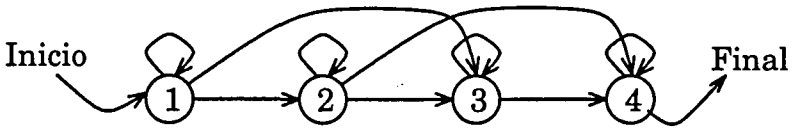

Figura 1.1: *Ejemplo de modelo oculto de Markov de cuatro estados del tipo empleado en el modelado de fonemas. Se permite transitar de un estado al mismo o a alguno de los siguientes, evitando, como máximo, un estado intermedio.*

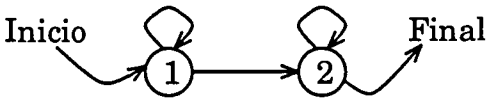

Figura 1.2: *Ejemplo de modelo oculto de Markov de dos estados del tipo empleado en el modelado de semifonemas. Nótese que el modelo equivalente para el fonema —la concatenación de dos modelos de semifonema— es un modelo de cuatro estados, ninguno de los cuáles puede ser evitado.*

los semifonemas, cada modelo es de dos estados, obligándose a visitar ambos de manera consecutiva y sin pemitir el salto desde el último al primero (véase las figura 1.1 y 1.2).

1.1.5 Parámetros optimizados mediante entrenamiento discriminativo

En todos los experimentos presentados en esta tesis se reestiman tres juegos de parámetros de los modelos acústicos:

1. Las probabilidades de emisión de símbolo de cada estado de los modelos (semi-continuos).
2. Las probabilidades de transición entre estados.
3. La ponderación logaritmica de cada una de las informaciones.

Inicialmente se consideró también la reestimación del cuantificador vectorial. Los resultados obtenidos reestimando sólo el cuantificador eran muy superiores a los originales, pero no superaban a los alcanzados cuando sólo se reestimaba el modelo de Markov. El mejor resultado se obtenía, sistemáticamente, reestimando tanto uno como otro, pero el beneficio respecto al caso de sólo reestimar el modelo de Markov era pequeño. Así pues, y dado que reestimar ambos representa unas necesidades de cálculo muy superiores —ya que se debe cuantificar cada señal de entrenamiento cada vez que va a ser usada—, se optó por sólo reestimar los parámetros de los modelos.

1.2 Sistemas de Reconocimiento del Habla

A principios de los 50's se proponen los primeros sistemas de reconocimiento automático del habla, basados en el aprovechamiento de las caraterísticas acústico fonéticas del habla [22, 85, 27]. Son sistemas muy simples cuyo objetivo es el reconocimiento de vocabularios reducidos, donde cada elocución es pronunciada de manera aislada, habitualmente por

un sólo locutor. A partir de estos inicios, y gracias tanto al desarrollo de nuevas y más sofisticadas herramientas matemáticas, como a la disponibilidad de equipos informáticos cada vez más potentes, el reconocimiento del habla ha experimentado una rápida evolución, especialmente desde mediados de los 70's. Esta evolución afecta, entre otros, a dos aspectos fundamentales del problema: el lenguaje de la tarea a reconocer, y las condiciones en que son pronunciadas las elocuciones a reconocer.

En cuanto a el lenguaje de la tarea, ya se comentó que los primeros sistemas propuestos sólo permitían el reconocimiento de palabras aisladas y extraídas de vocabularios de pequeño tamaño —los dígitos, las vocales, ciertas combinaciones silábicas, etc.—. El desarrollo de las técnicas de programación dinámica y, sobre todo, la introducción de los modelos ocultos de Markov, permitieron en la década de los 80's el desarrollo de algoritmos de reconocimiento de palabras conectadas [104, 76, 91]. En todos estos casos, los modelos acústicos son entrenados utilizando frases de la misma tarea que se pretende reconocer. Son lo que, en adelante, se denominará *modelos dependientes de la tarea*[2].

La necesidad de utilizar bases de datos dependientes de la tarea a reconocer implica limitaciones muy serias a la hora de diseñar el sistema de reconocimiento. En primer lugar, los vocabularios a reconocer deben ser de pequeño tamaño puesto que, en caso contrario, las necesidades de material de entrenamiento pueden ser muy elevadas. Por otro lado, es obvio que sólo se puede plantear el reconocimiento, con modelos dependientes de la tarea, de tareas conocidas con anterioridad al propio entrenamiento. Ahora bien, en casi todos los idiomas, las frases y palabras que las forman pueden ser consideradas como la concatenación de sonidos de corta duración y distinguibles entre sí mediante sus características acústicas —cuando menos para los oyentes conocedores del idioma—. Así, es habitual considerar que la sílaba como la unidad acústica básica en las lenguas orientales, mientras que el fonema lo sería para las lenguas indoeuropeas —como el castellano y resto de lenguas románicas, o el inglés y otras lenguas germánicas—. La utilización de estos sonidos, las ya mencionadas *unidades subléxicas*, como mecanismo para abordar el reconocimiento de cualquier tarea, constituye un hito fundamental en la evolución de los sistemas de reconocimiento del habla [20, 119, 65, 58]—.

Una característca fundamental del empleo de modelos de unidad subléxica es que permiten plantear el reconocimiento de cualquier tarea a partir de bases de datos de entrenamiento de propósito general [54]. Son, además, la base de funcionamiento de los sistemas de reconocimiento más ambiciosos en la actualidad: los de reconocimiento del habla espontánea y los sistemas de diálogo [118, 74]. En este tipo de sistemas, el locutor no es forzado a seguir ninguna estructura prefijada en la construcción de las frases a reconocer. Por el contrario, su interacción con el sistema de reconocimiento es natural y es el propio sistema de reconocimiento el encargado de resolver las situaciones no contempladas en el lenguaje.

Por lo que respecta a las condiciones en que se pronuncian las elocuciones a reconocer, han de considerarse múltiples cuestiones: el conjunto de locutores que va a usar el sistema, el modo de articulación de las palabras que forman cada frase, el ambiente de grabación, etc. En cuanto al conjunto de locutores, pueden distinguirse cuatro configuraciones distintas: mono-locutor, cuando el sistema es entrenado con elocuciones del mismo, y único,

[2]Según este criterio, que se mantendrá a lo largo de toda la tesis, son modelos dependientes de la tarea los que son entrenados utilizando bases de datos formadas por frases pertenecientes a la tarea a reconocer, con independecia de que los modelos en sí sean de palabras o no. Así, en el caso del reconocimiento de las cadenas de dígitos, son tan dependientes de la tarea los modelos de dígito entrenados con TIDIGITS, como los de fonema, o cualquier otra unidad subléxica, entrenados igualmente con ella.

locutor que va a usar el sistema [22, 85]; multi-locutor, idéntico al mono-locutor, pero destinado a un conjunto finito y predeterminado de locutores; independiente del locutor, cuando el locutor a reconocer no participa en el entrenamiento de los modelos acústicos [27]; y, finalmente, adaptado al locutor, consistente en un sistema de reconocimiento independiente del locutor modificado de manera que se optimice el reconocimiento para uno o varios locutores. En cuanto al modo de articulación de las palabras, los primeros sistemas de palabras conectadas exigían la presencia de pausas marcadas entre cada dos palabras de la frase a reconocer. En la actualidad, esta restricción tiene cada vez menos interés, y los esfuerzos de investigación suelen ir dirigidos al diseño de sistemas de reconocimiento en los que las frases son pronunciadas de manera natural (habla fluida). Finalmente, también merecen especial interés las condiciones en que se realiza la grabación de las elocuciones a reconocer. Así, la problemática asociada al reconocimiento del habla depende en gran medida del ruido ambiente. También merece especial atención los sistemas destinados al reconocimiento del habla a través del canal telefónico —como es el caso de los experimentos presentados en el apéndice B—.

Reconocimiento del habla continua. El término *habla continua* permite distintos significados en función del criterio de clasificación de los sistemas de reconocimiento empleado. Así, en un principio, habla continua indicaba la inexistencia de pausas entre palabras, en oposición al *habla conectada* que sí las presenta. Hoy en día parece más oportuno referirse a habla cuidadosa, fluida, espontánea, etc. para distinguir el modo de pronunciación, reservando el término habla continua para los sistemas de reconocimiento basados en unidades subléxicas —aunque la tarea a reconocer sea de palabras aisladas o conectadas— en oposición a los sistemas de reconocimiento dependientes de la tarea en los cuales se modela explícita —utilizando modelos de palabra—, o implícitamente —utilizando modelos de unidad subléxica, pero entrenados con una base de datos dependiente de la tarea a reconocer— cada una de las palabras del vocabulario. Los sistemas de reconocimiento del habla continua basados en unidades subléxicas constituyen el máximo reto en modelado acústico y son la puerta de acceso a los sistemas más ambiciosos de reconocimiento del habla espontánea, dictado automático y sistemas de diálogo.

1.2.1 Reconocimiento del habla con patrones

La primera aproximación al problema del modelado acústico en reconocimiento automático del habla consiste en seleccionar un conjunto de elocuciones típicas, llamados *patrones*, para cada palabra del vocabulario a reconocer. En este tipo de sistemas, el reconocimiento se realiza comparando la elocución a reconocer con cada uno de las patrones, dando como resultado la palabra del vocabulario cuyo patrón es más parecido. Como medida del parecido entre elocuciones hay diversas alternativas, siendo una de las más populares la distancia cepstral calculada a partir del modelo de predicción lineal o de un banco de filtros.

La utilización de patrones presenta un problema fundamental: dado que como representante de cada clase se toma una realización concreta, el modelado proporcionado por los patrones será muy dependiente de la duración y los locutores de las realizaciones seleccionadas. La dependencia de la duración —que imposibilita la comparación *trama a trama* de la elocución y el patrón si no son de la misma duración— se resolvió inicialmente aplicando técnicas de normalización de la duración [71], aunque el mayor avance se produce con la introducción de las técnicas de programación dinámica [105]. En cuanto a la

dependencia del locutor, puede evitarse utilizando múltiples patrones para cada palabra, seleccionando cuidadosamente los patrones de manera que se parezcan a las realizaciones del máximo número de locutores, promediando patrones, etc. [94]. Lamentablemente, el incremento en la capacidad de generalización de los sistemas basados en patrones se consigue a costa de un incremento importante de la complejidad del sistema y de sus necesidades de cálculo —que aumentan conforme más patrones son seleccionados para cada clase—, y la introducción de decisiones heurísticas en el algoritmo de programación dinámica —la penalización por no coincidencia temporal entre la elocución a reconocer y el patrón debe ser fijada a priori—.

1.2.2 Reconocimiento del habla con modelos ocultos de Markov

A finales de los 60's se presentan unas estructuras matemáticas que permiten superar las limitaciones de los patrones en su capacidad de generalizar las realizaciones disponibles en el entrenamiento [7, 6]: los modelos ocultos de Markov (**HMM**'s, por sus siglas en inglés). Los modelos ocultos de Markov constituyen un método estadístico en el cual la idea de parecido o distancia, usada en los patrones, es sustituida por la de probabilidad. Así, en lugar de asignar como resultado del reconocimiento la palabra del vocabulario más cercana a la elocución a reconocer, se pretende responder a la pregunta de cuál es la palabra más probable teniendo en cuenta las características acústicas de la elocución. La idea subyacente es considerar la voz como un proceso aleatorio paramétrico modelable como una cadena oculta de Markov cuyo modelo generador es estimable a partir de las muestras de entrenamiento [91].

Una cadena de Markov de orden O, $x = \{x(1)x(2)\dots\}$, es un proceso discreto que sólo puede tomar un número finito de valores —estados— diferentes y tal que su valor en cada instante sólo depende de los O precedentes. En concreto, si el orden es uno, tal y como es habitual en aplicaciones de reconocimiento del habla, entonces el valor del proceso en un instante determinado sólo depende del valor que tenía en el instante anterior. El modelo generador de este tipo de procesos, el modelo de Markov, es una máquina de estados finitos en la que la transición entre dos estados cualesquiera está gobernada por una probabilidad que sólo depende del estado de partida. Una cadena oculta de Markov es un proceso gobernado por un modelo de Markov, pero en el cual no se observa directamente el estado, sino que la observación es una función probabilística del mismo. Un ejemplo típico de modelos y cadenas ocultos de Markov lo constituye el clima. Podemos considerar el clima como un proceso que puede tomar cuatro valores: primaveral, veraniego, otoñal e inviernal. Posibles observaciones del proceso clima son la temperatura media y la cantidad de lluvia recogida durante un día. El modelo de Markov correspondiente tiene cuatro estados, uno para cada tipo de clima. Cada modelo está caracterizado por la probabilidad de tener cada uno de los climas conociendo el del día anterior, y por la probabilidad de que con ese clima se dé la pareja de valores de temperatura y lluvia observados.

Un modelo oculto de Markov, λ, se define mediante tres conjuntos de parámetros: la probabilidad de inicio, π_i, indica la probabilidad de que la cadena de Markov empiece en el estado i; el vector de probabilidades de transición del estado i, $\vec{a_i} = \{a_{ij}\}$, es la probabilidad de que, siendo i el estado de la cadena en el instante t, en el $t+1$ el estado sea j; finalmente, la probabilidad de emisión de símbolo es la función de distribución de probabilidad de las tramas de la observación en el estado, $B_i(x(t))$. Los modelos ocultos de Markov permiten una gran flexibilidad en la definición de la probabilidad de emisión

de símbolo, distinguiéndose los modelos discretos, semicontinuos y continuos de Markov; así como otras estructuras más complejas, como las híbridas formadas por un modelo de Markov dotado de una red neuronal para el cálculo de la probablidad de emisión de símbolo [5, 38, 42, 93].

Los modelos discretos y semicontinuos de Markov se basan en cuantificar vectorialmente las tramas de la observación. Como producto de la cuantificación vectorial, la observación en el instate t, $x(t)$, queda representada mediante una medida de la cercanía de la trama a cada uno de los C símbolos del cuantificador vectorial, $x(t) \rightarrow \vec{w}_t$ —en modelos discretos, sólo el valor del símbolo de centroide más cercano es distinto de cero; en semicontinuos, sólo suelen ser distintos de cero los $C' < C$ símbolos más cercanos—. La función de distribución de probabilidad se calcula utilizando un histograma de las observaciones vistas en el estado, $\vec{b}_i$. Tanto en el caso de los modelos discretos como en el de los semicontinuos, la probabilidad de emisión de símbolo puede expresarse como el producto escalar del vector de pesos de la observación multiplicado por el histograma del estado,

$$B_i(x(t)) = \vec{w}_t \cdot \vec{b}_i = \sum_j w_{tj} b_{ij} \tag{1.3}$$

Los modelos continuos, por el contrario, modelan la probabilidad de emisión de símbolo utilizando una combinación de contribuciones gaussianas distinta en cada estado,

$$B_i(x(t)) = \sum_j c_{ij} \mathsf{N}(x(t), m_{ij}, \sigma_{ij}). \tag{1.4}$$

Donde m_{ij} es la media de la j contribución gaussiana del estado i, σ_{ij}, la desviación típica de esa contribución, y c_{ij} el peso que se le da.

1.2.2.1 Utilización de los modelos ocultos de Markov

Para que los modelos de Markov resulten de utilidad, es necesario dar respuesta a tres cuestiones básicas: la evaluación, el alineado y el entrenamiento.

Evaluación de $\mathcal{P}(x/\lambda)$: probabilidades hacia adelante y hacia atrás. La evaluación consiste en el cálculo de la probabilidad con la que el modelo λ_i genera la observación x, esto es: $\mathcal{P}(x/\lambda_i)$. La estructura de los modelos de Markov permite realizar este cálculo de manera inductiva [7, 91]. El punto de partida es la probabilidad de que el tramo inicial de la observación $X_t = \{x(1)x(2)\ldots x(t)\}, t \leq T$, donde T es la duración de la observación, acabe en cada uno de los S estados del modelo. Es la denominada probabilidad *hacia adelante*, y se expresa habitualmente mediante la letra griega α,

$$\alpha(t,i) = \mathcal{P}(X_t, q_t = i/\lambda), \quad 1 \leq i \leq S \tag{1.5}$$

Esta probabilidad es conocida en el instante inicial, dado que es igual a la probabilidad de que la observación empiece en el estado, π_i, multiplicada por la probabilidad de emisión de símbolo del estado para la primera trama de la observación, $B_i(x(1))$. Además, conociendo su valor en un instante determinado, el cálculo para el instante siguiente es inmediato. Así pues, es posible calcular $\alpha(t,i)$ para todo t e i de manera inductiva. La probabilidad de la

observación entera es, simplemente, la suma de las probabilidades hacia adelante de todos los estados (que pueden ser salida) del modelo en la última trama de la elocución.

$$\mathcal{P}(x,\lambda) = \sum_i \mathcal{P}(X_T, q_t = i/\lambda) = \sum_i \alpha(T,i) \tag{1.6}$$

De manera análoga a las probabilidades hacia adelante, se definen las probabilidades *hacia atrás* como la probabilidad de la parte de la observación desde $t+1$ hasta el final, suponiendo que en t está en un estado determinado.

$$\beta(t,i) = \mathcal{P}(x(t+1)x(t+2)\dots x(T)/q_t = i,\lambda), \quad 1 \le i \le S \tag{1.7}$$

En este caso, el valor es conocido en el instante final de la observación, siendo igual a uno para todos los estados (que pueden ser final) del modelo. También en este caso, los valores para el resto de instantes de tiempo puede ser calculada de manera inductiva. La probabilidad de la observación dado el modelo puede calcularse a partir de las probabilidades hacia atrás en $t = 1$, $\beta(1,i)$, y las probabilidad de inicio, π_i.

$$\mathcal{P}(x/\lambda) = \sum_i \pi_i \beta(1,i) \tag{1.8}$$

Un resultado importante es que, a partir de las probabilidades hacia adelante y hacia atrás, es posible conocer la probabilidad de que una trama de señal ocupe un estado determinado. Esta probabilidad es igual a la probabilidad conjunta de llegar al estado en ese instante y proseguir hasta el final del modelo, a partir de él:

$$\mathcal{P}(x, q_t = i/\lambda) = \alpha(t,i)\beta(t,i), \quad 1 \le i \le S \tag{1.9}$$

Determinación del alineado entre x y λ: el algoritmo de Viterbi. El alineado consiste en determinar la mejor correspondencia entre las tramas de la observación, $x = \{x(1)x(2)\dots x(T)\}$ y los estados del modelo, λ. Esta correspondencia define la secuencia de estados o camino, $Q = (q_1 q_2 \dots q_T)$, óptimo para esa observación y ese modelo. Pueden aplicarse distintos criterios en la definición de la optimalidad del camino. Por ejemplo, podría aplicarse la ecuación 1.9, para calcular el estado que maximiza la probabilidad de cada una de las tramas de señal. El problema de esta definición es que la secuencia de estados no tiene porque ser posible. Esta situación se da cuando es cero la probabilidad de transición entre los estados *óptimos* en las tramas t y $t+1$. En lugar de seleccionar la secuencia óptima como la de estados óptimos para cada trama, es habitual considerar como camino óptimo a aquella sucesión de estados que genera la observación completa con mayor probabilidad. Este camino puede determinarse, de manera muy parecida al cálculo de las probabilidades hacia adelante, aplicando el denominado *algoritmo de Viterbi* [116, 28].

El algoritmo de Viterbi se basa en el uso de la probabilidad del mejor camino parcial, definida como la probabilidad de estar en cada uno de los estados en un instante determinado, suponiendo que el camino seguido desde el principio hasta ese instante es el de máxima probabilidad. Esta probabilidad, como ocurría con la probabilidad hacia adelante, es conocida en el instante inicial, y puede ser calculada para el resto de la observación de manera inductiva y eficiente, amparándose en el hecho que todo subcamino del camino óptimo es, a su vez, óptimo. A partir del conocimiento de las probabilidades del mejor camino parcial para todos los estados del modelo y todas las tramas de la observación,

es posible determinar la secuencia de estados que da lugar con mayor probabilidad a la observación. Así, el estado final del alineado óptimo entre la observación y el modelo es aquél cuya probabilidad del mejor camino parcial en la última trama es máxima. Por otro lado, y cualquiera que sea la trama en la que conocemos el estado del camino óptimo, el estado del camino óptimo en la trama anterior es el penúltimo estado del subcamino óptimo hasta este estado. El proceso de, partiendo del mejor estado en la última trama, determinar la secuencia óptima de estados *recordando* el mejor antecesor del último estado del camino óptimo determinado se denomina *backtracking*.

Entrenamiento de los modelos: algoritmo de Baum-Welch. El entrenamiento de los modelos consiste en el cálculo de sus parámetros a partir de las realizaciones contenidas en una base de datos. El problema del entrenamiento no cuenta con una solución análiticа cerrada que permita garantizar la obtención del óptimo. No obstante, sí existe un algoritmo que permite actualizar los parámetros de λ de manera que se alcance un máximo, al menos local, de su verosimilitud, $\mathcal{P}(x/\lambda)$: el algoritmo de Baum-Welch [7, 6]. El algoritmo de Baum-Welch se basa en la utilización de la función auxiliar de Baum, $Q(\lambda', \lambda) = \sum_q \mathcal{P}(x, q/\lambda') \log \mathcal{P}(x, q/\lambda)$, donde λ' representa los parámetros iniciales del modelo de Markov; λ, los deseados y q cada posible alineado entre la observación y el modelo. Puede demostrarse que un incremento en $Q(\lambda', \lambda)$ implica el incremento de $P(x/\lambda)$. Por tanto, la maximización en λ de la función de Baum implica un aumento de la verosimilitud del modelo. La ventaja de este planteamiento radica en que la función de Baum, contrariamente a lo que sucede con la propia verosimilitud, presenta un único punto crítico, que además es un máximo y que se puede determinar de manera cerrada. Esto garantiza que, en la reestimación iterativa de los parámetros del sistema, la verosimilitud del modelo aumente cada vez hasta alcanzar un máximo, como mínimo, local.

La aplicación del algoritmo de Baum-Welch lleva a que los parámetros del sistema se adapten de manera que reflejen las caraterísticas estadísticas de la población que representan [93]. Así, en cada iteración se calcula la probabilidad de que cada trama visite cada uno de los estados utilizando la ecuación 1.9 con los parámetros actuales. Los nuevos parámetros del modelo se calculan de manera que sean iguales a los derivados de estas probabilidades. Por ejemplo, la probabilidad de transición entre el estado i y el j se hace igual al número esperado de veces que se produce esa transición para el material de entrenamiento dividido por el número esperado de veces que se visita el estado i; y los parámetros de las funciones de emisión de símbolo en cada estado se calculan de manera que aproximen la función de densidad de probabilidad de las tramas de señal que lo visitan.

1.2.2.2 Aplicación de los HMM's al reconocimiento del habla

La aplicación de los modelos ocultos de Markov al reconocimiento del habla se basa en considerar que las observaciones de voz, $x = \{x(1)x(2)\dots x(T)\}$, se comportan como cadenas de Markov cuyos parámetros, λ, son distintos para cada palabra a reconocer $w_i \in W$. La ventaja de este planteamiento es que permite aplicar la regla de decisión de mínimo riesgo de Bayes para determinar la palabra reconocida, w_r, a partir de la probabilidad de la observación dados los distintos modelos, $\mathcal{P}(x/\lambda_i)$, y de la probabilidad a priori de cada palabra del vocabulario, $\mathcal{P}(w_i)$. Así, si efectivamente la observación se comporta como un proceso de Markov del cual conocemos sus parámetros para cada palabra

del vocabulario, λ_i, la regla de decisión de Bayes,

$$w_r : r = \arg\max_i (\mathcal{P}(w_i/x)) \tag{1.10}$$

puede expresarse como,

$$w_r : r = \arg\max_i (\mathcal{P}(w_i)\mathcal{P}(x/\lambda_i)) \tag{1.11}$$

En general, la voz no puede considerarse como un proceso de Markov y, por tanto, la aplicación del criterio de máxima verosimilitud no será equivalente a la regla de Bayes. Esta discrepancia será tratada en mayor profundidad en lo que queda de tesis, dado que da pie a la proposición de criterios alternativos en el entrenamiento, como es el caso del discriminativo. No obstante, los resultados obtenidos utilizando modelos ocultos de Markov entrenados con el algoritmo de Baum-Welch sí permiten concluir que la aproximación es razonablemente válida para muchas aplicaciones de reconocimiento del habla. Así, en ciertas tareas de reconocimiento de palabras aisladas —sí/no, los dígitos, etc.— es posible alcanzar tasas de reconocimiento correcto cercanas al 100%. En otras tareas, aunque las tasas alcanzadas no son tan espectaculares, los resultados también apuntan a su utilidad. Por ejemplo, con el sistema utilizado a lo largo de esta tesis, RAMSES (ver el apartado 1.1.3), el tanto por ciento de cadenas de dígito en inglés reconocidas correctamente utilizando modelos entrenados con el algoritmo de Baum-Welch es de más del 98%, si se utilizan modelos dependientes de la tarea, y del 94%, con modelos de fonema.

En el caso de los sistemas de reconocimiento de palabras aisladas, tanto el material de entrenamiento como el de reconocimiento puede estar formado por frases compuestas cada una de una única palabra. El entrenamiento de los modelos puede ser llevado a cabo utilizando el algoritmo de Baum-Welch directamente; y el reconocimiento puede ser realizado calculando la verosimilitud de cada unidad —utilizando la ecuación 1.6 ó el algoritmo de Viterbi—, y seleccionando la de mayor valor. En el caso de palabras conectadas —y también en el de utilizar unidades subléxicas— ni el material de entrenamiento ni el de reconocimiento está formado por unidades aisladas, sino por cadenas de ellas. Por tanto, el algoritmo de Baum-Welch ya no puede ser aplicado directamente. En su lugar puede optarse por tres soluciones distintas: determinar, automática o manualmente, la descomposición de la frase en sus unidades constitutivas y realizar el entrenamiento considerando cada segmento independientemente; utilizar los modelos disponibles para determinar los límites temporales de cada unidad, o *segmentación en unidades* de las frases, reestimar los modelos utilizando esta segmentación, volver a comenzar calculando de nuevo la segmentación asociada a los nuevos parámetros, e iterar hasta la convergencia, algoritmo *k-means* [43]; finalmente, y dado que la reestimación de los modelos de Markov es esencialmente lo mismo que optimizar el alineado entre los modelos y las frases de entrenamiento, puede plantearse la reestimación conjunta de modelos y segmentación construyendo el modelo compuesto de cada frase de entrenamiento a partir de la concatenación de los modelos de las unidades que la forman. Los parámetros de los modelos son reestimados, entonces, aplicando las ecuaciones de Baum-Welch a partir de los cómputos estadísticos obtenidos utilizando los modelos compuestos, entrenamiento conectado o *embedding training* [78]. Esta estrategía de modelado acústico posibilita el reconocimiento de cualquier tarea, sea de palabras aisladas, conectadas, o cualquier otra configuración.

En cuanto al reconocimiento de palabras conectadas tampoco es viable la aplicación directa de la ecuación 1.6 ya que eso implicaría unas necesidades de cálculo muy elevadas.

En su lugar se suele construir un modelo compuesto a partir de los modelos de todas las palabras en paralelo y formado de manera que, desde los estados correspondientes al final de cualquier palabra, puede saltarse al principio de cualquier otra, *level building* [76]. En el salto entre el final de una unidad y el principio de la siguiente es habitual incoporar algún tipo de penalización, de manera que sólo las frases válidas en la tarea sean permitidas —gramáticas regulares y estocásticas—. Utilizando el algoritmo de Viterbi, es posible conocer la secuencia de estados óptima la cual, al estar asociado cada estado con una única unidad, proporciona la cadena de unidades de máxima verosimilitud.

1.3 Reconocimiento del Habla Continua y Unidades Subléxicas

Existen ciertas tareas de gran importancia y uso frecuente que justifican el empleo de bases de datos de entrenamiento específicas a la tarea a reconocer —por ejemplo, el reconocimiento de dígitos y/o letras aislados o conectados—. Este tipo de tareas puede abordarse de múltiples maneras en función de si se emplean unidades de palabra o de algún tipo de unidad más pequeña, y en función de si se consideran o no los efectos de los contextos entre palabras; pero, en general, su reconocimiento se lleva a cabo a partir de modelos acústicos entrenados utilizando una base de datos específica a la tarea. La utilización de este tipo de bases de datos suele proporcionar resultados muy superiores a las obtenidas con bases de datos de propósito general. Por otro lado, lo habitual del reconocimiento de letras y dígitos hace que el coste de la grabación de bases de datos específicas se mantenga reducido en términos relativos.

Existen otras situaciones, por el contrario, en las que la utilización de una base de datos específica no es posible. Éste es el caso, por ejemplo, si el vocabulario de la tarea a reconocer no es conocido a la hora de realizar el entrenamiento de los modelos acústicos, o si puede variar en el tiempo. Tampoco suele tener interés práctico la grabación de bases de datos específicas para el reconocimiento de tareas muy especializadas, ya que el coste global del sistema depende en gran medida del coste —habitualmente, muy elevado— de grabación de la base de datos. En las situaciones en las que la grabación de bases de datos específicas no es posible o rentable, la única alternativa consiste en el empleo de modelos de unidad subléxica entrenados a partir de bases de datos de propósito general.

1.3.1 Modelado acústico utilizando unidades subléxicas

El modelado acústico utilizando unidades subléxicas consiste en considerar que toda realización oral puede ser descompuesta en una sucesión de sonidos diferentes tales que cada uno de ellos puede ser asociado de manera única a un símbolo —unidad subléxica— tomado de un conjunto finito. Las condiciones que deben cumplir las unidades subléxicas para resultar de utilidad en reconocimiento del habla son:

1. Deben constituir un conjunto finito y completo que permita una transcripción biunívoca de cualquier mensaje oral.

2. Las distintas unidades subléxicas deben ser distinguibles unas de otras a partir de sus caraterísticas acústicas.

3. Las características acústicas de las distintas realizaciones de una unidad acústica no deben depender del contexto concreto en el que se encuentrano.

El cumplimiento de las dos primeras condiciones es necesario para poder plantear el reconocimiento de cualquier mensaje oral a partir del reconocimiento encadenado de cada una de las unidades subléxicas que lo constituyen. Si también se cumple la tercera, entonces los modelos acústicos de las distintas unidades subléxicas serán esencialmente los mismos con independencia de las características de la base de datos de entrenamiento, y los entrenados con una de propósito general —en la cual cada unidad esté suficientemente representada— pueden ser utilizados para el reconocimiento de cualquier tarea.

Desde el punto de vista del entrenamiento y reconocimiento, un sistema de reconocimiento del habla continua basado en unidades subléxicas es equivalente a un sistema de reconocimiento de palabras conectadas, donde las palabras son sustituidas por las unidades correspondientes. Así, el entrenamiento puede llevarse a cabo de las tres maneras habituales en sistemas de palabras conectadas (ver el apartado anterior). En cuanto al reconocimiento utilizando unidades subléxicas, el algoritmo empleado puede variar pero, en todo caso, la base de funcionamiento es equivalente a construir el modelo de cada una de las palabras del léxico de la tarea mediante concatenación de los modelos correspondientes y, a continuación, utilizar los modelos compuestos en un sistema de reconocimiento de palabras conectadas [54].

Casi todos los idiomas permiten la definición de uno o más conjuntos de unidades subléxicas que cumplen, en mayor o menor medida, las tres condiciones arriba expuestas. La elección más inmediata de unidad subléxica para el reconocimiento del habla continua es el fonema. Todo mensaje oral puede ser representado en forma de cadena de fonemas, los cuales presentan características acústicas que, más o menos, permiten su diferenciación del resto. La correspondencia entre un mensaje oral y su transcripción en fonemas es casi biunívoca, sólo rompiéndose la biunivocidad en los casos de homofonía. De todos modos, la homofonía es inherente al propio mensaje y, en general, no depende de la elección de unidad subléxica. Despreciando los casos de homofonía, los fonemas constituyen el conjunto mínimo de unidades que permite representar completa y únicamente cualquier mensaje oral. Por lo tanto, forman el conjunto de unidades que mejor se puede modelar a partir del material de entrenamiento disponible. Sin embargo, el fonema presenta un grave inconveniente que limita sus prestaciones y dificulta su entrenamiento: la dependencia del contexto de sus características acústicas por culpa de la coarticulación.

1.3.2 La coarticulación acústica

La coarticulación es el efecto fonético por el cual las características acústicas de las distintas realizaciones de una cierta unidad fonética dependen del contexto concreto en que se encuentran. Pueden destacarse dos tipos de coarticulación entre sonidos: la coarticulación fonética y la contaminación introducida en el proceso de parametrización de la señal.

La coarticulación fonética consiste en la alteración de las características de un sonido en función de los que lo rodean por culpa de la imposibilidad fisiológica del locutor de cambiar instantáneamente de sonido. Pueden distinguirse dos fuentes distintas de coarticulación fonética: la inercia y la preparación. La inercia es el efecto por el cual ciertas características de un sonido dependen del que lo precede. En castellano hay varios casos de inercia, el más significativo consiste en la realización como aproximantes de las consonantes oclusivas

Unidad	Error	Sust	Inse	Borr	Acierto	Corr
Fonema	2,10	1,02	0,55	0,53	98,44	94,1
Dígito	0,51	0,28	0,11	0,12	99,61	98,4

Tabla 1.4: *Resultados del reconocimiento de TIDIGITS empleando modelos de fonema entrenados con TIMIT, así como de dígito entrenados con el corpus train de la propia TIDIGITS. Sólo se emplean las partes masculinas de los corpora respectivos.*

sonoras (/b/, /d/ y /g/) cuando vienen precedidas de pausa o consonante nasal. Por su parte, la preparación consiste en la anticipación de ciertas características del sonido siguiente. Por ejemplo, la asimilación, por las consonantes nasales en posición implosiva, del punto de articulación de la consonante que las sigue, o la sonorización del fonema /s/ cuando precede a consonante sonora.

Por otro lado, y aunque no se trata de un efecto fonético, el proceso de parametrización de la señal también provoca resultados semejantes a los de la coarticulación debido a la longitud finita de las ventanas de análisis. Así, y aún en el caso de transiciones *puras* entre unidades no coarticuladas fonéticamente, la parametrización de las tramas de señal cercanas a la transición puede depender simultáneamente de ambas unidades. Este efecto es especialmente significativo en el caso de características dinámicas —parámetros delta— o filtradas temporalmente [87]. Por ejemplo, la respuesta impulsional del operador segunda delta presenta un duración típica de unos 60ms, a lo cual habría que añadir la propia longitud de la ventana usada en la extracción de la característica estática. Dado que la duración media de los fonemas ronda los 100ms, buena parte de las tramas de señal parametrizada abarcarán segmentos de voz correspondientes a más de un fonema.

1.3.3 El fonema, el bifonema y el trifonema

La solución más sencilla al problema de la coarticulación consiste en garantizar que los modelos acústicos correspondientes a cada fonema son entrenados con suficientes realizaciones de cada uno de los posibles contextos en los cuales puede aparecer [54]. De este modo se pretende que el modelo *aprenda* a distinguir la unidad fonética en cualquier condición. La bondad de esta aproximación —en adelante: fonemas independientes del contexto o, simplemente, fonemas— queda de manifiesto si tenemos en cuenta que en reconocimiento libre de fonemas independientes del contexto y locutor se alcanza con facilidad una exactitud cercana al 70% en castellano —sobre un conjunto de 23 fonemas—, y al 60% en inglés —sobre un conjunto de 39—. No obstante, sus limitaciones son también evidentes al compararse las prestaciones alcanzadas en el reconocimento de tareas concretas con las que se alcanzan utilizando unidades dependientes de la aplicación. Por ejemplo, el tanto por ciento de cadenas de dígitos del corpus de reconocimiento de TIDIGITS reconocidas incorrectamente al utilizar modelos de fonema independientes del contexto entrenados con TIMIT se sitúa en torno al 5,9%. Si los modelos son de dígito y entrenados utilizando el corpus de entrenamiento de la propia TIDIGITS, la tasa de error baja hasta un 1,6% (ver la tabla 1.4).

Una alternativa a las unidades independientes del contexto consiste en modelar éste de manera explícita, utilizando unidades dependientes del contexto. En esta aproximación cada unidad subléxica es representada de manera independiente para cada uno de los

contextos distintos en los que puede aparecer [108, 52]. El objetivo consiste en garantizar que cada modelo acústico representa a un conjunto homogéneo de realizaciones cuyas características acústicas no dependen de ningún contexto más alejado de los explícitamente representados. Aparece un compromiso entre la precisión en la representación acústica —que podemos suponer aumenta con la longitud del contexto considerado— y la capacidad de entrenamiento de las unidades —conforme aumenta la longitud de los contextos representados, su número aumenta de manera exponencial, y la frecuencia de aparición de cada unidad disminuye en la misma medida—. Dos soluciones a este compromiso han sido ampliamente adoptadas: el bifonema y el trifonema [9]. En el caso del bifonema, el contexto considerado para cada fonema consiste únicamente de un fonema: bien el inmediatamente anterior, dando lugar a bifonemas por la izquierda; bien el inmediatamente posterior, bifonemas por la derecha. En el caso del trifonema, el contexto considerado abarca tanto el fonema inmediatamente anterior como el posterior.

1.3.4 El semifonema

Los resultados obtenidos usando bifonemas y trifonemas corroboran la importancia de incorporar la dependencia del contexto en los sistemas de reconocimiento del habla continua basados en unidades subléxicas. Sin embargo, la pequeña diferencia en los resultados obtenidos utilizando bifonemas o trifonemas, por ejemplo: en [53], apunta a que las mejores prestaciones de las unidades dependientes del contexto es debida, fundamentalmente, a que éstas modelizan explícitamente las transiciones. Así, tanto el bifonema como el trifonema aportan un modelado específico de cada transición. Sin embargo, la aportación específica del trifonema —la dependencia respecto a los otros contextos de los fonemas que forman la transición— se ve contrarrestada, en caso de existir, por su menor capacidad de entrenamiento.

Varios argumentos apoyan la importancia del modelado explícito de las transiciones entre fonemas:

1. Los efectos de coarticulación fonética no siempre son significativos pero, cuando lo son, suelen afectar en mayor medida a la zona de transición entre fonemas.

2. La contaminación debida al proceso de parametrización de la señal afecta sistemáticamente a la zona de transición. El resto del fonema puede verse o no afectado, en función de la longitud de las ventanas de análisis empleadas, pero la transición siempre lo estará.

3. La zona de transición entre fonemas es fundamental para la inteligibilidad del mensaje oral, siendo su modelado la base de la mayor parte de sistemas de síntesis de voz utilizados en la actualidad.

Si se admite que la ventaja fundamental de la dependencia del contexto radica en su modelado explícito de las transiciones entre fonemas, una unidad más conveniente que tanto el bifonema como el trifonema, para realizar este modelado es el semifonema [67, 68]. El semifonema es el producto de considerar que cada fonema A puede ser dividido en dos partes tales que la influencia de los sonidos que lo siguen no es remarcable en la primera de las partes, y la de los que lo preceden, en la segunda. Así, si el fonema A está precedido por el fonema B, y seguido del C, su descomposición en semifonemas será un semifonema inicial que tiene en cuenta la vecindad del fonema precendente, B–A, y uno final que tiene

en cuenta la vecindad del siguiente, A+C. Por tanto, el fonema *A* precedido de *B* y seguido por *C* da lugar a A ↔ B–A A+C.

La ventaja de este planteamiento es que podemos desvincular los efectos de los dos contextos del fonema, dándoles un tratamiento diferenciado en cada uno de los dos semifonemas. Desde el punto de vista de los semifonemas, y despreciando el efecto de los sonidos situados a más de un fonema de distancia, podemos distinguir tres tipos de contexto:

1. Dado que cada fonema es dividido en dos semifonemas, todo semifonema está en contacto, en uno de sus extremos, con otro semifonema correspondiente al mismo fonema. Denominamos a este contexto *contexto interno.*

2. En el extremo opuesto al contexto interno, el semifonema está en contacto con un fonema distinto. Es el *contexto inmediato.*

3. Finalmente, el fonema del que forma parte el semifonema está también en contacto con otro fonema, pero éste no lo está con el propio semifonema. Es su *contexto lejano.*

El efecto de la coarticulación provocada por cada uno de estos contextos es diferente. Así, el contexto interno no introduce ningún tipo de coarticulación, ya que se trata siempre del mismo fonema. Por otro lado, el contexto lejano sólo será relevante en aquellos casos en los que la coarticulación fonética abarca todo el fonema o éste es más corto que las ventanas de análisis utilizadas. Finalmente, un semifonema determinado siempre está en contacto con el mismo contexto inmediato, perteneciente a un fonema distinto. Despreciando los efectos de la coarticulación del contexto lejano frente al los del inmediato, se llega a una caraterización en la que cada fonema es dividido en un semifonema inicial que sólo depende del fonema anterior, y un semifonema final, que sólo lo hace del siguiente.

Desde el punto de vista de la transición, ésta queda modelada de manera biunívoca: cada aparición de la transición entre los fonemas A y B es representada mediante la concatenación de dos modelos que, a su vez, sólo pueden aparecer juntos y cuando sucede esa transición —A B ↔ A+B A–B—. Esta univocidad implica un aprovechamiento óptimo del material disponible en el entrenamiento de las transiciones, dado que el modelo conjunto de cada una —la concatenación de los dos semifonemas correspondientes— es entrenado con todas sus apariciones en la base de datos. Comparado con el semifonema, el trifonema sólo aporta el modelado explícito del contexto lejano, pero el diezmado del material de entrenamiento que introduce repercute en un peor modelado de los contextos cercanos y, por tanto, de las transiciones entre fonemas. En tanto en cuanto la explicitación de los contextos lejanos de la transición no compense su peor modelado acústico el semifonema proporcionará una mejor caraterización acústica que el trifonema.

Las principales ventajas del semifonemas frente a otros planteamientos basados en unidades subléxicas dependientes del contexto son:

1. Caracterización completa y biunívoca de las transiciones entre fonemas, aprovechando al máximo su aparición en el entrenamiento.

2. Modelado parcial de los efectos de coarticulación sobre los fonemas, incorporando explícitamente los más relevantes —aquéllos que afectan a la zona más próxima a la transición entre fonemas—.

Unidad	Error	Sust	Inse	Borr	Acierto	Corr
Fonema	2,10	1,02	0,55	0,53	98,44	94,1
Semifonema	1,27	0,45	0,41	0,41	99,14	96,3
Dígito	0,51	0,28	0,11	0,12	99,61	98,4

Tabla 1.5: *Resultados del reconocimiento de TIDIGITS empleando modelos de fonema y semifonema entrenados con TIMIT, así como de dígito entrenados con el corpus train de TIDIGITS. En estos experimentos se emplea el corpus completo de TIMIT, y no sólo el corpus train, para evitar la aparición de sobre-adaptación en el entrenamiento de los semifonemas.*

3. Al reducirse la amplitud del contexto caracterizado explícitamente —para cada semifonema sólo se considera el fonema vecino—, la disponibilidad de material de entrenamiento es muy superior al caso del trifonema.

4. Homogeneidad tanto en el modelado de los fonemas, como de las transiciones entre ellos.

5. Tratamiento simétrico de la representación de los contextos entre palabra, con independencia de que éstos se modelen explícitamente o no.

Los resultados publicados en distintas tareas en castellano [67, 69], demuestran la utilidad del uso del semifonema, proporcionando prestaciones superiores a las obtenidas con trifonemas, tanto seleccionados por umbral como por árboles de decisión. En el caso de la tarea abordada en esta tesis, el reconocimiento de TIDIGITS utilizando modelos acústicos entrenados a partir de TIMIT, los resultados son semejantes. La tabla 1.5 muestra los resultados del reconocimiento de las cadenas de dígitos de TIDIGITS utilizando tanto fonemas independientes del contexto como semifonemas, así como, para facilitar las comparaciones, los obtenidos utilizando modelos de dígito entrenados a partir del corpus de entrenamiento de la propia TIDIGITS. Como puede observarse en la tabla 1.5, el empleo de semifonemas dependientes del contexto proporciona una mejoría de casi el 40% en la tasa de error con respecto al resultado obtenido con fonemas independientes del contexto. Este resultado es muy interesante dado que, anteriormente, el semifonema sólo había sido probado en tareas de reconocimiento en castellano, una lengua con una fonética totalmente distinta a la del inglés. Así, mientras en la primera las elocuciones son fácilmente analizables en forma de cadena de fonemas relativamente poco coarticulados, en inglés los sonidos tienden a articularse en forma de grupos de fonemas de fronteras poco definidas. En cualquier caso, es de señalar que la utilización de modelos acústicos dependientes de la tarea —dígitos entrenados con TIDIGITS, en este caso— sigue proporcionando prestaciones muy superiores, aunque esta superioridad seguramente está tan relacionada con el mejor modelado proporcionado por los modelos dependientes de la tarea, como con la mayor coherencia entre las bases de datos utilizadas en entrenamiento y reconocimiento.

1.4 Entrenamiento Discriminativo para Reconocimiento del Habla

La gran popularidad de la utilización de modelos ocultos de Markov en el reconocimiento del habla se debe, en gran medida, a la conveniencia del criterio aplicado en su entrenamiento: la maximización de la verosimilitud. Este criterio proporciona una aproximación razonable al criterio de mínimo error de clasificación, siendo, al mismo tiempo, fácil de optimizar. Esta facilidad no se basa únicamente en la existencia de algoritmos muy potentes de entrenamiento —el adelante-atrás de Baum-Welch— sino también en el hecho que el modelo producido para una cierta unidad sólo depende de las características acústicas de los segmentos de voz cuyo contenido fonético es esa misma unidad. Esta propiedad permite la realización de entrenamiento conectado —o *embedded*—; así como la utilización de bases de datos de entrenamiento independientes de la tarea a reconocer.

Ahora bien, el criterio de máxima verosimilitud sólo es una aproximación del de mínimo error. En concreto el criterio de máxima verosimilitud es equivalente al de mínimo error si y sólo si el modelo utilizado es el correcto y, además, somos capaces de estimar correctamente sus parámetros. Pero la voz no es un proceso de Markov. Aún en el caso de que se disponga de la cantidad de material de entrenamiento suficiente para garantizar la práctica corrección de sus parámetros, el modelo empleado será incorrecto y no se podrá garantizar el cumplimiento del criterio de mínimo error. En estas condiciones puede ocurrir que exista un conjunto de parámetros tal que, aún sin cumplir el criterio de máxima verosimilitud, esté más cerca de cumplir el de mínimo error que aquél.

En sus distintas versiones, se ha dado en llamar como métodos de *entrenamiento discriminativo* (**ED**), a aquellos métodos tendentes a obtener modelos acústicos —sean éstos patrones, modelos de Markov o cualquier otra cosa— que sean lo más cercanos posible a la situación de mínimo error.

1.4.1 Sistemas de entrenamiento discriminativo

Los primeros sistemas de entrenamiento discriminativo propuestos se basan en la teoría de los clasificadores lineales. Entre ellos destacan dos: los algoritmos de reestimación de los cuantificadores lineales basados en el algoritmo *Learning Vector Quantisation*, **LVQ** [48]; y el entrenamiento correctivo [4, 2]. El LVQ ha sido empleado con éxito en el diseño de cuantificadores vectoriales que, integrados en sistemas de reconocimiento basados en modelos semi-continuos de Markov, proporcionan prestaciones muy superiores a las obtenidas con cuantificadores calculados con el algoritmo de Lloyd [45]. En este trabajo, el entrenamiento discriminativo sólo es utilizado en el diseño del cuantificador vectorial. Una vez obtenido éste, el entrenamiento de los modelos de Markov se lleva a cabo de la manera tradicional, utilizando el algoritmo de Baum-Welch. El algoritmo LVQ ha sido aplicado también en la reestimación de los modelos acústicos en sistemas de reconocimiento de palabras aisladas [30]. En este caso, el sistema de reconocimiento de palabras aisladas se equipara a un cuantificador vectorial cuyos parámetros son reestimados con LVQ3, obteniéndose unos resultados claramente superiores tanto a los obtenidos con modelos de máxima verosimilitud, como con modelos estimados utilizando entrenamiento correctivo. Por su parte, este último es una extensión heurística de las técnicas empleadas en problemas de clasificación lineal. A pesar de que se carece de una justificación teórica, los

resultados obtenidos son muy superiores a los alcanzados con entrenamiento de máxima verosimilitud.

En esta tesis no se considerará ninguno de estos dos métodos de entrenamiento discriminativo debido a que, aún cuando presentan un innegable interés en su aplicación a sistemas de reconocimiento dependiente de la tarea, su adecuación al entrenamiento de modelos acústicos de unidades subléxicas para su aplicación a sistemas de reconocimiento del habla continua está limitada por dos motivos: en primer lugar, su base de funcionamiento, la optimización de clasificadores lineales, puede representar adecuadamente cuantificadores vectoriales o sistemas de reconocimiento de palabras aisladas, pero guarda poca relación con los mecanismos involucrados en el reconocimiento de tareas concretas de habla continua; por otro lado, ambas técnicas han quedado relegadas a un segundo plano con la propuesta del entrenamiento de mínimo error de clasificación [41], el cual no sólo aborda directamente la regla de decisión realmente empleada durante el reconocimiento, sino que, además, presenta una fundamentación teórica sólida.

1.4.2 Entrenamiento discriminativo basado en la optimización de funciones de calidad

El entrenamiento de mínimo error de clasificación (**MCE**) [41], es semejante tanto al de máxima información mutua (**MMIE**) [3], como al de mínima confusibilidad [81] —propuesta propia de esta tesis—. En los tres casos se define una función continua y diferenciable que representa la calidad del sistema —en general, la función de coste—. A continuación se minimiza la función de coste con la ayuda de un algoritmo de optimización de funciones de múltiples variables. En el caso de la máxima información mutua, la función de coste es una medida derivada de la teoría de la información que representa el grado de ambigüedad entre el contenido fonético de una frase y su caracterización acústica. En los otros dos casos se utiliza una medida *suavizada* del número de errores que el sistema puede cometer —el número total de errores, en entrenamiento de mínimo error de clasificación; y el número de errores posibles, en el de mínima confusibilidad—.

Dada una elocución de entrenamiento $x_n \in X$ —que puede tratarse de una trama de señal, unidad subléxica, cadena de unidades o frase completa; pero a la cual nos referiremos, en general, como palabra—, otorgamos a cada posible reconocimiento de x_n permitido por la gramática W una puntuación de asignación $g_i(x_n, \Lambda) = g_i(x_n, \lambda_i)$ donde $\lambda_i \in \Lambda$ es el modelo acústico de la palabra reconocida $w_i \in W$. En el caso de modelos ocultos de Markov, $g_i(x_n, \lambda_i)$ suele ser la probabilidad de x_n dado λ_i, $g_i(x_n, \lambda_i) = \mathcal{P}(x_n/\lambda_i)$, o bien, tal y como será el caso en lo que sigue, su logaritmo. Para simplificar el cálculo de la probabilidad, suele considerarse sólo el alineado óptimo entre la elocución y el modelo $Q(x_n, \lambda_i)$. Así pues, la expresión de la puntuación de asignación habitualmente empleada es:

$$g_i(x_n, \lambda_i) = \log \mathcal{P}(x_n/\lambda_i, Q(x_n, \lambda_i)) \tag{1.12}$$

Aplicando como regla de decisión del reconocimiento el criterio de máxima verosimilitud, la palabra reconocida por el sistema, w_r, al procesar la elocución x_n es:

$$w_r \ : \ r = \arg\max_k \{g_k(x_n, \lambda_k)\} \tag{1.13}$$

Por tanto, si x_n es un representante de la palabra w_i, lo cual denotaremos mediante[3] x_n^i, entonces el resultado del reconocimiento es correcto si y sólo si $w_i : i = \arg\max_k\{g_k(x_n^i, \lambda_k)\}$. A partir de la puntuación asignada a la palabra correcta y cada una de sus posibles confusiones se construye una función de coste de la elocución, $l(x_n^i, \Lambda)$, que representa de algún modo la probabilidad de reconocerla incorrectamente[4]. La función de coste del sistema, $\mathcal{L}(X, \Lambda)$, es igual a la suma de las funciones de coste calculadas para cada una de las elocuciones que participan en el entrenamiento:

$$\mathcal{L}(X, \Lambda) = \sum_n l(x_n, \Lambda) \tag{1.14}$$

Con la notación empleada para las elocuciones de entrenamiento correspondientes a cada una de las palabras, y teniendo en cuenta que x_n^i sólo está definido para aquéllas cuyo contenido acústico es w_i, la función de coste global del sistema puede reescribirse en la forma de un doble sumatorio —para cada palabra de la tarea y todas las elocuciones de entrenamiento pertenecientes a la misma—:

$$\mathcal{L}(X, \Lambda) = \sum_i \sum_n l(x_n^i, \Lambda) \tag{1.15}$$

Finalmente, tanto en mínimo error de clasificación como en máxima información mutua, esta función de coste global es minimizada mediante algún algoritmo de optimización —generalmente, uno basado en búsqueda de gradiente—. La principal diferencia entre estos dos tipos de entrenamiento discriminativo radica en la manera como se construye la función de coste para cada una de las elocuciones que participan en el entrenamiento.

1.4.2.1 Entrenamiento de mínimo error de clasificación

En entrenamiento de mínimo error de clasificación, la función de coste de la elocución se construye de manera que, siendo continua, vale apróximadamente uno si existe alguna palabra con mayor verosimilitud que la correcta, y cero en caso contrario. Se trata, por tanto, de una función de cómputo de error. La función de coste se construye en dos pasos: en primer lugar se determina, entre todos los posibles impostores, aquél de mayor verosimilitud. Es necesario, por tanto, aplicar el operador $\max(\cdot)$. Seguidamente, la función de cómputo de error es equivalente al operador *menor que* aplicado a las verosimilitudes de la palabra correcta y de la hipótesis incorrecta de mayor verosimilitud.

$$l(x_n^i, \Lambda) = 1\left(\mathcal{P}(x_n^i/\lambda_i) < \max_{j \neq i}\{\mathcal{P}(x_n^i/\lambda_j)\}\right) \tag{1.16}$$

Donde el operador $1(\cdot)$ devuelve uno si su argumento es cierto y cero en caso contrario. Ahora bien, esta definición de la función de coste presenta dos inconvenientes que desaconsejan su uso. En primer lugar, el operador $\max(\cdot)$ sólo depende del valor de su

[3]El superindice en x_n^i debe considerarse como un atributo de x_n y no como un ordinal en X. En las expresiones donde aparezca esta expresión, sólo hay que considerar los términos para los cuales x_n es una realización de w_i.

[4]En el caso del entrenamiento de máxima información mutua, la función construida es una medida de calidad, no de coste. No obstante, basta con sustituir esta medida por su opuesto —en cierto sentido, una medida de la ambigüedad en la caracterización acústica del contenido fonético—, para convertir el problema en la minimización de una función de coste.

argumento máximo. Por tanto, la única hipótesis útil en entrenamiento será la primera, no aprovechándose el conocimiento de otras hipótesis de menor verosimilitud. Para permitir que otras hipótesis participen en el entrenamiento se utiliza una versión suavizada del operador max(·):

$$\max_{j\neq i}\{\mathcal{P}(x_n^i/\lambda_j)\} \approx \mathcal{P}_M(x_n^i,\Lambda) = \left[\frac{1}{|W|-1}\sum_{j\neq i}\mathcal{P}(x_n^i/\lambda_j)^\eta\right]^{\frac{1}{\eta}} \tag{1.17}$$

Donde $|W|$ es el número de palabras en el vocabulario W, y η es un parámetro de control mayor que cero que determina el grado de aproximación al operador max(·): conforme aumenta de valor, el error en la aproximación disminuye, llegando a anularse para $\eta \to \infty$. A partir de la probabilidad aproximada de la hipótesis incorrecta de verosimilitud máxima y de la verosimilitud de la palabra correcta, se define la medida de error de clasificación como la diferencia entre sus logaritmos: $d_i(x_n^i,\Lambda) = -g_i(x_n^i,\lambda_i) + \log \mathcal{P}_M(x_n^i,\Lambda)$. Si esta medida es negativa —$g_i(x_n^i,\lambda_i) > \log \mathcal{P}_M(x_n^i,\Lambda) \approx \max_{j\neq i}\{\log \mathcal{P}(x_n^i/\lambda_j)\}$—, la elocución es reconocida correctamente. Por contra, si $d_i(x_n^i,\lambda)$ es positiva, la elocución es reconocida incorrectamente. Por tanto, la condición de error puede representarse mediante el resultado de comparar la medida de error de clasificación con cero. Ahora bien, la comparación con cero no es una función continua y su empleo dificultaría enormemente la optimización de la función de coste. En su lugar se utiliza la sigmoide, definida como:

$$1\{\mathcal{P}_M(x_n^i,\Lambda) > \mathcal{P}(x_n^i/\lambda_i)\} = 1\{d_i(x_n^i,\Lambda) > 0\} \approx \frac{1}{1+e^{-\gamma d_i(x_n^i,\Lambda)}} \tag{1.18}$$

Aunque ésta es la definición habitualmente utilizada, en esta tesis se expresará de forma ligeramente distinta echando mano de las funciones hiperbólicas:

$$1\{\mathcal{P}_M(x_n^i,\Lambda) > \mathcal{P}(x_n^i/\lambda_i)\} = 1\{d_i(x_n^i,\Lambda) > 0\} \approx \frac{1}{2}\left(1+\tanh\frac{d_i(x_n^i,\Lambda)}{d_0}\right) \tag{1.19}$$

Esta segunda notación —idéntica a la anterior para $d_0 = \gamma^{-1}$— presenta dos ventajas: proporciona una notación más compacta tanto de la función de coste como de sus derivadas; y está gobernado por un parámetro de significado más evidente. Así, si la medida del error de clasificación es muy superior en valor absoluto a d_0, la sigmoide se comporta como un limitador; si es de valor absoluto muy inferior, su comportamiento es casi lineal. Dado que la función de coste sólo es sensible a variaciones en los parámetros si su gradiente es distinto de cero, y éste se anula cuando $|d_i(x_n^i,\Lambda)| \gg d_0$, el efecto de elegir una d_0 finita es el de eliminar del entrenamiento discriminativo tanto las elocuciones para las cuales el margen de protección frente al error es muy grande, como aquéllas en las que el error es inevitable.

Con estas dos alteraciones sobre la fórmula teórica de la función de cómputo de error de elocución (1.16), la función de coste[5] utilizada en entrenamiento de mínimo error de

[5]Esta expresión proporciona la contribución de una realización concreta, x_n^i, de la palabra w_i. La función de coste del sistema se obtiene sumando para todas las realizaciones de cada una de las palabras que participan en el entrenamiento (ecuación 1.15).

clasificación queda:

$$\begin{aligned} l(x_n^i, \Lambda) &= \frac{1}{2}\left(1 + \tanh \frac{d_i(x_n^i, \Lambda)}{d_0}\right) \\ d_i(x_n^i, \Lambda) &= -\log \mathcal{P}(x_n^i/\lambda_i) + \frac{1}{\eta} \log \left(\frac{1}{|W|-1} \sum_{j \neq i} \mathcal{P}(x_n^i/\lambda_j)^\eta \right) \end{aligned} \tag{1.20}$$

La minimización de esta expresión lleva, en [19], a una reducción de más del 25% de la tasa de error en el reconocimiento de la tarea de TIDIGITS —aunque utilizando modelos de dígito dependientes del contexto y entrenados con el corpus de train de la propia TIDIGITS—. En este experimento, el resultado base, con modelos de máxima verosimilitud, es del 0,97% de cadenas erróneas, reestimando los modelos según el criterio de mínimo error de clasificación, la tasa de error baja hasta el 0,72%.

1.4.2.2 Entrenamiento de máxima información mutua

Dada la elocución x_n^i, representante de la palabra w_i perteneciente al vocabulario W, y el modelo de ésta $\lambda_i \in \Lambda$, la información mutua entre la observación y su contenido se define como [3]:

$$I_\Lambda(x_n^i, W) = \log \frac{\mathcal{P}(x_n^i, \lambda_i)}{\mathcal{P}(x_n^i)\mathcal{P}(w_i)} = \log \mathcal{P}(x_n^i/\lambda_i) - \log \sum_k \mathcal{P}(x_n^i/\lambda_k)\mathcal{P}(w_k) \tag{1.21}$$

Donde el primer sumando es la puntuación de asignación otorgada a la palabra correcta, y el segundo es una medida de la ambigüedad con la que se reconoce la hipótesis correcta. Así, la información mutua es máxima si la única hipótesis con verosimilitud no nula es la correcta, decreciendo conforme otras hipótesis presentan verosimilitudes cercanas a la de ésta. La expresión de la información mutua en (1.21) es muy semejante a la de la función de cómputo de error empleada en entrenamiento de mínimo error de clasificación (1.20), cambiada de signo. En tanto en este último el objetivo es maximizar la diferencia entre la verosimilitud de la palabra correcta y una aproximación de la del error más probable, en máxima información mutua el objetivo es maximizar la diferencia entre la verosimilitud de la correcta y un valor de fondo calculado a partir de las verosimilitudes y probabilidades *a priori* de cada una de las palabras del vocabulario. Esta verosimilitud de fondo es igual al valor esperado de la probabilidad de generación de la elocución de entrenamiento suponiendo que el sistema no proporciona ninguna información acerca del contenido de la palabra a reconocer a partir de su modelado acústico.

En [84] se muestra la aplicación del entrenamiento de máxima información mutua a la reestimación de los modelos de dígito empleados en el reconocimiento de TIDIGITS. En este caso, el 1,01% de error inicial del sistema de referencia entrenado con Baum-Welch se reduce al 0,89% —una reducción del 12%—.

1.4.2.3 Sensibilidad de MCE y MMIE. El asombrado de hipótesis

Una característica común al entrenamiento de mínimo error de clasificación y de máxima información mutua es su tendencia a despreciar la influencia de las hipótesis erróneas de verosimilitud inferior a la de mayor valor. Este efecto, en adelante *asombrado de hipótesis*, se pone de manifiesto analizando la sensibilidad de las funciones empleadas en ambos

métodos. La sensibilidad de una función con respecto a un cierto conjunto de parámetros es igual a la norma del gradiente de su logaritmo. Dado que el gradiente del logaritmo de una función es igual al gradiente de la función dividido por su propio valor, la sensibilidad de una función es indicativo del beneficio relativo que podemos esperar obtener alterando los valores del conjunto de parámetros involucrado.

Sensibilidad de la función de cómputo de error en MCE. La sensibilidad de la función de cómputo de error con respecto a los modelos correspondientes a los distintos errores determinados en el reconocimiento de las N hipótesis más verosímiles para la secuencia de entrenamiento x_n^i es:

$$
\begin{aligned}
S_{\lambda_{j\neq i}}(l(x_n^i,\Lambda))\big| &= \left\|\nabla_{\lambda_j}\log l(x_n^i,\Lambda)\right\|_{j\neq i} = \\
&= \left\|\frac{\nabla_{\lambda_j} l(x_n^i,\Lambda)}{l(x_n^i,\Lambda)}\right\|_{j\neq i} \\
&= \frac{1}{2d_0 l(x_n^i,\Lambda)\cosh^2(d_i(x_n^i,\Lambda)/d_0)}\left\|\nabla_{\lambda_j} d_i(x_n^i,\Lambda)\right\|_{j\neq i} \\
&= \frac{1}{2d_0 l(x_n^i,\Lambda)\cosh^2(d_i(x_n^i,\Lambda)/d_0)}\left\|\nabla_{\lambda_j}\log \mathcal{P}_M(x_n^i,\Lambda)\right\|_{j\neq i} \\
&= \frac{1}{2d_0 l(x_n^i,\Lambda)\cosh^2(d_i(x_n^i,\Lambda)/d_0)}\frac{1}{(|W|-1)\eta\left[\mathcal{P}_M(x_n^i,\Lambda)\right]^\eta}\left\|\nabla_{\lambda_j}\mathcal{P}(x_n^i,\lambda_j)^\eta\right\|_{j\neq i} \\
&= \frac{1}{2d_0(|W|-1)l(x_n^i,\Lambda)\cosh^2(d_i(x_n^i,\Lambda)/d_0)}\left[\frac{\mathcal{P}(x_n^i,\lambda_j)}{\mathcal{P}_M(x_n^i,\Lambda)}\right]^\eta\left\|\nabla_{\lambda_j}\log\mathcal{P}(x_n^i,\lambda_j)\right\|_{j\neq i} \\
&= \mathcal{F}(d_i(x_n^i,\Lambda))\left[\frac{\mathcal{P}(x_n^i,\lambda_j)}{\mathcal{P}_M(x_n^i,\Lambda)}\right]^\eta S_{\lambda_j}(\mathcal{P}(x_n^i,\lambda_j))\Big|_{j\neq i} \qquad (1.22)
\end{aligned}
$$

En la expresión (1.22) se pone de manifiesto la dependencia de la sensibilidad de la función de cómputo de error en MCE de tres factores de comprotamiento muy distinto. El primero de los factores, $\mathcal{F}(d_i(x_n^i,\Lambda)) = 1/2d_0(|W|-1)l(x_n^i,\Lambda)\cosh^2(d_i(x_n^i,\Lambda)/d_0)$, sólo depende de la medida de error de clasificación $d_i(x_n^i,\Lambda)$ y es común, por tanto, a todas las hipótesis erróneas consideradas para la realización x_n^i. El término fundamental de este primer factor es $1/\cosh^2(d_i(x_n^i,\Lambda)/d_0)$. Este término tiende a cero para $|d_i(x_n^i,\Lambda)| \gg d_0$ alcanzando su valor máximo en $d_i(x_n^i,\Lambda) = 0$, es decir, cuando la estimación de la verosimilitud de la hipótesis más probable es igual a la de la palabra correcta $g_i(x_n^i,\lambda_i) = \log\mathcal{P}_M(x_n^i,\Lambda)$ (ver la figura 1.3). Como consecuencia, la contribución al entrenamiento de aquellas elocuciones de entrenamiento para las que $d_i(x_n^i,\Lambda) \gg d_0$ es prácticamente cero, y otro tanto ocurre con aquéllas tales que $d_i(x_n^i,\Lambda) \ll d_0$. El efecto de este término es centrar los esfuerzos de la reestimación en la eliminación de los errores que pueden ser evitados con una menor perturbación de los parámetros del sistema.

El segundo factor en (1.22), $\left[\mathcal{P}(x_n^i,\lambda_j)/\mathcal{P}_M(x_n^i,\Lambda)\right]^\eta$, actúa como un *distribuidor* del esfuerzo de reestimación entre las distintas hipótesis consideradas, siendo máximo para la hipótesis de mayor verosimilitud siempre que $\eta > 0$. En el caso en que $\eta \to \infty$ sólo se reestiman los parámetros correspondientes a la palabra errónea de mayor verosimilitud (ver la figura 1.4. Como se discute más abajo, este término es el responsable de la aparición del asombrado de hipótesis. Finalmente, el último de los factores en la expresión (1.22), $S_{\lambda_j}(\mathcal{P}(x_n^i,\lambda_j))$, es igual a la sensibilidad de la probabilidad de que el modelo

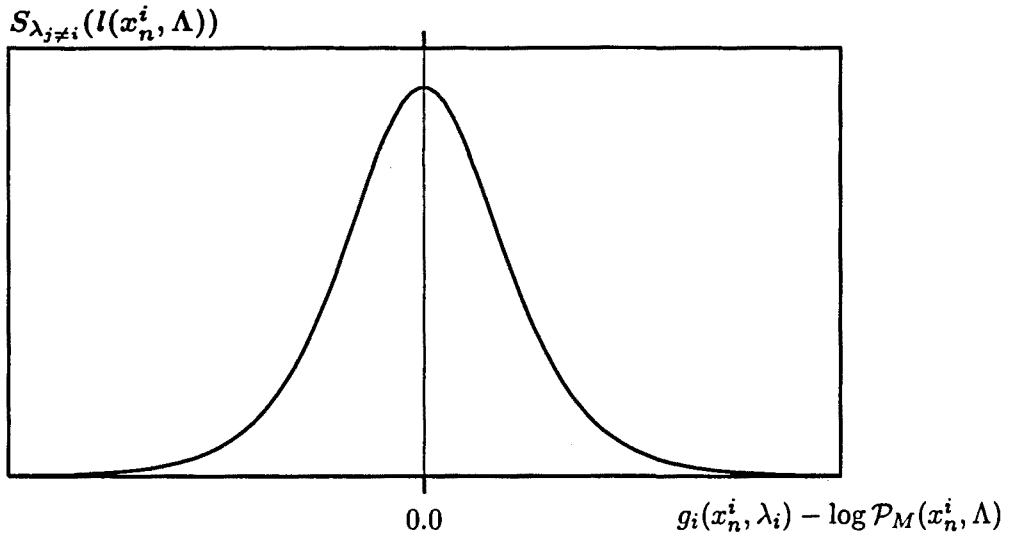

Figura 1.3: *Sensibilidad de la función de cómputo de error usada en MCE respecto a la diferencia entre el logaritomo de la verosimilitud de la palabra correcta,* $g_i(x_n^i, \lambda_i)$*, y la estimación del de la de valor máximo,* $\log \mathcal{P}_M(x_n^i, \Lambda)$*. El valor máximo de la sensibilidad se produce para las elocuciones en las que ambos valores son semejantes,* $g_i(x_n^i, \lambda_i) - \log \mathcal{P}_M(x_n^i, \Lambda) \approx 0$*, llegando a anularse cuando el error es muy poco probable,* $g_i(x_n^i, \lambda_i) - \log \mathcal{P}_M(x_n^i, \Lambda) \gg 0$*, o cuando es inevitable,* $g_i(x_n^i, \lambda_i) - \log \mathcal{P}_M(x_n^i, \Lambda) \ll 0$.

considerado genere la realización. Es en este último término donde se aborda directamente la eliminación de cada una de las hipótesis erróneas, modulándose su influencia en función de la utilidad de la eliminación del conjunto de errores para la palabra, así como de la relación entre la verosimilitud del error y la de la primera hipótesis.

Sensibilidad de la información mutua. De manera análoga al caso de la función de cómputo de error en entrenamiento de mínimo error de clasificación, la sensibilidad de la información mutua puede expresarse como el producto de tres factores de significado distinto:

$$\begin{aligned} S_{\lambda_j}(I_\Lambda(x_n^i, W))\big|_{j\neq i} &= \left\|\nabla_{\lambda_j} \log I_\Lambda((x_n^i, W)\right\|_{j\neq i} = \\ &= \left\| \frac{\nabla_{\lambda_j} I_\Lambda(x_n^i, W)}{I_\Lambda(x_n^i, \Lambda)} \right\|_{j\neq i} = \\ &= \frac{1}{2I_\Lambda(x_n^i, \Lambda)} \frac{\mathcal{P}(x_n^i/\lambda_j)\mathcal{P}(w_j)}{\sum_k \mathcal{P}(x_n^i/\lambda_k)\mathcal{P}(w_k)} S_{\lambda_j}(\mathcal{P}(x_n^i, \lambda_j))\Big|_{j\neq i} \end{aligned} \tag{1.23}$$

Como en el caso de (1.22), la sensibilidad de la información mutua depende directamente de la sensibilidad de la probabilidad de generación de la realización x_n^i por el modelo λ_j, modulada por dos términos: uno constante y común para todas las hipótesis consideradas, $1/2I_\Lambda(x_n^i, \Lambda)$, y otro que reparte el esfuerzo de reestimación entre ellas, $\mathcal{P}(x_n^i/\lambda_j)\mathcal{P}(w_j)/\sum_k \mathcal{P}(x_n^i/\lambda_k)\mathcal{P}(w_k)$. Este término de reparto es análogo al que había en entrenamiento de mínimo error de clasificación para el valor del parámetro $\eta = 1$. Como en MCE, el máximo se produce para la hipótesis de mayor verosimilitud, aunque en este caso

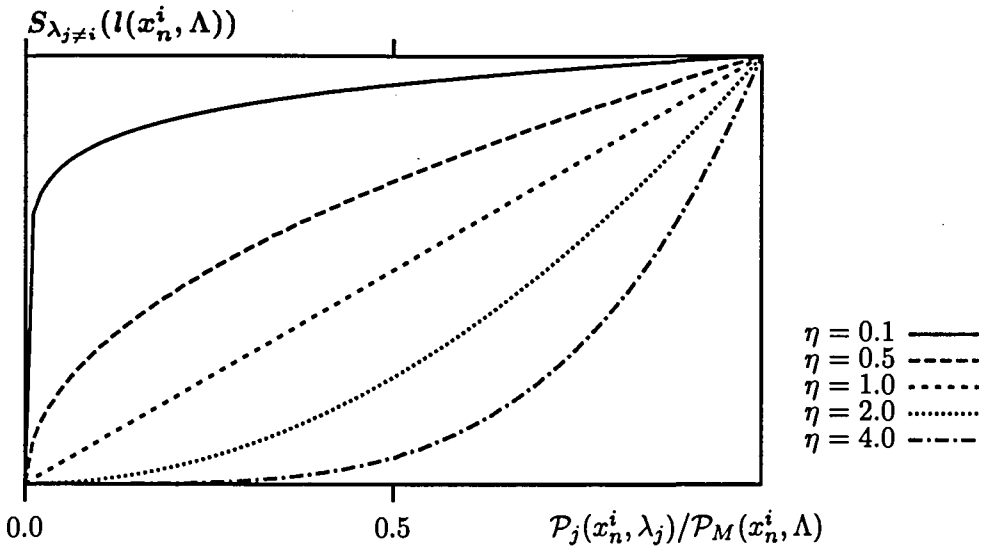

Figura 1.4: *Sensibilidad de la función de cómputo de error usada en MCE respecto a la relación entre la verosimilitud de la hipótesis errónea considerada, $\mathcal{P}_j(x_n^i, \lambda_j)$, y la estimación de la de valor máximo, $\mathcal{P}_M(x_n^i, \Lambda)$, para distintos valores del parámetro η. Con independencia de la relación con la verosimilitud de la palabra correcta, una hipótesis sólo tiene influencia en el gradiente de la función de cómputo de error si su verosimilitud es comparable a $\mathcal{P}_M(x_n^i, \Lambda)$. En caso contrario, $\mathcal{P}_j(x_n^i, \lambda_j) \ll \mathcal{P}_M(x_n^i, \Lambda)$, su influencia llega a anularse.*

el reparto es lineal en función de la probabilidad de cada hipótesis. Un fenómeno característico del entrenamiento de máxima información mutua: la sensibilidad del método crece sin control conforme la información mutua de la elocución de entrenamiento disminuye [98]. Es decir, el método de la máxima información mutua centra el esfuerzo de la reestimación en el aumento de la información mutua de las elocuciones más ambiguas. Este fenómeno es indeseable porque representa que el algoritmo puede depender en demasía de elocuciones poco representativas del conjunto de la población: aquéllas para las cuales la probabilidad de ser generadas por su propio modelo es muy inferior a la de serlo por algún otro, $\sum_k \mathcal{P}(x_n^i/\lambda_k) \gg \mathcal{P}(x_n^i/\lambda_i)$.

El asombrado de hipótesis. El término de ponderación de la sensibilidad de ambos métodos en función de la relación entre la probabilidad del error considerado y el resto de hipótesis implica que el proceso de optimización no trata por igual a las distintas hipótesis erróneas para una misma elocución de entrenamiento. En concreto, y en ambos casos, el esfuerzo se centra en la reducción de la probabilidad de comisión de la hipótesis más probable. Como resultado, cuanto mayor sea la probabilidad de ésta, menor será el esfuerzo dedicado a la erradicación del resto. Este efecto se produce con independencia de la relación entre la verosimilitud de la palabra correcta y la de las distintas hipótesis erróneas. Así, si hay una hipótesis con probabilidad muy superior al resto, sólo ésta participará activamente en el entrenamiento *asombrando* a todas las demás. En la figura 1.4 puede observarse la dependencia de la sensibilidad de la función de cómputo de error respecto a los parámetros de una cierta hipótesis en función de la diferencia entre los logaritmos de la verosimilitud de esa hipótesis y la estimación realizada de la verosimilitud de la hipótesis más probable.

El asombrado de las hipótesis de menor probabilidad por parte de las de mayor probabilidad es consecuencia directa de la naturaleza de las funciones optimizadas. En concreto, la función de cómputo de error en entrenamiento de mínimo error de clasificación pretende ser una aproximación a lo que su nombre indica: el número de elocuciones que se reconoce incorrectamente. Esta función sólo depende de la relación entre las verosimilitudes de la palabra correcta y de la hipótesis más verosímil. Siempre que esta última tenga mayor probabilidad de generar la elocución que la correcta, tendremos un error. El hecho de que haya otras posibles alternativas erróneas de mayor verosimilitud que la correcta no tiene ninguna implicación en el cómputo global: la contribución sigue siendo un sólo error. Este comportamiento no es sino un fiel reflejo del mecanismo de decisión adoptado en el reconocimiento: el resultado del reconocimiento es la palabra con mayor puntuación de asignación, $g_k(x_n^i/\lambda_k)$. En consecuencia, la eliminación del error más probable es mucho más beneficioso, desde el punto de vista del cómputo global de errores, si éste es el único error con mayor probabilidad que la palabra correcta, que si hay más de uno. De hecho, en tanto no consigamos eliminar todos y cada uno de los errores posibles, reducir su número no tiene ninguna repercusión en la función de coste global.

Ahora bien, para que el sistema de entrenamiento refleje fielmente los mecanismos del reconocimiento no sólo es necesario usar una función de coste adecuada, sino también que las elocuciones de entrenamiento y las hipótesis erróneas consideradas sean representativas de las que se pueden encontrar en condiciones reales de funcionamiento del sistema de reconocimiento. Es decir, el material de entrenamiento debe estar formado por frases pertenecientes a la tarea a reconocer, y la gramática de la tarea debe ser conocida para poderla utilizar en la determinación de las hipótesis erróneas. Si éste es el caso, el asombrado de hipótesis es conveniente, dado que es consecuencia de la fidelidad en el modelado de la regla de decisión del sistema de reconocimiento. Si no lo es, y el sistema de entrenamiento no puede reflejar fielmente las condiciones del reconocimiento —por ejemplo, porque no se conoce la tarea a reconocer o no se dispone de material de entrenamiento específico—, el asombrado de hipótesis puede no estar justificado. En el próximo capítulo se estudia el entrenamiento discriminativo de unidades subléxicas utilizando bases de datos independientes de la tarea a reconocer. Como se discutirá ahí, en este tipo de entrenamiento el esfuerzo dado a la eliminación de una hipótesis determinada debería depender en mayor medida de la importancia que el error tiene en el reconocimiento de la tarea, que no de la probabilidad del resto de hipótesis erróneas.

1.4.2.4 Comparativa de los sistemas de entrenamiento discriminativo propuestos en la literatura

Los trabajos publicados con los distintos esquemas de entrenamiento discriminativo prueban, en todos los casos, la oportunidad de este planteamiento de entrenamiento frente al convencional de máxima verosimilitud. Con cualquiera de los sistemas mencionados en los apartados enteriores —LVQ, entrenamiento correctivo, MCE y MMIE— es posible reducir, en algún caso espectacularmente, los errores en el reconocimiento. Existen, no obstante, diferencias entre ellos, tanto de concepto como de prestaciones. Así, el LVQ es una herramienta muy potente, y sólidamente fundamentada, para el diseño de cuantificadores vectoriales [48, 47, 88], pero de complicada extensión al entrenamiento de los modelos acústicos. Por otro lado, el entrenamiento correctivo, aunque sí permite abordar el entrenamiento de los modelos acústicos —de hecho está orientado a él—, es

una técnica heurística carente de argumentación teórica sólida [3]. Finalmente, tanto el entrenamiento de máxima información mutua, como el de mínimo error de clasificación, están sólidamente fundamentados —el MMIE en conceptos de teoría de la información, y el MCE en estadísticos—. No obstante, el primero, MMIE, ha reportado repetidamente peores resultados que el (teóricamente más débil) entrenamiento correctivo [2]. También son peores que los obtenidos con MCE [98]. Así pues —desde el punto de vista de la obtención de altas prestaciones en reconocimiento manteniendo el máximo rigor matemático—, la mejor alternativa para el entrenamiento discriminativo de modelos ocultos de Markov parece ser la proporcionada por el criterio de mínimo error de clasificación, La razón de esta superioridad se basa en que es el único que aborda —de manera directa— la minimización del número esperado de errores del sistema.

Capítulo 2

Entrenamiento de Mínima Confusibilidad de Unidades Subléxicas en Segmentos Acústicos de Longitud Limitada

Los métodos de entrenamiento discriminativo han demostrado ser una herramienta muy potente en el modelado acústico de palabras y otras unidades dependientes de la tarea a reconocer. No obstante, su aplicación a los sistemas de reconocimiento del habla continua basados en unidades subléxicas independientes de la tarea ha resultado ser mucho más complicada.

La principal limitación de los sistemas de reconocimiento dependientes de la tarea es la necesidad de disponer de bases de datos específicas para realizar el entrenamiento de los modelos acústicos. Aunque ciertas tareas —como el reconocimiento de dígitos o letras— mantienen su interés y justifican el uso de este tipo de base de datos, la utilidad de las técnicas de entrenamiento discriminativo puede ser puesta en entredicho en tanto no resulten también provechosas en el modelado acústico de unidades subléxicas independientes de la tarea. El objetivo de este capítulo es la extensión de los métodos de entrenamiento disciminativo a este tipo de sistema.

En primer lugar se analiza la problemática específica del entrenamiento discriminativo de unidades subléxicas para el reconocimiento del habla continua. Se considera una aproximación al problema en tres pasos: establecimiento del material de entrenamiento, definición de una función de coste sobre ese material y optimización de la función de coste mediante algún algoritmo adecuado. Se propone el uso de la función de confusibilidad estimada a partir de segmentos acústicos de longitud limitada los cuales, como se verá a lo largo del capítulo, permiten abordar la minimización de la confusibilidad dependiente de la tarea utilizando bases de datos independientes de la misma. El entrenamiento discriminativo independiente de la tarea es tratado a partir de la adaptación a una tarea suficientemente general como para abarcar cualquier posible tarea: el idioma. Finalmente, la utilidad de las propuestas realizadas se comprueba en un escenario especialmente complicado: la optimización de sistemas de reconocimiento basados en unidades subléxicas dependientes del contexto. Este tipo de sistemas es el que, hoy por hoy, proporciona la alternativa más potente en modelado acústico para el reconocimiento del habla continua.

Así pues, cualquier mejoría adicional no sólo es epecialmente difícil de lograr, sino que también implica alcanzar las máximas prestaciones posibles con la tecnología actual.

2.1 Entrenamiento Discriminativo y Reconocimiento Automático del Habla Continua

La principal ventaja de los sistemas de reconocimiento del habla continua basados en unidades subléxicas independientes de la tarea no es tanto la capacidad para reconocer cualquier tarea con el mismo conjunto de modelos, como el hecho que éstos son entrenados con una base de datos de propósito general y, por tanto, barata. Por desgracia es, precisamente, el uso de bases de entrenamiento independientes de la tarea la principal dificultad que aparece al intentar aplicar las técnicas de entrenamiento discriminativo en la reestimación de los modelos acústicos de unidad subléxica para el reconocimiento del habla continua. Esto es así debido a que —en sus diferentes versiones— estas técnicas se basan en aumentar la probabilidad de que el sistema reconozca la frase correcta y no sus posibles competidoras en la tarea. El procedimiento seguido consiste, habitualmente, de una fase en la que se determinan las verosimilitudes y alineados óptimos de la frase correcta y los N errores más probables para cada elocución de entrenamiento; seguida de la fase de reestimación en la cual se utilizan estos alineados y verosimilitudes para actualizar los parámetros del sistema de manera que la diferencia entre la verosimilitud de la frase correcta y la de los N errores más probables sea lo más grande posible.

Si el sistema es dependiente de la tarea, y se dispone de una base de datos específica, la obtención de hipótesis erróneas, alineados y verosimilitudes, y la consiguiente aplicación de entrenamiento discriminativo son inmediatas. Por contra, en el caso de utilizar una base de datos de propósito general, y aún conociendo el lenguaje de la tarea a reconocer, la minimización del número de errores cometido en el reconocimiento de una tarea concreta sólo puede ser acometido de manera indirecta. Es decir, no se pueden considerar sólo errores permitidos por la tarea, ya que, para ello, es necesario que el material de entrenamiento esté formado por frases específicas de la tarea, y ésta debe ser utilizada en la determinación de los posibles errores para evitar considerar hipótesis no permitidas. En lugar de ello, es necesario definir una medida de la capacidad del sistema en discriminar las distintas unidades subléxicas a partir de una base de datos de propósito general, de tal manera que la optimización de esa medida lleve a la minimización de la tasa de error del sistema durante el reconocimiento de la tarea. Ahora bien, aunque pueden postularse múltiples alternativas, no cualquier medida de calidad del sistema de reconocimiento del habla continua estimable utilizando una base de datos de propósito general, por plausible que ésta sea, tiene por qué reflejar adecuadamente el comportamiento del sistema en tareas reales de reconocimiento del habla continua. Por ejemplo, es discutible la utilidad de la minimización del número de errores cometidos en clasificación fonética —reconocimiento aislado de cada uno de los fonemas que forman la base de datos, conociendo y respetando la segmentación correcta en unidades—. Si bien la utilización de entrenamiento discriminativo permite reducir la tasa de error en clasificación de fonemas considerablemente [98, 81], esta mejora no sólo no se traslada al reconocimiento del habla continua, sino que las prestaciones incluso empeoran [81] (en [98] sólo se aportan resultados en la tarea de clasificación de fonemas).

En sus dos variantes más populares —el entrenamiento de mínimo error de clasificación

y el de máxima información mutua—, así como en la propuesta propia de esta tesis, el entrenamiento de mínima confusibilidad, el entrenamiento discriminativo de los modelos acústicos pueden descomponerse en tres fases diferentes:

1. Establecimiento del material de entrenamiento. En entrenamiento discriminativo el material a emplear consiste de errores cometidos en reconocimiento. Es preciso, por tanto, establecer una tarea concreta en la cual determinarlos.

2. Definición de la función de coste. A partir de la información proporcionada por los distintos errores considerados, se formula una función de coste que representa la calidad global del sistema.

3. Mediante algún algoritmo de optimización, habitualmente de búsqueda de gradiente, se reestiman los parámetros del sistema de manera que la función de coste se minimice.

En este capítulo se abordan las dos primeras de estas fases, tratándose el tema de la optimización de la función de coste en el próximo.

2.1.1 Material de entrenamiento: segmentos acústicos de longitud limitada

La problemática de la elección del material de entrenamiento para la aplicación de técnicas de entrenamiento discriminativo a sistemas de reconocimiento del habla continua es radicalmente distinta tanto al entrenamiento de máxima verosimilitud, como al entrenamiento discriminativo dependiente de la tarea.

En entrenamiento de máxima verosimilitud, el modelo de cada unidad es entrenado únicamente a partir de segmentos acústicos correspondientes a ella. Suponiendo que las distintas realizaciones de la unidad no dependen del contexto en que se encuentran —bien porque se desprecia su influencia, en el caso de unidades subléxicas independientes del contexto; bien porque esa influencia es modelada explícitamente, unidades dependientes del contexto—, los modelos obtenidos para cada unidad no dependerán de las características de la base de entrenamiento, sirviendo una de propósito general para el entrenamiento de los modelos de todas las unidades. Esta caraterística del entrenamiento de máxima verosimilitud es consecuencia directa de la forma polinómica de primer orden de coeficientes constantes de la función auxiliar de Baum, lo cual permite descomponer el gradiente en términos independientes para cada unidad. En entrenamiento discriminativo de sistemas de reconocimiento del habla continua, por contra, el entrenamiento de una unidad no puede independizarse de las unidades circundantes dado que los errores cometidos en el reconocimiento del habla continua no suelen ceñirse a los límites concretos de las unidades. Así, no sólo aparecen errores de sustitución entre unidades —como en el caso de la clasificación de fonemas—, sino que también aparecen errores de inserción y sustitución, y, en cualquier caso, no existe ninguna garantía de que la segmentación correcta en unidades sea respetada, aún en el caso que la secuencia correcta de unidades sí lo sea.

Por otro lado, en entrenamiento discriminativo dependiente de la tarea, la consideración de todos los errores posibles a cometer por el sistema puede abordarse a partir de frases de la tarea enteras. Dado que el material disponible es específico y el lenguaje de la tarea conocido, un reconocimiento de las N hipótesis más probables proporciona información

precisa de los errores cometidos en condiciones reales de funcionamiento, siendo cada uno de estos errores confusiones entre segmentos acústicos de principio y fin conocidos: los de la propia frase. Los sistemas de reconocimiento del habla continua, sin embargo, son entrenados a partir de bases de datos formadas por frases que no responden a ninguna tarea concreta. Por el contrario, el criterio de diseño de la base de datos suele ser garantizar que cada unidad aparezca un número suficiente de veces en cada contexto posible, lo cual, en la práctica, entra en contradicción con la imposición de ningún tipo de restricción a las frases que la forman.

En entrenamiento discriminativo de unidades subléxicas no existe una definición clara para el material de entrenamiento, y no siempre se cumple la restricción de únicamente utilizar bases de datos independientes de la tarea a reconocer. El objetivo de la elección debe ser el planteamiento de una tarea de reconocimiento que, utilizando una base de datos de propósito general, represente suficientemente las prestaciones del sistema en el reconocimiento de cualquier tarea real. Entre las alternativas utilizadas, pueden destacarse las siguientes:

1. Reconocimiento de frases de habla continua dependientes de una tarea[1].
2. Clasificación de tramas aisladas de señal.
3. Clasificación de unidades subléxicas aisladas.
4. Decodificación acústico fonética en frases de habla continua independientes de la tarea.
5. Decodificación acústico fonética en segmentos acústicos de longitud limitada.

2.1.1.1 Frases de habla continua dependientes de una tarea

El primer esquema de entrenamiento discriminativo propuesto para habla continua consiste en utilizar una base de datos dependiente de una tarea de habla continua, y minimizar alguna función de pérdida definida a partir de los errores cometidos en el reconocimiento de esa misma tarea. En cierto modo el atractivo de la utilización de bases de datos de entrenamiento independientes de la tarea se pierde al utilizar una base de datos específica. No obstante, esto sólo es cierto si el sistema va a ser utilizado en el reconocimiento de la tarea utilizada en el entrenamiento, pudiéndose plantear la conveniencia de emplear los modelos óptimos en una tarea concreta para el reconocimiento de cualquier otra. Así, son posibles tres modos distintos de aplicar entrenamiento discriminativo utilizando frases de habla continua dependientes de la tarea:

1. Entrenamiento dependiente, reconocimiento dependiente.
2. Entrenamiento dependiente, reconocimiento independiente.
3. Entrenamiento independiente, reconocimiento independiente.

Entrenamiento dependiente, reconocimiento dependiente. En este modo de operación, la base de datos de entrenamiento debe estar formada por frases pertenecientes a la misma tarea a reconocer. Se pierde, por tanto, la independencia de la tarea de la base de entrenamiento —y, en cierto modo, su propia calificación de sistema de habla continua— ya

[1]Como se explica más adelante, esta alternativa también puede contemplarse como entrenamiento independiente de la tarea a reconocer si ésta no es la misma que la empleada en el entrenamiento.

que las unidades subléxicas empleadas pasan a ser unidades dependientes de la tarea y, en principio, no útiles para el reconocimiento de una distinta. Esta estrategia ha sido aplicada repetidamente en el reconocimiento de una tarea típica: la *Resource Management* de DARPA, de 997 palabras [90], usando el bigrama de palabras de perplejidad 60. Un ejemplo lo proporciona [59]. En este trabajo, el bigrama de palabras es usado para determinar los $N = 6$ errores más probables de cada frase. A partir de estos errores, los modelos de Markov se reestiman aplicando entrenamiento correctivo. El resultado obtenido es notable: partiendo de un 6,3% de frases incorrectas se pasa a un 4,9% —una reducción de más del 20%—.

Entrenamiento dependiente, reconocimiento independiente. La necesidad de bases de datos específicas para el entrenamiento discriminativo de modelos de unidades subléxicas en tareas de reconocimiento del habla continua no presenta mayor inconveniente si los modelos obtenidos en la optimización de una cierta tarea son igualmente útiles en el reconocimiento de cualquier otra. En este caso, se puede plantear la obtención de modelos acústicos independientes de la tarea —en el sentido de que se usarán en el reconocimiento de cualquier tarea de habla continua desconocida a priori— pero entrenados a partir de una base de datos dependiente de una conocer con antelación los límites de cada una de las unidades. Estos límites pueden determinarse de manera manual o mediante alineado forzado del contenido acústico de la frase con su contenido fonético empleando el algoritmo de Viterbi (ver el apartado 1.2.2.1).

Desde el punto de vista de su utilización en un sistema de entrenamiento discriminativo, la minimización del número de errores en clasificación de unidades subléxicas, al igual que ocurre con la clasificación de tramas de señal, presenta la ventaja de facilitar la obtención de los N errores más probables. Esta característica simplifica enormemente la aplicación de entrenamiento discriminativo alcanzándose mejoras considerables en la tasa de clasificación —13% de reducción de la tasa de error en clasificación de fonemas en alemán [98]; 24% de reducción en inglés usando TIMIT [81]—. No obstante, esta mejoría tan sustancial no se traduce en mejora alguna cuando los modelos obtenidos de este modo son aplicados a tareas de reconocimiento del habla continua. Por el contrario, en [81] los mismos modelos que reducen en un 24% la tasa de error en clasificación de fonemas, con respecto a los resultados obtenidos utilizando entrenamiento de máxima verosimilitud, empeoran en casi un 40% —del 10,9% al 15%— la tasa de error de frases cuando son utilizados en el reconocimiento de las cadenas de dígitos de TIDIGITS (ver la tabla 2.1).

Este comportamineto puede achacarse al hecho que todos los errores que aparecen en clasificación de unidades subléxicas pertenecen a un tipo concreto de error: los que se producen cuando una unidad subléxica es sustituida por otra respetando estrictamente los límites temporales de la correcta. En habla continua no sólo hay errores de confusión entre unidades, sino que también aparecen errores de inserción y borrado de unidades. Además, incluso cuando el error producido es una confusión entre unidades subléxicas, los límites en los cuales se reconoce la unidad incorrecta no tienen porque coincidir con los de la correcta. Así pues, aunque el tipo concreto de error minimizado en el entrenamiento se reduce considerablemente, ese beneficio sólo se trasladará al reconocimiento del habla continua si no se produce un incremento de los otros tipos de error.

Aunque la minimización de la tasa de error en clasificación de unidades subléxicas no resulta de interés en la optimización de los modelos de Markov de sistemas de reconocimiento del habla continua, sí puede resultarlo si el parámetro optimizado está

Experimento	Tarea	Error	Sust	Inse	Borr	Acierto	Clas
Base	DAF	39,9	21,2	12,3	6,4	72,4	61,1
fono	DAF	38,7	24,6	2,4	11,7	63,7	70,5
Experimento	**Tarea**	**Error**	**Sust**	**Inse**	**Borr**	**Acierto**	**Corr**
Base	digitos	3,7	1,3	1,5	0,9	97,8	89,1
fono	digitos	5,4	2,9	0,7	1,8	95,3	85,0

Tabla 2.1: *Resultados obtenidos en DAF, clasificación de fonemas y el reconocimiento de TIDIGITS, empleando modelos de fonema entrenados con Baum-Welch (**Base**), y según MCE en clasificación de fonemas aislados (**fono**). La columna **Clas** muestra el resultado en clasificación de fonemas aislados. Se observa que éste mejora considerablemente aplicando MCE. Sin embargo, el reconocimiento de las cadenas de dígitos empeora sensiblemente.*

suficientemente alejado de la regla decisión empleada en el reconocimiento. Es decir, si el parámetro optimizado es suficientemente general a ambos sistemas de reconocimiento —el de clasificación de unidades y el de reconocimiento del habla continua— como para que el incremento en la capacidad de discriminación en un caso resulte beneficioso también en el otro. Por ejemplo, ambos sistemas comparten tanto la etapa de extracción de características como la de cuantificación vectorial. Si se consigue disminuir el número de errores en clasificación fonética aumentando la capacidad de discriminación del cuantificador, es posible que ese incremento de discriminación también resulte beneficioso en cualquier otra tarea. Por ejemplo, en otro trabajo de Reichl [97], el objeto del entrenamiento discriminativo es la optimización de la fase de extracción de características en clasificación de fonemas. En este caso, la reducción de la tasa de error en clasificación sí resulta en una reducción análoga de la tasa de error en deocodificación acústico fonética.

2.1.2 Decodificación acústico fonética en frases de habla continua independientes de la tarea

La decodificación acústico fonética consiste en el reconocimiento libre de unidades subléxicas. Es decir, en la determinación de la sucesión de unidades que mejor representa la frase a reconocer. En caso de usar una gramática, ésta es fonotáctica —permite con igual probabilidad aquellas sucesiones de unidades válidas en el idioma, e impide cualquier otra sucesión—, o estocástica —bigramas, trigramas y, en general, n-gramas, que asignan una probabilidad a cada posible sucesión de unidades—.

La optimización de la tasa de error en decodificación acústico fonética resulta un tanto más complicada que en los casos de clasificación de tramas y unidades subléxicas ya que, ahora, la obtención de los N errores más probables requiere de la realización de un reconocimiento de múltiples hipótesis en habla continua [63]. Una vez obtenidas las distintas hipótesis, el sistema se reestima de igual modo que si se tratara de un sistema de palabras aisladas, donde cada hipótesis representa una palabra del vocabulario. La principal ventaja de optimizar los modelos de las unidades subléxicas en la tarea de decodificación acústico fonética radica en el hecho que esta tarea sí tiene en cuenta todos los errores que pueden aparecer en habla continua: de sustitución, inserción, borrado y segmentación; no apareciendo errores de imposible comisión —al contrario de lo que ocurre

HIP	sil	f	ao	r	ey	t	s	ih	k	s	eh	v	en	sil
1	sil	f	ao	r	ey	t	s	ih	g	s	ow		en	sil
2	sil	f	ao	r	ey	t	s	ih	g	s	ow		n	sil
3	sil	f	ao	r	ey	t	s	ih	g	s	ah		en	sil
4	sil	f	ao	r	ey		ch	ih	g	s	ow		en	sil
5	sil	f	ao	r	ey	t	s	ih	g	s	ah		n	sil
6	sil	f	ao	r	ey		ch	ih	g	s	ow		n	sil
7	sil	f	ao	r	ey		ch	ih	g	s	ah		en	sil
8	sil	f	ao	r	ey	t	s	ih	g	s	ow		ng	sil
9	sil	f	ao	r	ey	t	s	eh	g	s	ow		en	sil
10	sil	f	ao	r	ey		ch	ih	g	s	ah		n	sil
11	sil	f	ao	r	ey		ch	ih	g	s	ow		ng	sil
12	sil	f	ao	r	ey	t	s	ih	g	s	ah		ng	sil

Tabla 2.2: *Resultado del reconocimiento de las 12 hipótesis más problables en DAF de la frase de TIDIGITS* **`test/man/sw/4867a`**. *También se muestra, en negrita, la secuencia correcta. Puede observarse que casi todas las hipótesis son idénticas en la mayor parte de la elocución. Especialmente en los segmentos en los que el reconocimiento en primera hipótesis es correcto. También se observa la existencia de segmentos erróneos poco relacionados entre sí: el que afecta al fonema /t/ de eight, y el que afecta a la parte central de seven.*

en clasificación de tramas—. Así, la eliminación total de los errores en decodificación acústico fonética es garantía del reconocimiento correcto —salvando posibles problemas de homofonía— de cualquier tarea de habla continua.

La utilización de la tasa de decodificación acústico fonética, no obstante, presenta dos grandes inconvenientes en su utilización en entrenamiento discriminativo (ver la tabla 2.2):

1. Mal aprovechamiento del material de entrenamiento: por muchas hipótesis que generemos, éstas suelen estar formadas por variaciones y combinaciones de los errores cometidos por las de mayor probabilidad. Como consecuencia, en buena parte de la frase —aquellos segmentos en que se ha reconocido en primera hipótesis la secuencia correcta de fonemas— las distintas hipótesis no incurrirán en error, dando lugar a una mala protección frente a *casi-errores*; y donde sí aparecen errores, las distintas hipótesis sólo se diferencian en unas pocas unidades.

2. Errores que efectan a partes distintas de la frase son agrupados en un único error en la función de coste minimizada. No se minimiza el número de errores cometido a nivel fonético, sino a nivel de frase completa.

De los dos inconvenientes planteados a la utilización de la tasa de decodificación acústico fonética, el primero, el mal aprovechamiento del material de entrenamiento, es solucionable aumentando el tamaño de la base de datos. Dado que el tamaño de las bases de datos es cada vez mayor, puede esperarse que el mal aprovechamiento del material de entrenamiento deje de ser un problema algún día. Por el contrario, el hecho que la función de error empleada agrupe en un único error a nivel de frase todos los errores cometidos

a nivel fonético en el reconocimiento de la frase, aunque afecten a segmentos claramente diferenciados, tiene consecuencias más difícilmente evitables.

En primer lugar, la utilización de la tasa de decodificación de frases provoca la aparición del efecto de *arrastre de errores*, esto es: la dependencia cruzada entre errores cuya única relación es la de aparecer en la misma frase. Por ejemplo, si en la decodificación acústico fonética de una frase sólo se cometen dos errores, uno al principio de la frase y el otro al final, siendo correctamente reconocido un número elevado de fonemas entre ambos, puede suponerse que los dos errores no guardan ninguna relación entre sí, es decir: la probabilidad de comisión de uno de ellos no depende de la probabilidad con la que se comete el otro. Sin embargo, la consideración de la tasa de error de frase provoca el agrupamiento de ambos, haciendo que la sensibilidad de la función optimizada a variaciones en los paramétros del sistema que intervienen en uno de los errores dependa de la probabilidad del otro (ver el apartado 1.4.2.3).

En segundo lugar, y consecuencia directa del agrupamiento de errores en la función de coste, la sensibilidad de las funciones de coste utilizadas depende en gran medida del número total de errores cometidos en la frase. Así, dado que la medida de discriminación usada es la diferencia entre los logaritmos de la probabilidad de la frase reconocida y la correcta, si ambas coinciden en algún segmento, la diferencia de logaritmos en él será cero. Por tanto, si el número de errores de decodificación cometidos es pequeño, la diferencia de los logaritmos de la probabilidad de la frase correcta y las hipótesis erróneas tenderá a ser pequeña. Por el contrario, si el número de errores es elevado —por ejemplo, porque la longitud de la frase también es mayor— la diferencia de los logaritmos será elevada. La consecuencia de esto es distinta para el método de mínimo error de clasificación y el de máxima información mutua. En el primero, la sensibilidad de la función de coste es máxima para aquellas frases tales que la probabilidad de la hipótesis de mayor valor es próxima a la probabilidad de la frase correcta. Por tanto, la utilización del método de mínimo error de clasificación en decodificación acústico fonética de frases completas centrará sus esfuerzos en eliminar los errores de las frases con menor número de errores en primera hipótesis —por ejemplo, las más cortas—. En el caso de la máxima información mutua, la sensibilidad crece conforme aumenta la ambigüedad en la codificación de la frase, es decir: conforme mayor es la diferencia entre la probabilidad de las hipótesis erróneas y la de la frase correcta. Por tanto será máxima para las frases con mayor número de errores —las frases más largas, entre otras—.

Los sistemas objeto de reestimación mediante técnicas de entrenamiento discriminativo son sistemas que, inicialmente, suelen presentar tasas de error de frase en decodificación acústico fonética muy elevadas. En la mayor parte de los experimentos ninguna frase es decodificada completamente bien. Esto se debe al pobre modelado acústico proporcionado por las unidades subléxicas —en torno al 40% de tasa de error de fonemas—. Cuando el número de fonemas de una frase es elevado —en bases de datos de propósito general suele ser de unos 40 fonemas por frase—, la probabilidad de reconocer todos ellos correctamente es, prácticamente, despreciable. Por ejemplo, usando RAMSES y modelos acústicos de fonemas entrenados con el algoritmo de Baum-Welch, en ningún experimento se ha conseguido reconocer, ni tan siquiera, una frase enteramente bien. Utilizando modelos entrenados discriminativamente, se ha llegado a alcanzar un 2% de frases reconocidas correctamente sobre el propio entrenamiento —algo menos de la mitad en experimentos independientes del locutor—, pero este resultado puede considerarse anecdótico y, además, todas las frases reconocidas correctamente pertenecen al grupo de las más cortas —menos

de treinta fonemas por frase—.

No se conoce de la existencia de ningún trabajo donde se presenten los resultados del reconocimiento de frases completas en decodificación acústico fonética, pero ya se ha comentado que la tasa de reconocimiento en esta tarea suele rondar el cero más rotundo. En su lugar suele utilizarse la tasa de error de fonemas, definida como la suma de las tasas de sustitución, borrado e inserción. Los resultados publicados utilizando esta medida de error demuestran que, a pesar de los inconvenientes mencionados, el entrenamiento discriminativo permite mejorar de manera notoria la decodificación acústico fonética. En dos de estos trabajos, además, se proporcionan resultados utilizando los modelos obtenidos de este modo en tareas concretas de reconocimiento del habla continua [53, 81], con resultados igualmente satisfactorios. En [81] se alcanza una reducción del 22% en la tasa de decodificación acústico fonética independiente del locutor de las frases de TIMIT —se pasa del 39,9% utilizando máxima verosimilitud, al 31.2% utilizando mínimo error de clasificación—. Los mismos modelos reducen en un 10% la tasa de error en el reconocimiento de las cadenas de dígitos de TIDIGITS —del 10,9% de cadenas erróneas al 9,0%—. En [53], una reducción del 33% en la tasa de decodificación acústico fonética[2] redunda, así mismo, en reducciones en torno al 10% de la tasa de error en dos tareas de reconocimiento del habla continua distintas: el reconocimiento de cadenas de dígitos y letras, y el de nombres de ciudades del estado de Nueva Jersey.

2.1.3 Decodificación acústico fonética en segmentos acústicos de longitud limitada

De las distintas alternativas de material de entrenamiento comentadas hasta el momento —clasificación de tramas o de fonemas, reconocimiento dependiente de la tarea, y decodificación acústico fonética— sólo la utilización de los errores cometidos en decodificación acústico fonética de frases completas cumple simultáneamente tres condiciones necesarias para que su minimización sea de interés en la reestimación de los modelos de Markov para tareas de reconocimiento del habla continua, objetivo de esta tesis:

1. Posibilidad de entrenamiento de los modelos de Markov: la clasificación tanto de tramas de señal como de unidades subléxicas pueden ser de utilidad en la optimización de la extracción de características o cuantificación, pero presenta dificultades a la hora de reestimar los propios modelos.

2. Utilización de bases de datos independientes de la tarea: la optimización de tareas concreta de habla continua sólo demuestran su utilidad en el reconocimiento de la misma tarea que se utiliza en el entrenamiento.

3. Buena representación de los errores reales en habla continua: cualquier error posible en reconocimiento del habla continua es posible en decodificación acústico fonética, y se puede conseguir —utilizando gramáticas fonotácticas o estocásticas— limitar los errores considerados en esta última a errores posibles en el idioma.

Ya se han comentado, no obstante, las dos grandes limitaciones que presenta la minimización de la tasa de error de frase en decodificación acústico fonética de cara a optimizar los modelos de Markov: el pobre aprovechamiento del material de entrenamiento

[2]En este caso, la tasa de error en decodificación acústico fonética es estimada sobre el propio conjunto de frases de entrenamiento y no sobre un conjunto independiente del locutor y del texto.

disponible —de cada frase sólo se extrae información acerca de unos pocos errores— y la dependencia simultánea de la función utilizada de errores que afectan a partes distintas de una misma frase —de hecho se minimiza el número de frases mal decodificadas, no el número de fonemas erróneos—. Ambos problemas se acentúan con la longitud de las frases de entrenamiento. Por tanto, sus efectos pueden evitarse, por ejemplo, utilizando una base de entrenamiento formada por frases de corta longitud. Así, en [53] la base empleada consiste de frases de dos a cuatro palabras.

Los mismos sistemas que proporcionan resultados tan pobres en decodificación acústico fonética permiten alcanzar tasas de reconocimiento mucho más elevadas al introducirse la gramática de la tarea a reconocer. Por ejemplo, la tasa de error de fonemas en decodificación acústico fonética de TIDIGITS obtenida utilizando el sistema empleado a lo largo de esta tesis se sitúa en el entorno del 40%. Si las mismas frases son reconocidas utilizando la gramática de las cadenas de dígito, la tasa de error de fonemas baja hasta menos del 2%. En esta situación, la mayor parte de la frase es reconocida correctamente. Además, en los trozos de la frase en los que se comete error, éste sólo afecta a unas pocas unidades subléxicas. Ambas situaciones sugieren el uso de una aproximación segmental en la caracterización de los errores cometidos: la utilización de segmentos formados por unas pocas unidades subléxicas —en adelante, segmentos acústicos de longitud limitada (**SALL**)—.

2.1.3.1 Aproximación segmental basada en segmentos acústicos de longitud limitada

La aproximación segmental basada en segmentos acústicos de longitud limitada surge al considerar el resultado del reconocimiento de una frase de habla continua como la concatenación de segmentos formados por unas pocas unidades subléxicas tales que, o bien el resultado del reconocimiento en el segmento es correcto, o bien se ha cometido algún error —incluyendo posibles errores en la segmentación en unidades—. Para tomar en consideración posibles errores en la segmentación es necesario que los límites de los segmentos considerados correctos coincidan con los límites reales de las unidades correspondientes. En general esta condición es díficil de garantizar. Ahora bien, si se considera que la segmentación correcta es la correspondiente al alineado forzado de máxima verosimilitud entre la frase y los modelos acústicos [13], podemos suponer que las unidades reconocidas correctamente tenderán a hacerlo dentro de los límites correctos, los cuales pueden ser determinados utilizando el algoritmo de Viterbi visto en el apartado1.2.2.1. En concreto, podemos suponer que toda transición entre dos unidades correctamente reconocidas se produce por el mismo punto que el alineado forzado con el contenido correcto, independientemente de los errores cometidos en otras partes de la frase. Los límites de los segmentos a ser considerados como correctos comienzan en una transición entre unidades correctas —incluyendo la segunda de estas unidades, pero no la primera— y acaban en otra —incluyendo sólo la primera—, abarcando el mayor número posible de unidades correctas en su interior. Los límites de los segmentos incorrectos son los resultantes de dividir por las transiciones entre unidades correctas los segmentos no considerados como correctos. Es decir, no podemos tener dos segmentos correctos seguidos, pues darían lugar a un sólo segmento correcto más grande; pero sí podemos tener dos o más segmentos incorrectos consecutivos, si la transición entre ellos se produce entre unidades correctas. Por otro lado, los segmentos incorrectos pueden presentar unidades

HIP	sil	f	ao	r	ey	t	s	ih	k	s	eh	v	ax	n	sil
1	sil	f	ao	r	ey	t	s	ih	k	*s*	*ow*	*n*	*ay*	*n*	sil
2	sil	f	ao	r	ey	t	s	ih	k	s	eh	v	ax	n	sil
3	sil	f	ao	r	ey	t	s	ih	k	*s*	*w*	*ah*	*n*		sil
4	sil	f	ao	r	ey	t	s	ih	k	*s*	*ow*	*w*	*ah*	*n*	sil
5	sil	f	ao	r	ey	t	s	ih	k	*s*	*ow*	*ow*	*n*	*ay* *n*	sil
6	sil	f	ao	r	ey	t	s	ih	k	*s*		*ow*			*sil*
7	sil	f	ao	r	ey	t	s	ih	k	*s*	*w*	*ah*	*n*	*ay* *n*	sil
8	sil	f	ao	*r ow*	*ey*	t	s	ih	k	*s*	*ow*	*n*	*ay*	*n*	sil
9	sil	f	ao	*r ow*	*ey*	t	s	ih	k	s	eh	v	ax	n	sil
10	sil	f	ao	r	ey	t	s	ih	k	*s*	*ow*		*ow*		*sil*
11	sil	f	ao	*r ow*	*ey*	t	s	ih	k	*s*	*w*	*ah*	*n*		sil
12	sil	f	ao	*r ow*	*ey*	t	s	ih	k	*s*	*ow*	*w*	*ah*		*sil*

Tabla 2.3: *Segmentación en SALL correctos e incorrectos, en cursiva, del resultado del reconocimiento de las 12 hipótesis más problables de la frase de TIDIGITS* **`test/man/sw/4867a`** *utilizando la gramática de las cadenas de dígito. También se muestra, en negrita, la secuencia correcta.*

reconocidas correctamente en su interior, pero no cadenas de dos o más. La tabla 2.3 muestra la segmentación a que da lugar esta aproximación para las doce hipótesis más probables en el reconocimiento de una frase de TIDIGITS.

El planteamiento de esta aproximación es dividir cada una de las frases de entrenamiento en segmentos de unas pocas unidades subléxicas de tal manera que los distintos errores considerados en una estimación de las N hipótesis más probables proporcione información acerca de los errores en los que se podría ver involucrado ese mismo segmento en el caso de aparecer en una frase de la tarea a reconocer. Con las definiciones dadas para los segmentos correctos e incorrectos, los límites de éstos en reconocimiento del habla continua coinciden con los obtenidos mediante alineado forzado de la frase con su contenido, siendo esta operación independiente de la tarea. De este modo podemos segmentar las frases de entrenamiento a partir de su contenido acústico, obteniendo segmentos semejantes a los que hubiéramos obtenido en el caso de realizar esta operación sobre una base de datos específica. Para ello se fija el número de unidades subléxicas de que se compone cada segmento y se generan todos los segmentos posibles a partir de cada una de las frases de entrenamiento. Dado que el conjunto de errores posibles en un segmento de longitud L incluye todos los errores posibles en un subsegmento del mismo —téngase en cuenta que, aunque aumentemos la amplitud de los segmentos incorrectos incorporando a los mismos las unidades correctas consecutivas por sus extremos, el error considerado en el segmento no varía—, la minimización del número de errores en segmentos de longitud L también conlleva la minimización de errores en segmentos de longitud inferior a L. Por tanto, la longitud de los segmentos debe ser suficientemente grande como para permitir la aparición de todos los errores posibles en reconocimiento del habla continua.

La utilización de segmentos de longitud limitada en entrenamiento discriminativo lleva a una situación análoga a la que se da en entrenamiento de máxima verosimilitud de unidades subléxicas utilizando bases de datos independientes de la tarea. En este tipo

de entrenamiento cada modelo es entrenado únicamente con tramas de señal correspondientes a la unidad subléxica a la cual pertenece. Considerando que esta asignación se produce con independencia del contexto en que se halla la unidad, los modelos obtenidos entrenando con material correspondiente a una tarea determinada son igualmente válidos en cualquier otra, y es posible orientar el diseño de la base de datos de entrenamiento a la correcta representación de las distintas unidades subléxicas con total independencia de la tarea en la cual van a ser empleadas. En el caso de entrenamiento discriminativo, el objetivo es obtener una buena representación de los errores que se pueden cometer, no de las propias realizaciones de las distintas unidades. La aproximación segmental basada en cadenas de unidades subléxicas se apoya en considerar que los errores cometidos en un segmento determinado no dependen de los cometidos en los segmentos vecinos. Si se cumple esta independencia, entonces la minimización del número de errores cometido en el reconocimiento de ese segmento será beneficiosa en todas aquellas situaciones reales de reconocimiento del habla continua en las que pueda aparecer ese mismo error.

Cabe señalar que recientemente, en el ICSLP'98, O'Neill *et al.* propusieron un planteamiento similar (las *multi-phone strings*) como unidad subléxica para el reconocimiento del habla continua [86]. En concreto, se propone el uso de cadenas de dos fonemas. Aunque los objetivos de ese trabajo son muy distintos a los de esta tesis, los razonamientos que llevan a los autores a su propuesta son muy parecidos a los aplicados para defender el uso de segmentos acústicos de longitud limitada en entrenamiento discriminativo, así como a los que llevan a la utilización de semifonemas dependientes del contexto.

2.1.3.2 Comparación del entrenamiento de mínimo error en decodificación acústico fonética usando frases completas y segmentos acústicos de longitud limitada

La utilización de segmentos acústicos de longitud limitada, frente a la de frases completas, tiene varias ventajas de cara a su utilización en entrenamiento discriminativo. Estas ventajas se deben, en general, a la menor longitud de los segmentos sobre los cuales se va a realizar la determinación de las N hipótesis más probables.

1. El número de errores distintos considerado es muy superior dividiendo la frase en segmentos de menor longitud ya que se fuerza el reconocimiento de N hipótesis para cada uno de ellos independientemente.

2. Proporciona mayor protección frente a *casi errores* ya que se fuerza el reconocimiento de posibles confusiones incluso para los segmentos reconocidos correctamente en primera hipótesis.

3. Evita el efecto *arrastre*, esto es: la consideración simultánea de dos o más errores que afectan a partes distintas de la elocución sin guardar relación entre sí.

4. Al optimizar el reconocimiento en segmentos de longitud menor a la de la propia frase, el criterio se aproxima a la minimización de la tasa de error de fonemas.

5. Si la longitud de todos los segmentos es la misma, la dependencia de la sensibilidad de la función de coste respecto la longitud de la frase es neutralizada.

Como única desventaja de los segmentos de longitud limitada a L unidades frente al uso de frases completas en la aplicación de entrenamiento discriminativo hay que señalar su

Experimento	Tarea	Error	Sust	Inse	Borr	Acierto
Base	DAF	39,9	21,2	12,3	6,4	72,4
frase	DAF	31,2	20,8	3,0	7,4	71,8
segm	DAF	30,8	18,6	7,3	4,9	76,5

Experimento	Tarea	Error	Sust	Inse	Borr	Acierto	Corr
Base	digitos	3,7	1,3	1,5	0,9	97,8	89,1
frase	digitos	3,2	1,4	0,6	1,2	97,4	91,0
segm	digitos	2,9	1,2	1,1	0,6	98,2	91,8

Tabla 2.4: *Resultados obtenidos en DAF y el reconocimiento de TIDIGITS empleando modelos de fonema entrenados con Baum-Welch (**Base**), según MCE en DAF de frases completas (**frase**), y según MCE en DAF de segmentos de cinco fonemas (**segm**). Aunque ambos esquemas discriminativos consiguen mejorar el resultado del reconocimiento en ambas tareas, en el caso de usar frases completas, la mejoría se debe, fundamentalmente, a la reducción del número de errores de inserción. (En estos experimentos no se intenta equilibrar la tasa de inserción y borrado.)*

incapacidad para representar adecuadamente errores que afecten a cadenas de más de L unidades. No obstante, incluso en el caso de realizar decodificación acústico fonética, la mayor parte de los segmentos erróneos —según la definición de los mismos dada más arriba— presentan una longitud inferior a seis fonemas. Por lo tanto, una segmentación de la base de entrenamiento en segmentos de cinco fonemas basta para considerar casi todos los posibles errores en habla continua. Además es previsible que la minimización del número de errores en segmentos de longitud L, aunque no toma en consideración de manera explícita errores de más de L unidades, sí resultará beneficioso en su erradicación, ya que, en primera aproximación, estos últimos pueden descomponerse en la concatenación de segmentos erróneos de menor longitud, probablemente solapados.

La tabla 2.4 muestra los resultados en decodificación acústico fonética (**DAF**) y reconocimiento de las cadenas de dígitos (**digitos**) obtenidos utilizando entrenamiento de mínimo error de clasificación en frases completas (**frase**) y segmentos de cinco fonemas (**segm**). Se observa que el resultado obtenido en DAF es muy semejante en ambos casos —en torno al 22% de reducción en la tasa de error de fonemas respecto a los modelos de máxima verosimilitud—, aunque ligeramente mejor para el caso de los segmentos de cinco fonemas. Sin embargo, el pequeño beneficio aportado por la utilización de segmentos en DAF se traduce en una mejoría mucho más importante cuando los modelos son utilizados en el reconocimiento de las cadenas de dígitos —la reducción en la tasa de frases incorrectas es del 17%, utilizando frases completas; y del 25%, utilizando segmentos—. La explicación a este comportamiento desemejante puede achacarse al tratamiento dado a los errores de inserción y borrado en uno y otro caso. Así, mientras la reducción en la tasa de error en decodificación acústico fonética obtenida utilizando frases completas es debida, fundamentalmente, a la reducción del número de fonemas insertados —la tasa de inserción se reduce en nueve puntos, mientras que la tasa de acierto empeora ligeramente—, en el caso de utilizar segmentos de cinco fonemas, la tasa de inserción también se reduce —aunque sólo en cinco puntos—, pero la tasa de acierto aumenta considerablemente —más

de tres puntos—[3].

2.2 Función de Coste: Criterio de Mínima Confusibilidad

Tanto en entrenamiento de mínimo error de clasificación como de máxima información mutua, el procedimiento seguido consiste en la formulación de una función continua y diferenciable que representa la calidad del sistema. La reestimación de los parámetros del sistema se realiza de manera que se optimice el valor de esta función calculada sobre el material de entrenamiento —accediendo a un mínimo local de la función cómputo de error en MCE, o un máximo de la información mutua en MMIE—. En la presentación de ambos métodos se dio especial importancia a una característica de ambas funciones de calidad: la existencia del *asombrado de hipótesis*, que, si bien está plenamente justificada en entrenamiento discriminativo dependiente de la tarea, tiene efectos indeseables a la hora de aplicar entrenamiento discriminativo a unidades subléxicas independientes de la tarea para su utilización en sistemas de reconocimiento del habla continua.

2.2.1 Asombrado de hipótesis en entrenamiento discriminativo independiente de la tarea

El asombrado de hipótesis consiste en el hecho que el esfuerzo con que se aborda la eliminación de cada una de las hipótesis erróneas depende de la verosimilitud del resto. Este efecto es común al entrenamiento de máxima información mutua y al de mínimo error de clasificación, siendo mucho más marcado en este último, así que la presente discusión se centrará en él.

En entrenamiento de mínimo error de clasificación, la función optimizada es especialmente sensible a la verosimilitud de la hipótesis errónea más probable (ver el apartado 1.4.2.3). Esta dependencia es doble. En primer lugar, si la primera hipótesis presenta una verosimilitud muy superior o muy inferior a la de la correcta, la participación de la elocución, junto con el conjunto de sus posibles confusiones, en el entrenamiento discriminativo de los modelos será mínima. En segundo lugar, si la verosimilitud de una hipótesis es muy inferior a la de la primera, el modelo correspondiente a la hipótesis tampoco se reestimará. Combinando ambas influencias, si una hipótesis presenta una verosimilitud muy inferior a la primera, la cual a su vez es muy superior a la de la palabra correcta, el modelo correspondiente a esa hipótesis no se reestimará con independencia de la relación entre las verosimilitudes de la hipótesis y de la palabra correcta. Este comportamiento es producto de la propia definición de la función de cómputo de error y está plenamente justificado en entrenamiento discriminativo dependiente de la tarea. Así, si la regla de decisión utilizada en el entrenamiento es, como mínimo, semejante a la empleada en el reconocimiento, las situaciones en las que existe una hipótesis de verosimilitud muy superior a la de la palabra correcta corresponden, en general, a casos que podemos considerar como degenerados: ruido, mala pronunciación, etc. En estos casos

[3]Es importante reseñar que en esta serie de experimentos no se intenta equilibrar las tasas de inserción y borrado. Esta estrategia resulta sumamente beneficiosa y es de fácil integración y ajuste en condiciones reales de reconocimiento. Así, su utilización permite aumentar la tasa de cadenas de dígitos correctas del experimento **Base** desde los 89,1% de la tabla 2.4, hasta un 92,5% —superior a cualquiera de los esquemas de entrenamiento discriminativo considerados—. Más adelante, los experimentos presentados en esta tesis sí incluyen esta compensación, observándose igualmente el beneficio de la utilización de entrenamiento discriminativo, con el aliciente añadido de partir de un experimento de referencia de mayores prestraciones.

es más conveniente dejar la elocución fuera del entrenamiento, tal y como apuntan Reichl y Ruske [98], a intentar evitar la confusión. Por otro lado, una elocución es reconocida erróneamente si y sólo si la primera hipótesis es de mayor verosimilitud que la correcta. Así pues es razonable no abordar la eliminación de un error en tanto en cuanto haya otros errores de mayor verosimilitud, ya que son estos últimos los que, en todo caso, marcarán la existencia de un error de clasificación o no.

En entrenamiento discriminativo de unidades subléxicas para su aplicación al reconocimiento del habla continua, el efecto del asombrado de hipótesis es muy distinto. En este caso el desconocimiento de la tarea en la que van a ser empleados los modelos acústicos hace imposible que las reglas de decisión empleadas en entrenamiento y reconocimiento sean iguales. En su lugar es necesario utilizar una tarea de reconocimiento independiente de la que se va a reconocer, y calculable a partir de una base de entrenamiento balanceada fonéticamente. La tarea habitualmente utilizada, como se vio en el apartado 2.1.1, es la decodificación acústico fonética, bien sobre frases completas, bien sobre segmentos acústicos. Ahora bien, los errores cometidos en decodificación acústico fonética son muy distintos a los que encontramos en el reconocimiento del habla continua. En decodificación acústico fonética el resultado del reconocimiento es la cadena de unidades subléxicas de mayor verosimilitud sin restricciones gramaticales o, como mucho, aplicando una gramática estocástica o fonotáctica. En este tipo de reconocimiento es determinante el modelado acústico de las unidades subléxicas, el cual, en generar, podemos considerar como bastante pobre. Así, la tasa de error en la decodificación acústico fonética independiente del locutor de TIMIT es del orden del 40% si se utilizan modelos de fonema. Además, es habitual que un fonema o grupo de fonemas presente una larga lista de hipótesis incorrectas de mayor verosimilitud que la secuencia correcta, sin que ello signifique que la elocución sea defectuosa. Sin embargo, muchos de estos errores son de imposible comisión una vez se introducen las restricciones gramaticales de una tarea concreta de reconocimiento de habla continua. Así, los mismos modelos de fonema que provocan una tasa de error del 40% en el reconocimiento de TIMIT, presentan una tasa de error de fonema de sólo el 2% en el reconocimiento de TIDIGITS con las restricciones gramáticales propias de las cadenas de dígitos. Se puede concluir, al menos en el caso dado de ejemplo, que en reconocimiento del habla continua las características de la tarea son tan importantes o más que el propio modelado acústico de las unidades subléxicas.

Desde el punto de vista de la decodificación acústico fonética, la introducción de restricciones gramaticales puede ser interpretada como la *desaparición* de determinadas hipótesis de la lista considerada. Así, existirán combinaciones de sonidos aceptadas por la gramática empleada en la decodificación acústico fonética que no serán admitidas por la gramática de la tarea. El efecto de considerar estas hipótesis tanto en entrenamiento de mínimo error de clasificación como en entrenamiento de máxima información mutua es doble: en primer lugar, es evidente que se intentará evitar errores de imposible comisión; pero desconociendo la tarea concreta en que será empleado el sistema, todo lo que podemos hacer es pretender eliminar todos los errores posibles y, de este modo, eliminar también aquéllos susceptibles de ser cometidos cuando la tarea se concreta. De mayor gravedad es la influencia de estas hipótesis provocada por el efecto de asombrado. Debido a él, la influencia de cada una de las hipótesis depende de la verosimilitud del resto, especialmente de las de mayor verosimilitud. En concreto, si una hipótesis presenta una verosimilitud muy superior al resto, éstas no participarán en la reestimación. Si, además, la primera hipótesis no va a ser admitida por la gramática, estaremos perdiendo la oportunidad de aprovechar

la elocución para eliminar otros errores que, aún siendo de menor verosimilitud, sí son permitidos.

2.2.2 Entrenamiento de mínima confusibilidad

Los efectos del asombrado de hipótesis en entrenamiento discriminativo de sistemas de reconocimiento del habla continua indican la necesidad de plantear una función de coste que no presente este problema. Tal función debe garantizar que la influencia de una hipótesis no depende de la verosimilitud del resto. Para ello, debe ser definida de manera que las derivadas cruzadas con respecto a parámetros de modelos correspondientes a errores distintos sea cero. En cualquier otro caso, el valor del gradiente respecto de un determinado error dependerá del resto de errores y aparecerá el asombrado de hipótesis. La función de coste para la elocución de entrenamiento x^i debe ser, por tanto, de la forma:

$$l(x^i, \Lambda) = \sum_{j \neq i} l_j(x^i, \lambda_i, \lambda_j) \tag{2.1}$$

Donde cada error contribuye a la función de coste sin dependencias cruzadas entre errores distintos. Una posible elección de $l_j(x^i, \lambda_i, \lambda_j)$ es la función de error utilizada en entrenamiento de mínimo error de clasificación, pero involucrando sólo a los modelos de las palabras i y j. La función de coste, en adelante confusibilidad, es un cómputo del número de errores posibles, de manera que una misma elocución del entrenamiento contribuye a la función de coste global de manera proporcional al número de confusiones con verosimilitud mayor que la de la palabra correcta.

$$l_j(x^i, \Lambda) = l_j(x^i, \lambda_i, \lambda_j) = \frac{1}{2}\left(1 + \tanh \frac{\log \mathcal{P}(x^i/\lambda_j) - \log \mathcal{P}(x^i/\lambda_i)}{d_0}\right) \tag{2.2}$$

La confusibilidad es muy semejante a la función de cómputo de error empleada en entrenamiento de mínimo error de clasificación, existiendo una situación en la que ambos métodos son equivalentes: cuando de las N hipótesis consideradas, $M \leq N$ tienen verosimilitudes semejantes entre sí, y las $N - M$ restantes, verosimilitudes muy inferiores tanto a la de las M de mayor valor como a la de la correcta. Este es el caso que se da, por ejemplo, cuando sólo una de las posibles hipótesis erróneas presenta verosimilitud comparable a la palabra correcta. También ocurre esta situación cuando todas las hipótesis erróneas presentan verosimilitudes muy parecidas entre sí, como sucede cuando se calculan las N hipótesis más probables en decodificación acústico fonética de frases de larga longitud. Si, además, consideramos palabras equiprobables, ambos métodos son también muy parecidos al entrenamiento de máxima información mutua.

2.2.2.1 Propiedades de la función de confusibilidad

Sensibilidad de la función de confusibilidad. La sensibilidad de la confusibilidad con respecto a los modelos correspondientes a los distintos errores determinados en el reconocimiento de las N hipótesis más verosímiles para la secuencia de entrenamiento

x_n^i es:

$$
\begin{aligned}
S_{\lambda_{j\neq i}}(l(x_n^i,\Lambda))\big| &= \left\|\nabla_{\lambda_j}\log l(x_n^i,\Lambda)\right\|_{j\neq i} = \\
&= \left\|\frac{\nabla_{\lambda_j} l(x_n^i,\Lambda)}{l(x_n^i,\Lambda)}\right\|_{j\neq i} \\
&= \left\|\frac{\nabla_{\lambda_j}\sum_{k\neq i} l_k(x_n^i,\lambda_k)}{l(x_n^i,\Lambda)}\right\|_{j\neq i} \\
&= \left\|\frac{\nabla_{\lambda_j} l_j(x_n^i,\lambda_j)}{l(x_n^i,\Lambda)}\right\|_{j\neq i} \\
&= \frac{1}{2d_0 l(x_n^i,\Lambda)\cosh^2(d_i j(x_n^i,\Lambda)/d_0)}\left\|\nabla_{\lambda_j}[\log\mathcal{P}(x_n^i/\lambda_j)-\log\mathcal{P}(x_n^i/\lambda_i)]\right\|_{j\neq i} \\
&= \frac{1}{2d_0 l(x_n^i,\Lambda)\cosh^2(d_i j(x_n^i,\Lambda)/d_0)}\left\|\nabla_{\lambda_j}\log\mathcal{P}(x_n^i,\lambda_j)\right\|_{j\neq i} \\
&= \frac{1}{2d_0 l(x_n^i,\Lambda)\cosh^2(d_i j(x_n^i,\Lambda)/d_0)} S_{\lambda_j}(\mathcal{P}(x_n^i,\lambda_j))\big|_{j\neq i}
\end{aligned}
\tag{2.3}
$$

Donde $d_i j(x_n^i,\Lambda) = \log\mathcal{P}(x^i/\lambda_j) - \log\mathcal{P}(x^i/\lambda_i)$ es una medida de la posibilidad de que se reconozca como w_j la elocución correspondiente a la w_i, x_n^i. La expresión de la ecuación 2.3 es muy semejante a las ya vistas para la función de cómputo de error en MCE, ecuación 1.22, y la información mutua en MMIE, ecuación 1.23, especialmente la primera. Así, la principal diferencia que aparece entre la sensibilidad de la función de confusibilidad y la de cómputo de error es la ausencia, en la primera, del término de reparto de la sensibilidad entre las distintas hipótesis erróneas. En entrenamiento de mínima confusibilidad, la sensibilidad de cada hipótesis se relaciona con la sensibilidad de la probabilidad de generación de la elocución mediante términos constantes y un término, $1/\cosh^2(d_i j(x_n^i,\lambda_i,\lambda_j))$, que vale cero si la probabilidad de generación de la secuencia por el modelo incorrecto es muy distinta a la del modelo correcto, siendo máxima cuando ambas probabilidades son idénticas. El efecto de este término, que también aparece en MCE, pero no en MMIE (ver el apartado 1.4.2.3), es el de centrar el esfuerzo de la reestimación en aquellas confusiones producidas o evitadas con un pequeño margen de probabilidad. Dado que la variación en la probabilidad de generación necesaria para evitar estas confusiones es bajo, es de preveer que la modificación de los parámetros necesaria para evitarlas será pequeña. Por tanto, la confusibilidad es especialmente sensible a aquellas confusiones que pueden ser evitadas con menor esfuerzo.

Comparación del entrenamiento de mínima confusibilidad con MCE y MMIE. Las ventajas de utilizar la función de confusibilidad en lugar del cómputo de errores o la información mutua en entrenamiento discriminativo de sistemas de reconocimiento del habla continua son tres: en primer lugar, la ya comentada eliminación del asombrado de hipótesis; en segundo lugar, el hecho de enfocar el esfuerzo de reestimación en los errores más fácilmente eliminables garantiza que el máximo número de posibles errores en reconocimiento del habla continua es eliminado con la mínima perturbación de los parámetros del sistema; finalmente, el tratamiento independiente de cada uno de los

posibles errores facilita la inclusión de algún tipo de ponderación de los mismos cuando la gramática de la tarea es conocida, aunque sólo se disponga para el entrenamiento de los modelos acústicos una base de datos independiente de la tarea. Así, si del conocimiento de la tarea se concluye que la confusión entre dos segmentos v_i y v_j es mucho más grave que la confusión entre v_k y v_l, parece razonable que en la reestimación de los modelos se dé más importancia a los casos en los que aparece la primera confusión que en los que lo hace la segunda. Este mecanismo de adaptación será tratado en profundidad en el próximo apartado.

Como única desventaja de la utilización del criterio de mínima confusibilidad frente al de mínimo error de clasificación hay que señalar que no se adopta directamente la misma regla de decisión utilizada en el reconocimiento. Si bien la renuncia a reflejar la regla de decisión está justificada —y, de hecho, el autor la considera necesaria— para el caso en que los materiales de entrenamiento y reconocimiento son de características distintas, en el caso de no ser así, y disponer de una base de entrenamiento específica a la tarea a reconocer, siempre será mejor minimizar el error de clasificación que la confusibilidad. Ello es debido a que una mejora en la tasa de error de clasificación siempre implica una mejora en la tasa de reconocimiento del sistema, en tanto que podemos disminuir la confusibilidad disminuyendo la verosimilitud de un cierto número de hipótesis erróneas pero de verosimilitud no máxima, sin que ello se traduzca en una mejora de la tasa de reconocimiento si las hipótesis de máxima verosmilitud no son evitadas. De hecho esto es una consecuencia de la necesidad del asombrado de hipótesis en entrenamiento discriminativo dependiente de la tarea, necesidad que se convierte en inconveniente cuando el entrenamiento discriminativo se realiza a partir de una base de datos independiente de la misma.

2.2.2.2 Resultados obtenidos con entrenamiento de mínima confusibilidad en el reconocimiento de TIDIGITS

El entrenamiento de mínima confusibilidad fue presentado inicialmente en [81] conjuntamente con la utilización de segmentos acústicos de longitud limitada. En la tabla 2.5 se detallan los resultados de ese artículo que permiten comparar las prestaciones obtenidas utilizando entrenamiento de mínima confusibilidad y de mínimo error de clasificación a partir de las frases de TIMIT y utilizando esos modelos en el reconocimiento de TIDIGITS. Dado que ambos métodos son casi coincidentes cuando la tarea optimizada en el entrenamiento es la decodificación acústico fonética de frases completas, sólo se dan los resultados para cadenas de cinco fonemas con los de los extremos forzados a ser reconocidos correctamente. El tanto por ciento de cadenas reconocidas incorrectamente pasa de un 10,9% de los modelos entrenados aplicando el criterio de máxima verosimilitud, a un 8,2% si se minimiza el error de clasificación —una mejora del 25%—, y a un 7,4% si se minimiza la confusibilidad —una mejora del 32%—. Este resultado confirma la utilidad del empleo de la función de confusibilidad en entrenamiento de unidades subléxicas para el reconocimiento del habla continua.

Experimento	Tarea	Error	Sust	Inse	Borr	Acierto
Base	DAF	39,9	21,2	12,3	6,4	72,4
MCE	DAF	30,8	18,6	7,3	4,9	76,5
EMC	DAF	31,4	17,8	8,5	5,1	77,1

Experimento	Tarea	Error	Sust	Inse	Borr	Acierto	Corr
Base	digitos	3,7	1,3	1,5	0,9	97,8	89,1
MCE	digitos	2,9	1,2	1,1	0,6	98,2	91,8
EMC	digitos	2,5	1,0	0,9	0,6	98,4	92,6

Tabla 2.5: *Resultados obtenidos en DAF y el reconocimiento de TIDIGITS empleando modelos de fonema entrenados aplicando **MCE** y **EMC** en segmentos de cinco fonemas. A pesar de la pequeña diferencia entre ambos criterios de optimización, el resultado obtenido con EMC es ligeramente superior al obtenido con MCE.*

2.3 Adaptación a la Tarea Aplicando Entrenamiento de Mínima Confusibilidad en Bases de Datos de Propósito General

En los apartados precedentes se han presentado y argumentado dos modificaciones sobre los sistemas estándar de entrenamiento discriminativo para su adaptación a la problemática específica del entrenamiento de modelos subléxicos en su aplicación al reconocimiento del habla continua: el uso de segmentos acústicos de longitud limitada y de la función de confusibilidad. La línea argumental seguida no pretende —tampoco lo conseguiría— demostrar la necesidad de ambas, aunque su empleo se justifica en el hecho que tampoco las otras alternativas presentadas hasta la fecha —en particular, la minimización del error de clasificación en decodificación acústico fonética de frases completas— dan una respuesta satisfactoria y, frente a ellas, la utilización tanto de segmentos acústicos de longitud limitada, como de la función de confusibilidad, presentan ventajas que las hacen especialmente interesantes. En este apartado se parte de la base de que la confusibilidad es una medida de calidad como cualquier otra y de que los segmentos acústicos de longitud limitada permiten abordar la minimización de la tasa de error de fonemas —y no de frases— en decodificación acústico fonética, aunando una buena representación de los errores cometidos en habla continua con la facilidad de obtención de tales segmentos en bases de entrenamiento independientes de la tarea. A partir de estas premisas, se plantea la minimización de la confusibilidad dependiente de la tarea, la cual puede ser aproximada mediante la confusibilidad definida a partir de segmentos acústicos provinientes de una base de datos independientes de la tarea y del conocimiento de la gramática de ésta. De este modo se consigue la *adaptación* de los modelos acústicos a una tarea concreta sin necesidad de disponer de una base de datos específica. Finalmente, y como una extensión de esta adaptación al caso en que la tarea es desconocida, se propone la realización de entrenamiento discriminativo independiente de la tarea como una adaptación al idioma, considerando éste como una tarea de tareas que engloba a cualquier otra. Los resultados obtenidos en el entrenamiento de modelos de fonema para su utilización en el reconocimiento de cadenas de dígitos avalan plenamente ambas aproximaciones, obteniéndose de su acumulación —primero se obtienen modelos

independientes de la tarea y a continuación se adaptan a la misma— una reducción a la mitad del número de errores cometido empleando modelos entrenados mediante el criterio de máxima verosimilitud.

2.3.1 Adaptación a la tarea usando el criterio de mínima confusibilidad y bases de datos ilimitadas

Considérese, en primer lugar, que la base de datos de entrenamiento es lo suficientemente grande como para abarcar suficientes elocuciones de entrenamiento de cada posible frase de cualquier tarea de una lengua. En ese caso, la base de datos sería independiente de la tarea, pero permitiría, al mismo tiempo, realizar un entrenamiento dependiente de la misma. Basta para ello con sólo considerar en la función a optimizar los errores permitidos por la tarea, es decir: aquéllos en los que una frase válida en la tarea es confudida con otra igualmente válida.

Sea Λ el conjunto de parámetros del sistema a reestimar; W, la gramática de la tarea; $x_n^i \in X$, la n-ésima elocución de entrenamiento, indicando el superindice que esta elocución es un representante de la palabra $w_i \in W$; y $g_j(x_n^i) = \log \mathcal{P}(x_n^i/\lambda_j)$ el logaritmo de la probabilidad de generación de la elocución x_i^n por el modelo λ_j siguiendo el alineado de máxima probabilidad entre ambos. El error $[w_i \rightarrow w_{j\neq i}]$ es posible para x_n^i si, y sólo si, la probabilidad de generación de la elocución por λ_j es mayor que por λ_i. Se define la posibilidad de error, $\mathcal{E}_{ij}(x_n^i, \Lambda) = \mathcal{E}_{ij}(x_n^i, \lambda_i, \lambda_j)$, como uno, si $j \neq i$ y el error es posible, y cero, en caso contrario:

$$\mathcal{E}_{ij}(x_n^i, \Lambda) = \begin{cases} 1 & g_j(x_n^i) \geq g_i(x_n^i) \\ 0 & g_j(x_n^i) < g_i(x_n^i) \end{cases} \qquad \forall\, j \neq i \tag{2.4}$$

$$\approx \frac{1}{2}\left(1 + \tanh \frac{g_j(x_n^i) - g_i(x_n^i)}{G_0}\right)\Bigg|_{j\neq i} \tag{2.5}$$

Donde se utiliza la aproximación dada por la ecuación 2.5 para garantizar la continuidad de la función —necesaria para poder utilizar un algoritmo de búsqueda de gradiente en su optimización—. Se define la *confusibilidad individual* de una elocución como el número total de errores que es posible cometer en el reconocimiento de una frase concreta y aplicando una gramática determinada. Siendo $1(\cdot)$ una función que devuelve uno si su argumento es cierto y cero en caso contrario, $1([w_i \rightarrow w_j] \in W)$ es un indicador de que la comisión de la confusión entre w_i y w_j puede aparecer en el reconocimiento de la tarea W. Utilizando este indicador, la confusibilidad individual de la elocución x_n^i en la tarea utilizando los modelos Λ, puede expresarse en función de las posibilidades de comisión de todas las posibles confusiones de la elocución de entrenamiento, con independencia de que éstas puedan ocurrir o no en el reconocimiento de la tarea:

$$\mathcal{CI}(x_n^i, \Lambda, W) = \sum_{j\neq i} \mathcal{E}_{ij}(x_n^i, \lambda_i, \lambda_j) 1([w_i \rightarrow w_j] \in W) \tag{2.6}$$

A partir de la confusibilidad individual de todas las elocuciones de entrenamiento representantes de la palabra w_i, es posible estimar la esperanza del número de errores de posible comisión para las elocuciones de entrenamiento de la misma, la *confusibilidad de clase*, $\mathcal{CC}$. Siendo $X^i = \{x_n^i\}$ el subconjunto de la base de entrenamiento formado por todos los representantes de la palabra w_i, y $|X^i|$ el número de elementos del subconjunto, la

confusibilidad de clase de la palabra w_i vale:

$$CC_i(X^i, \Lambda, W) = E\{CI(x_n^i, \Lambda, W)\} = \quad (2.7)$$

$$= \frac{1}{|X^i|} \sum_{x_n^i \in X^i} CI(x_n^i, \Lambda, W) \quad (2.8)$$

La *confusibilidad global* —o, simplemente, confusibilidad— del sistema se define como el número esperado de errores distintos que éste puede cometer en el reconocimiento de la tarea:

$$C(X, \Lambda, W) = \sum_i f_W(w_i) CC_i(X^i, \Lambda, W) \quad (2.9)$$

Donde $f_w(w_i)$ es la frecuencia de aparición de la palabra w_i en la gramática W. Agrupando las ecuaciones 2.5-2.9, llegamos a:

$$C(X, \Lambda, W) = \sum_i f_W(w_i) \frac{1}{|X^i|} \sum_{x_n^i \in X^i} \sum_{j \neq i} \mathcal{E}_{ij}(x_n^i, \lambda_i, \lambda_j) 1([w_i \rightarrow w_j] \in W) \quad (2.10)$$

En esta última expresión, la confusibilidad del sistema combina tres tipos distintos de información: el término $\mathcal{E}_{ij}(x_n^i, \lambda_i, \lambda_j)$ es el único que depende de las características acústicas de las frases de entrenamiento y de los modelos acústicos; $|X^i|$ es el número de apariciones de realizaciones de cada palabra del vocabulario en la base de datos de entrenamiento y sólo depende de la estructura de ésta; finalmente, tanto el término $f_W(w_i)$ como la condición explícita $1([w_i \rightarrow w_j] \in W)$, sólo dependen de las características de la tarea a reconocer. Reordenando el orden de los sumatorios en la ecuación 2.10, se obtiene:

$$C(X, \Lambda, W) = \sum_i \sum_{x_n^i \in X^i} \sum_{j \neq i} \frac{f_W(w_i) 1([w_i \rightarrow w_j] \in W)}{|X^i|} \mathcal{E}_{ij}(x_n^i, \lambda_i, \lambda_j) \quad (2.11)$$

$$= \sum_i \sum_{x_n^i \in X^i} \sum_{j \neq i} \frac{\mathcal{R}(w_i, w_j, W)}{|X^i|} \mathcal{E}_{ij}(x_n^i, \lambda_i, \lambda_j) \quad (2.12)$$

Donde $\mathcal{R}(w_i, w_j, W) = f_W(w_i) 1([w_i \rightarrow w_j] \in W)$ —en adelante, la *relevancia* del error $[w_i \rightarrow w_j]$ en la tarea W— es una medida del efecto que tiene, en el reconocimiento de la tarea W, la confusión entre w_i y w_j. Si la gramática permite el error, $f_W(w_i)$ es distinto de cero, $1([w_i \rightarrow w_j] \in W)$ vale uno y el numero esperado de errores es proporcional a la frecuencia de aparición en la tarea de la palabra confundida. Si la palabra de entrenamiento no aparece en la tarea, $f_W(w_i) = 0$, o si la gramática no permite la comisión del error, $1([w_i \rightarrow w_j] \in W) = 0$, la contribución neta del error a la confusibilidad del sistema es nula.

2.3.2 Adaptación a la tarea usando el criterio de mínima confusibilidad en segmentos de longitud limitada

La ecuación 2.11 proporciona un mecanismc de adaptación a la tarea utilizando para ello una base de datos independiente de la misma. De hecho, la aplicación de esta ecuación es equivalente a descartar cualquier frase de entrenamiento de imposible aparición en la tarea a reconocer, permitiendo, así mismo, sólo el reconocimiento de frases válidas en la

tarea. Evidentemente, para que este esquema sea provechoso es necesario que la base de datos contenga realizaciones bastantes de cada posible frase en la tarea, pero, dado que ésta es desconocida en el momento de diseñar la base de datos de entrenamiento, esta condición implica que la base de datos debe ser de dimensiones gigantescas para poder abarcar cualquier posible tarea —aunque existen propuestas al respecto, éstas son más teóricas que prácticas [14]—. Ahora bien, la posibilidad de que se cometa un error léxico —entre palabras del vocabulario— sólo depende de la relación entre la verosimilitud de la palabra correcta y la de la incorrecta. Esta relación entre verosimilitudes puede ser calculada utilizando tanto la frase entera como a partir de las verosimilitudes obtenidas en cada uno de los segmentos en que ésta puede ser dividida según el criterio apuntado en el apartado 2.1.3.1. Dado que la definición dada en la ecuación 2.5 de la posibilidad de comisión de error es una función continua y monótonamente creciente de la relación de verosimilitudes, es posible determinar la posibilidad de error a nivel léxico a partir de las posibilidades de error de cada uno de los segmentos. Así, siendo T el número de segmentos en los que se divide la confusión entre la elocución de entrenamiento w_i y la reconocida w_j; $s^k_{n,t}$ cada uno de los T segmentos en que se divide la elocución de entrenamiento; y v_k el contenido acústico del segmento $s^k_{n,t}$, la posibilidad de comisión del error léxico es, en general:

$$\mathcal{E}_{ij}(x^i_n, \lambda_i, \lambda_j) = \mathcal{F}(\mathcal{E}_{kl}(s^k_{n,t}, \lambda_k, \lambda_l) : \quad t = 1 \ldots T) \tag{2.13}$$

Por ejemplo, la confusión entre los dígitos en inglés /five/ y /nine/ —a nivel subléxico: [f ay v] y [n ay n]— depende de la posibilidad de que [f ay] se confunda con [n ay], y de que [ay v] lo haga con [ay n]. Si ninguna de las dos confusiones segmentales es posible, tampoco la confusión léxica lo será. Si ambas confusiones son posibles, el error léxico también lo será. En el resto de casos, una de las confusiones es posible y la otra no, la relación entre la posibilidad de los errores segmentales y la del error léxico es menos evidente y dependerá del margen con el que se reconoce un segmento y otro. Dada la no linealidad de la función de posibilidad de error tanto a nivel léxico como al segmental, la forma de $\mathcal{F}(\cdot)$ es relativamente intrincada. Ahora bien, existen dos situaciones en las que la relación entre la posibilidad del error léxico y la de los segmentos necesarios para su comisión es inmediata: si sólo hay un segmento erróneo, la posibilidad de comisión del error léxico es idénticamente igual a la del error segmental; si es necesario más de un error segmental para provocar el error léxico, pero las verosimilitudes de los segmentos erróneos son muy parecidas a las de los segmentos correctos —y debido a que la función de posibilidad definida en 2.5 es casi lineal cuando su argumento es próximo a cero—, entonces la posibilidad de error léxico es aproximadamente igual a la suma de las posibilidades de error de cada uno de los segmentos. La importancia de estos dos casos —un único error segmental provoca el error léxico, o todos los errores segmentales presentan verosimilitudes muy semejantes a las de los segmentos correctos—, justifica la aproximación de la posibilidad de error léxico como la suma de las posibilidades de los errores segmentales necesarios. Así, la ecuación 2.13 puede aproximarse mediante:

$$\mathcal{E}_{ij}(x^i_n, \lambda_i, \lambda_j) \approx \sum_t \mathcal{E}_{kl}(s^k_{n,t}, \lambda_k, \lambda_l) \tag{2.14}$$

Esta exprexión es equivalente a considerar que cada error léxico contribuye a la confusibilidad global del sistema tantas veces como errores segmentales sean necesarios para

cometerlo. Así, la confusión entre /five/ y /nine/, representa dos errores distintos a nivel segmental ([f ay] → [n ay] y [ay v] → [ay n]).

Para poder expresar la confusibilidad dependiente de la tarea únicamente en función del material acústico presente en la base de datos y de su estructura y la de la gramática de la tarea a reconocer, es necesario realizar una nueva aproximación consistente en suponer que las características acústicas de los segmentos no dependen del contexto en el que se encuentran. En el ejemplo de la confusión de /five/ y /nine/, esta aproximación equivale a considerar que las características de las partícula inicial [f ay] son esencialmente las mismas, provengan de la palabra /five/, o de cualquier otra —/fight/, /fine/, etc.—. En ese caso, la confusión de /fight/ con /night/ ([f ay t] → [n ay t]), aporta tanta información acerca de la posibilidad de confudir la parte inicial de la palabra como la que hubiera proporcionado la propia disponibilidad de la misma. Es decir, podemos conocer la esperanza de la posibilidad de comisión de cada error léxico a partir de las esperanzas de la posibilidad de comisión de error entre cada uno de los segmentos necesarios.

$$E\{\mathcal{E}_{ij}(x_n^i, \lambda_i, \lambda_j)\} \approx E\{\sum_t \mathcal{E}_{kl}(s_{n,t}^k, \lambda_k, \lambda_l)\} \tag{2.15}$$

$$\approx \sum_t E\{\mathcal{E}_{kl}(s_{n,t}^k, \lambda_k, \lambda_l)\} \tag{2.16}$$

Introduciendo esta aproximación en la expresión 2.8, y suponiendo, además, que la confusión entre cada par de palabras w_i y w_j da lugar siempre a la misma descomposición en T segmentos acústicos $[w_i \rightarrow w_j] \Rightarrow [v_l(1) \rightarrow v_k(1) \quad v_l(2) \rightarrow v_k(2) \quad \ldots \quad v_l(T) \rightarrow v_k(T)]$—, se obtiene la relación entre la confusibilidad de clase de cada palabra y la esperanza de la posibilidad de comisión de cada error segmental:

$$CC_i(X^i, \Lambda, W) = E\{\mathcal{CI}(x_n^i, \Lambda, W)\} = \tag{2.17}$$

$$\approx \sum_{w_i \in W} \sum_{j \neq i} \sum_t E\{\mathcal{E}_{kl}(s_{n,t}^k, \lambda_k, \lambda_l)\} \tag{2.18}$$

Esta expresión es equivalente al sumatorio para todos los segmentos a que da lugar cada posible confusión de cada realización de la palabra w_i. En principio, estos segmentos dependen de la confusión cometida, ya que sus fronteras vienen determinadas por la presencia de dos o más unidades subléxicas correctamente reconocidas (ver el apartado 2.1.3.1). Una alternativa a este procedimiento consiste en considerar todos los posibles segmentos de cada frase de entrenamiento, $s_k \in X^i$, estimar todos los posibles reconocimientos en cada uno de ellos, v_l, y utilizar únicamente aquellas confusiones entre segmentos que participan en la confusión entre palabras. Para ello se vuelve a echar mano del operador $1(\cdot)$, que devuelve uno si y sólo si su argumento es cierto. Así, $1([v_k \rightarrow v_l] \subset [w_i \rightarrow w_j])$, vale uno si la confusión entre los segmentos v_k y v_l participa en la confusión entre las palabras w_i y w_j. Utilizando esta expresión, se tiene :

$$CC(X^i, \Lambda, W) \approx \frac{1}{|X^i|} \sum_{s_k \in X^i} \sum_{j \neq i} \left[\sum_{l \neq k} \mathcal{E}_{kl}(s_k, \lambda_k, \lambda_l) 1([v_k \rightarrow v_l] \subset [w_i \rightarrow w_j]) \right] 1([w_i \rightarrow w_j] \in W) \tag{2.19}$$

Reordenando los sumatorios y agrupando términos constantes:

$$CC(X^i, \Lambda, W) \approx \sum_{s_k \in X^i} \sum_{l \neq k} \left[\frac{\sum_{j \neq i} 1([v_k \rightarrow v_l] \subset [w_i \rightarrow w_j]) 1([w_i \rightarrow w_j] \in W)}{|X^i|} \right] \mathcal{E}_{kl}(s_k, \lambda_k, \lambda_l) \tag{2.20}$$

Y, utilizando esta expresión de la confusibilidad de clase, la confusibilidad global del sistema vale:

$$C(X, \Lambda, W) \approx \sum_i f_W(w_i) \sum_{s_k \in X^i} \sum_{l \neq k} \left[\frac{\sum_{j \neq i} 1([v_k \rightarrow v_l] \subset [w_i \rightarrow w_j]) 1([w_i \rightarrow w_j] \in W)}{|X^i|} \right] \mathcal{E}_{kl}(s_k, \lambda_k, \lambda_l) \tag{2.21}$$

En esta expresión sólo el término $\mathcal{E}_{kl}(s_k, \lambda_k, \lambda_l)$ depende de las características de los segmentos y modelos acústicos. Es, por tanto, equivalente a la expresión dada para la confusibilidad dependiente de la tarea (ecuación 2.11), sustituyendo las frases de entrenamiento x_n por los segmentos que las forman, s_k; y utilizando un término de relevancia para cada uno de los posibles errores segmentales que sólo depende de los segmentos confundidos y las características de la tarea a reconocer:

$$C(X, \Lambda, W) \approx \sum_{s_k \in X} \sum_{l \neq k} \frac{\mathcal{R}(v_k, v_l, W)}{|s_k|} \mathcal{E}_{kl}(s_k, \lambda_k, \lambda_l) \tag{2.22}$$

Donde la la relevancia dada a cada error segmental $[v_k \rightarrow v_l]$ es igual a la frecuencia de aparición de dicho error en el conjunto formado por los $|W| \times |W|$ errores léxicos posibles.

2.3.3 Cálculo aproximado de la relevancia. Adaptación al Idioma

Aunque la expresión de la ecuación 2.22 requiere el cálculo de la relevancia de cada posible error segmental, éste puede simplificarse notoriamente utilizando dos aproximaciones: en primer lugar, puede considerarse que cualquier segmento válido en la tarea puede ser confundido con cualquier otro. En ese caso, la frecuencia de aparición del error segmental es igual al producto de las frecuencias de aparición del segmento y su confusión. Por otro lado, ambas frecuencias de aparición pueden estimarse aproximadamente usando gramáticas estocásticas. En la experimentación realizada a lo largo de esta tesis, se consideró independencia entre los segmentos correctos y reconocidos, y la frecuencia de aparición de ambos se calculó utilizando el bigrama de las unidades subléxicas que aparecen en la tarea, B_T [80]. Así mismo, la frecuencia de aparición de los segmentos en la base de datos también se obtuvo a partir del bigrama del material de entrenamiento, B_E. Es decir, se realizó la aproximación:

$$\frac{\mathcal{R}(v_k, v_l, W)}{|s_k|} \approx \frac{f_W(v_k) f_W(v_l)}{f_E(v_k)} \approx \frac{B_T(v_k) B_T(v_l)}{B_E(v_k)} \tag{2.23}$$

La expresión de la relevancia de cada error segmental requiere del conocimiento previo de la tarea a reconocer. Si ésta no es conocida a priori, puede considerarse que el objetivo del entrenamiento será minimizar el número de errores posibles en cualquier tarea o, lo que es lo mismo, en una tarea que abarque todas las posibles tareas del idioma —algo así como la base de datos considerada al principio del apartado 2.3.2—. Aunque una base de datos

que abarque todas las posibles tareas es irrealizable, sí es posible conocer la estadística de su contenido fonético, dado que coincidirá con la propia estadística del idioma. Así pues, la expresión 2.23 proporciona un mecanismo de adaptación a la tarea cuando la estadística de ésta es conocida, pero, también, un mecanismo para realizar entrenamiento discriminativo independiente de la tarea, cuando sólo se conoce la estadística del idioma. Es lo que, en adelante, se denominará *adaptación al idioma*.

2.3.4 Reconocimiento de TIDIGITS utilizando modelos de fonema entrenados con TIMIT

Tanto el esquema de adaptación a la tarea como el de adaptación al idioma han sido probados en el reconocimiento de las cadenas de TIDIGITS utilizando modelos acústicos de fonema entrenados con el corpus masculino de entrenamiento de la base de datos TIMIT. El punto de partida utilizado es el sistema entrenado aplicando el criterio de máxima verosimilitud. La confusibilidad del sistema es minimizada utilizando segmentos de cinco fonemas en los cuales ambos extremos son forzados a ser reconocidos correctamente. En cada segmento se determinan las doce hipótesis de mayor verosimilitud y, a partir de ellas, el gradiente de la función de confusibilidad. En esta serie de experimentos se optó por equilibrar las tasas de inserción y borrado penalizando las transiciones entre unidades. Esta modificación reporta un beneficio considerable —compárense los resultados obtenidos por los experimentos **base** de las tablas 2.5 y 2.6—, siendo fácil de incorporar a los sistemas reales de reconocimiento del habla. En reconocimiento, el factor de penalización se determina empíricamente; en entrenamiento de mínima confusibilidad se determina según se explica en el apartado A.2.2.

Se han probado cuatro configuraciones de entrenamiento distintas:

Base Modelos acústicos entrenados según el criterio de máxima verosimilitud.

Inde Modelos acústicos entrenados aplicando entrenamiento de mínima confusibilidad considerando igual relevancia a todas las posibles confusiones.

Lang Modelos acústicos entrenados aplicando entrenamiento de mínima confusibilidad independiente de la tarea, pero adaptado al idioma.

Adap Modelos acústicos entrenados aplicando entrenamiento de mínima confusibilidad adaptado a la tarea.

Todo Modelos acústicos entrenados aplicando entrenamiento de mínima confusibilidad, primero independiente de la tarea —el resultado de **Lang**—, y adaptado seguidamente a la tarea del mismo modo que **Adap**.

La tabla 2.6 muestra los resultados obtenidos con cada una de estas cuatro configuraciones. Es remarcable el hecho que todas las configuraciones de entrenamiento discriminativo permiten reducir la tasa de error del sistema. Ahora bien, la introducción de la relevancia de los errores en el idioma resulta de crucial importancia en la reducción de la tasa de error independiente de la tarea —compárense los resultados de **Inde** y **Lang**—. De hecho, el experimento de adaptación al idioma proporciona un mejor resultado que el adaptación a la propia tarea a reconocer. Este comportamiento puede deberse al poco aprovechamiento del material de TIMIT cuando se introducen las restricciones gramaticales de la tarea de reconocimiento de cadenas de dígito. Así, todo segmento con algún fonema, o transición

Experimento	Error	Sust	Inse	Borr	Acierto	Corr
Base	2,7	1,1	0,8	0,8	98,1	92,5
Inde	2,5	1,5	0,5	0,5	98,0	93,0
Lang	2,2	1,1	0,6	0,5	98,4	94,0
Adap	2,3	1,0	0,7	0,6	98,3	93,6
Todo	1,8	0,9	0,5	0,4	98,7	94,9

Tabla 2.6: *Resultados obtenidos en el reconocimiento de TIDIGITS empleando modelos de fonema entrenados con TIMIT. La mejora obtenida por la adaptación al idioma (**Lang**) y a la tarea (**Adap**) se acumulan al aplicarlas consecutivamente (**Todo**), proporcionando el mejor resultado obtenido hasta la fecha utilizando modelos de fonema en esta tarea.*

entre fonemas, no permitido en la tarea es descartado del entrenamiento, quedando sólo un 5% del material de TIMIT disponible para la adaptación a la tarea. No obstante, el resultado obtenido por la combinacón de entrenamiento adaptado al idioma seguido de adaptación a la tarea, **Todo**, proporciona un resultado muy superior al obtenido por cualquiera de ellos por separado, corroborando la utilidad de ambos planteamientos. Cabe señalar que el resultado obtenido por la estrategia combinada es el mejor que se ha conseguido nunca utilizando RAMSES con modelos de fonema independientes de la tarea y del contexto entrenados con este corpus.

2.4 Entrenamiento Discriminativo de Unidades Subléxicas Dependientes del Contexto

En el apartado precedente se ha visto como las técnicas de entrenamiento discriminativo, inicialmente propuestas para sistemas de reconocimiento dependientes de la tarea, permiten así mismo mejorar notoriamente las prestaciones de los sistemas de reconocimiento del habla continua, tanto si la tarea es conocida a priori como si no, utilizando bases de datos de entrenamiento de propósito general. En concreto, se ha comprobado la utilidad de la aplicación del criterio de mínima confusibilidad calculada sobre segmentos acústicos de corta longitud en la reestimación de HMM's de fonemas independientes del contexto. Un procedimiento distinto de incrementar igualmente las prestaciones de este tipo de sistemas consiste en el modelado explícito del contexto en el que aparece cada unidad acústica; es decir, la utilización de unidades subléxicas dependientes del contexto. Como se verá a lo largo de este apartado, la aplicación por separado de cada una de estas técnicas proporciona resultados semejantes, muy superiores a los obtenidos por sistemas basados en modelos de fonemas entrenados mediante el criterio de la máxima verosimilitud. Dada esta situación, parece evidente que la acumulación de ambas técnicas, esto es: el entrenamiento discriminativo de unidades subléxicas dependientes del contexto, debería proporcionar resultados aún mejores a los obtenidos por cada una de ellas independientemente. No obstante, y aunque en teoría los algoritmos de entrenamiento discriminativo no debieran verse afectados por el uso de un tipo u otro de unidad subléxica, este tipo de entrenamiento resulta mucho más complicado de llevar a cabo.

2.4.1 Entrenamiento de mínima confusibilidad de semifonemas

En principio, el hecho de utilizar unidades dependientes o no del contexto no debe implicar ninguna modificación sustancial a la hora de aplicar entrenamiento discriminativo. Los mismos algoritmos empleados, por ejemplo, en el entrenamiento de mínima confusibilidad de modelos de fonemas, deberían ser igualmente útiles a la hora de entrenar modelos de semifonema. No obstante, las unidades dependientes del contexto presentan una problemática muy específica que hace especialmente complicado la consecución de mejoras adicionales en las prestaciones de reconocimiento mediante la aplicación de entrenamiento discriminativo. Las razones de esta dificultad son varias:

1. El número de unidades a ser entrenado es muy superior en el caso de considerar explícitamente el contexto, con lo que el número de parámetros a estimar es también muy superior —en el caso de la experimentación con TIMIT presentada en esta tesis, esto es casi un millón de parámetros, utilizando semifonemas, frente a los 150.000 de los fonemas—.

2. La cantidad de material de entrenamiento necesario para entrenar discriminativamente este número superior de parámetros crece enormemente, debido a que el entrenamiento discriminativo considera errores entre palabras y no las realizaciones de éstas directamente.

3. Ambas técnicas tienen un comportamiento semejante: el entrenamiento discriminativo tiende a mejorar el modelado de las unidades en aquellos contextos en los que mayor número de errores se comete; la dependencia del contexto tiende a mejorar el modelado en todos ellos.

4. Las unidades dependientes del contexto proporcionan una tasa de error tan baja que conseguir mejoras adicionales resulta mucho más complicado.

Estas dificultades pueden resumirse en dos: el material de entrenamiento necesario es muy superior; y, tal vez, el entrenamiento discriminativo no sea capaz de reducir la tasa de error por debajo de lo que ya lo hace la dependencia del contexto, y viceversa. Como consecuencia, en ocasiones, los resultados obtenidos acumulando ambas técnicas han quedado, de hecho, por debajo de los alcanzados al aplicar una de ellas solamente [53].

2.4.2 Resultados experimentales utilizando modelos de semifonema entrenados con TIMIT

La problemática apuntada en el apartado anterior se pone de manifiesto en el entrenamiento de mínima confusibilidad aplicada a modelos de semifonema. Así, en la tarea de reconocimiento de TIDIGITS usando TIMIT, alguno de los parámetros que gobiernan el algoritmo cuando la unidad fonética empleada es el fonema deben ser alterados para que el método sea de utilidad también en el entrenamiento de semifonemas. Las modificaciones introducidas y el motivo de su introducción son las siguientes:

1. El corpus **test** de TIMIT —formado por 896 frases— ha sido añadido al corpus **train** —de 2608 frases— para realizar el entrenamiento de los modelos acústicos. Con esta modificación se consigue aumentar considerablemente la capacidad de entrenamiento del sistema, fundamental en el entrenamiento de unidades dependientes del contexto,

tanto si se aplica entrenamiento convencional de máxima verosimilitud, como si se aplica entrenamiento discriminativo, pero sobre todo en este último caso.

2. Aún con la utilización del corpus completo de TIMIT, se han detectado problemas de adaptación excesiva. Estos problemas han podido aliviarse considerablemente haciendo que el tamaño de la época —esto es: el número de frases que intervienen en cada actualización de los parámetros— aumente. En los experimentos aquí presentados se pasó de las alrededor de 250 frases utilizadas con los fonemas independientes del contexto, a unas 1000 frases por época con los semifonemas.

3. En el caso de adaptación a la tarea, el material útil presente en la base de entrenamiento es excesivamente pequeño —utilizando un bigrama, sólo el 2% de las transiciones posibles entre semifonemas aparece en la tarea de los dígitos—. Para aumentar la cantidad de material aprovechado en la adaptación a la tarea se ha optado por:

   - Suavizar el modelo del lenguaje de la tarea. Para ello se ha añadido a la probabilidad de transición entre unidades un término proporcional a la probabilidad de que esa transición se dé en el entrenamiento. Este término de suavizado sólo se añade a las transiciones que tienen como origen o destino una unidad presente en la tarea y abundantes en el entrenamiento.
   - Reducción de la longitud de los segmentos. En lugar los cinco fonemas, con los de los extremos forzados, utilizados en entrenamiento de fonemas, la longitud del segmento se ha reducido a 3 semifonemas sin forzar los extremos.

Una modificación adicional del entrenamiento de semifonemas con respecto al de fonemas es la estructura de los modelos de Markov. Así, en tanto en aquél cada fonema se representa con un modelo de cuatro estados de los cuales es posible evitar uno de los dos centrales, al utilizar semifonemas el modelo utilizado para cada uno consiste de dos estados, ninguno de los cuales puede ser evitado. De este modo, cada fonema queda representado por un modelo de cuatro estados, sin que se pueda evitar ninguno de ellos.

Todas estas modificaciones con respecto al caso del entrenamiento de fonemas hacen que también el experimento de referencia —entrenamiento de máxima verosimilitud de modelos de fonema— vea alterados los resultados obtenidos. Así pues, es necesario repetir también el experimento de referencia de manera que los resultados sean homogéneos.

El conjunto de semifonemas empleado se ha obtenido aplicando un algoritmo de selección de unidades basado en combinar la capacidad de generalización de los árboles lógicos basados en rasgos fonéticos, con la homegeidad proporcionada por los algoritmos aglomerativos [66]. El número de agrupamientos resultantes utilizando el corpus masculino completo de TIMIT es 344. Los agrupamientos presentan 300 realizaciones, como mínimo, cada uno. A modo de ejemplo, la tabla 2.7 muestra alguno de los agrupamientos resultantes de este proceso.

Se comparan ocho posibles configuraciones de entrenamiento en condiciones homogéneas:

BaseFon Modelos de fonema entrenados según el criterio de máxima verosimilitud.

LangFon Modelos de fonema entrenados aplicando entrenamiento de mínima confusibilidad independiente de la tarea (pero adaptado al idioma).

ix+C(1)	ix+m,ix+n,ix+ng
ix+C(3)	ix+s,ix+z
m+C(1)	m:,m+s,m+z
m+C(5)	m+p,m+b
C-t(0)	:t,C-t,k-t,f-t,p-t,b-t,l-t
C-t(1)	ih-t,ix-t,uw-t,ax-t,ow-t,er-t,r-t,uh-t
C-t(2)	eh-t,ae-t,ah-t,aw-t,aa-t,ao-t
C-t(3)	n-t,en-t
C-t(4)	s-t,zh-t
C-t(5)	ey-t,ay-t,iy-t
...	...

Tabla 2.7: *Ejemplos de los agrupamientos de semifonemas empleados en la experimentación. Estos agrupamientos se han obtenido aplicando el algoritmo de selección de unidades descrito en [66].*

AdapFon Modelos de fonema entrenados aplicando entrenamiento de mínima confusibilidad adaptado a la tarea.

TodoFon Modelos de fonema entrenados aplicando entrenamiento de mínima confusibilidad, primero independiente de la tarea —el resultado de **LangFon**—, y adaptado a continuación a la tarea del mismo modo que **AdapFon**.

BaseSefo Modelos de semifonema entrenados según el criterio de máxima verosimilitud.

LangSefo Modelos de semifonema entrenados aplicando entrenamiento de mínima confusibilidad independiente de la tarea.

AdapSefo Modelos de semifonema entrenados aplicando entrenamiento de mínima confusibilidad adaptado a la tarea.

TodoSefo Modelos de semifonema entrenados aplicando entrenamiento de mínima confusibilidad, primero independiente de la tarea y después adaptado a la misma.

2.4.2.1 Reconocimiento de TIDIGITS utilizando modelos de semifonema entrenados con TIMIT

La tabla 2.8 muestra los resultados obtenidos con cada una de estas configuraciones. El tanto por ciento de cadenas reconocidas erróneamente aparece representado en la gráfica 2.1.

A la vista de estos resultados puede concluirse:

1. La utilización del semifonema como unidad acústica resulta muy beneficiosa frente al uso del fonema. Así, en el caso de entrenamiento convencional de máxima verosimilitud, se obtiene una reducción de casi el 40% en la tasa de error de cadenas —compárense BaseFon y BaseSefo—.

2. El entrenamiento discriminativo de los modelos de semifonema proporciona resultados claramente superiores a los obtenidos utilizando máxima verosimilitud, tanto en independencia de la tarea como en adaptación a la misma.

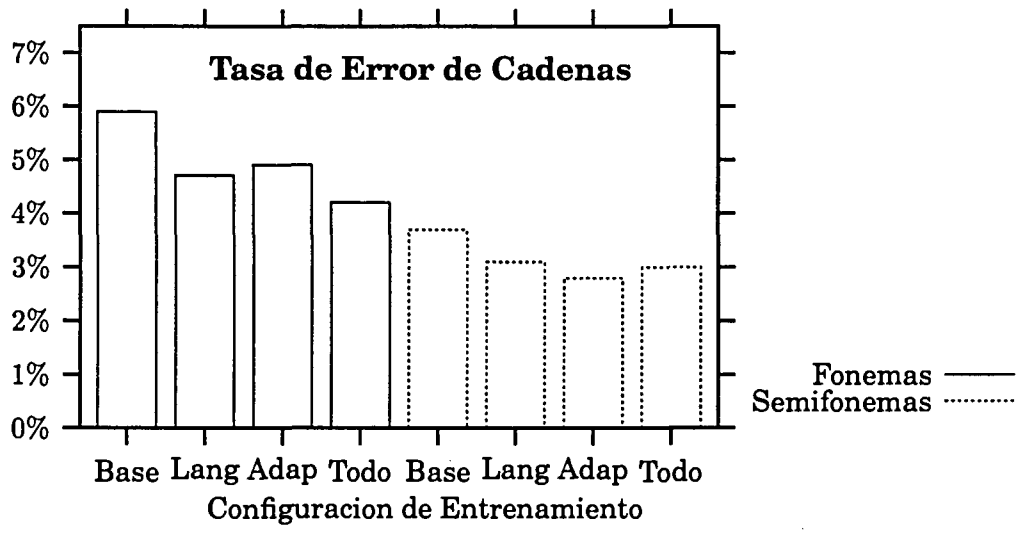

Figura 2.1: *Tanto por ciento de cadenas erróneas en el reconocimiento de TIDIGITS empleando modelos de fonema y semifonema entrenados con TIMIT.*

Configuración	Error	Sust	Inse	Borr	Acierto	Corr
BaseFon	2,10	1,02	0,55	0,53	98,44	94,1
LangFon	1,69	0,88	0,38	0,43	98,69	95,3
AdapFon	1,76	0,92	0,39	0,45	98,63	95,1
TodoFon	1,48	0,68	0,39	0,41	98,91	95,8
BaseSefo	1,27	0,45	0,41	0,41	99,14	96,3
LangSefo	1,10	0,41	0,35	0,34	99,24	96,9
AdapSefo	0,96	0,35	0,31	0,29	99,36	97,2
TodoSefo	1,05	0,38	0,35	0,32	99,30	97,0

Tabla 2.8: *Resultados obtenidos en el reconocimiento de TIDIGITS empleando modelos de fonema y semifonema entrenados con el corpus completo TIMIT.*

3. La adaptación a la tarea puede realizarse sobre la base de los modelos entrenados según el criterio de mínima confusibilidad independiente de la tarea pero, aunque se consigue acumular los beneficios de una y otro, el resultado final no supera el obtenido adaptando directamente los modelos de máxima verosimilitud, AdapSefo.

4. El experimento AdapSefo proporciona los mejores resultados obtenidos hasta la fecha en esta tarea utilizando como sistema de entrenamiento/reconocimiento RAMSES.

Es importante hacer notar una aparente contradicción en el comportamiento de los fonemas y los semifonemas frente a la adaptación al idioma y a la tarea: en tanto el fonema se beneficia en mayor medida de la adaptación al idioma, el semifonema lo hace al revés. El origen de este comportamiento puede achacarse al distinto aprovechamiento del material de entrenamiento disponible en uno y otro caso. Así, en adaptación al idioma, todo el material de entrenamiento es aprovechable ya que todas las frases (se supone que)

son válidas en el idioma. Ahora bien, el número de unidades y parámetros a reestimar es muy superior en el caso de los semifonemas, con lo que el aprovechamiento del material disponible es muy inferior al de los fonemas. Sin embargo, al introducir las restricciones propias de la tarea de las cadenas de dígito sólo las cadenas de posible aparición participan en el entrenamiento. Éstas son, en esencia, las mismas para fonemas y semifonemas —y, con el suavizado explicado más arriba, incluso más para estos últimos—. Por otro lado, el número de modelos y parámetros a reestimar de manera efectiva cuando se introducen las restricciones propias de la tarea se reduce mucho más para los semifonemas que para los fonemas, ya que sólo se reestiman las unidades presentes en la tarea, hasta casi igualarse —20 fonemas frente a 61 semifonemas (equivalentes a 30 fonemas y medio)—. Por tanto, la pérdida de material efectivamente aprovechado en el entrenamiento es mucho má severa en el caso de los modelos de fonema, que en el de los de semifonema.

2.4.2.2 Decodificación acústico fonética independiente del locutor utilizando modelos de semifonema

Otro experimento igualmente significativo es el de la decodificación acústico fonética. Aunque a lo largo de esta tesis se ha hecho hincapié en la aplicación de los modelos acústicos obtenidos al reconocimiento de tareas reales de habla continua, una vez ilustrada la utilidad de los planteamientos propuestos —la minimización de la confusibilidad estimada en segmentos acústicos de longitud limitada—, el mantenimiento de la restricción a la tarea de las cadenas de dígito presenta el inconveniente de su poca representatividad. Esto es especialmente grave en el caso de utilizarse modelos de semifonema, ya que sólo la sexta parte de los mismos interviene en la tarea. En estas condiciones la influencia del modelo del lenguaje aumenta frente a la del modelado acústico y las interacciones entre ambas pueden provocar que una mejora en este último, sin atender a las características del primero, no resulte en el beneficio esperado. La decodificación acústico fonética, aunque no puede catalogarse como una *auténtica* tarea de reconocimiento del habla continua, sí está libre de esta dependencia del modelo del lenguaje. Además, la tasa de error de fonemas —aunque no, por ahora, la de frases— en decodificación acústico fonética proporciona una medida razonable de la capacidad de discriminación de los modelos acústicos en cualquier condición.

En experimentos de decodificación acústico fonética sólo tienen sentido comparar las prestaciones de las configuraciones de entrenamiento independientes de la tarea —experimentos **BaseFon**, **LangFon**, **BaseSefo** y **LangSefo**—. La tabla 2.9 muestra los resultados obtenidos con estas cuatro configuraciones.

A la vista de los resultados puede concluirse el elevado beneficio obtenido tanto al aplicar entrenamiento discriminativo o explicitación de los contextos separadamente, como al entrenar discriminativamente los modelos acústicos de unidades deopendientes del contexto —semifonemas, en este caso—. Una primera conclusión de los resultados es que, aunque tanto **LangFon** como **BaseSefo** permiten reducir notablemente la tasa de error de **BaseFon**, la diferenciación de contextos es más benefiosa que el entrenamiento discriminativo —una disminución del 15% frente a una del 12%—. No obstante, la diferencia de prestaciones es relativamente pequeña, y el coste a pagar es un incremento notable de la complejidad (número de unidades) del sistema. Finalmente, la aplicación de entrenamiento discriminativo a los modelos de semifonema **LangSefo** ha resultado en una disminución adicional del error del 8,5% sobre los modelos de semifonemas entrenados

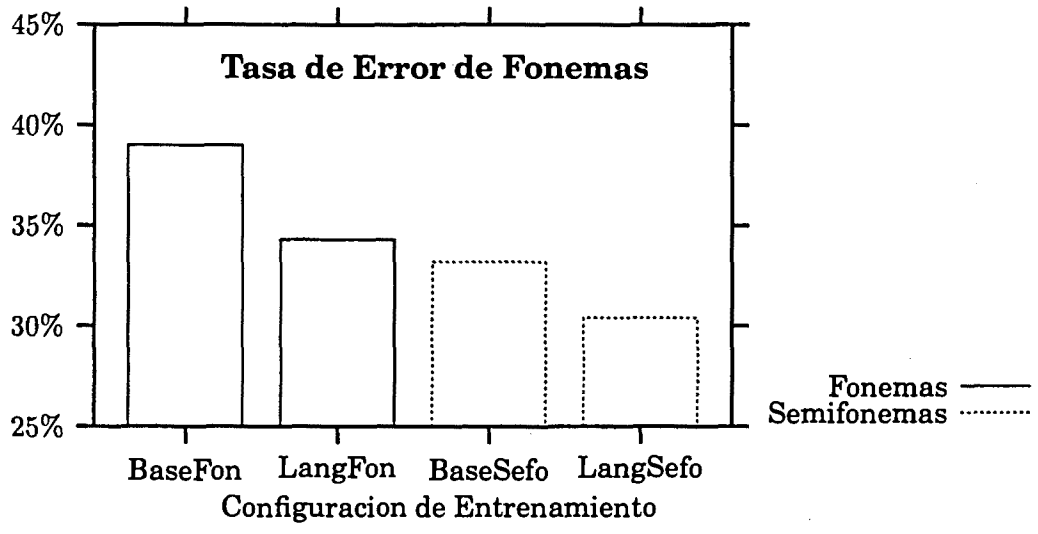

Figura 2.2: *Tasa de error de fonemas en decodificación acústico fonética independiente del locutor de TIMIT utilizando modelos de fonema y semifonema. Nótese que el origen de la escala no es cero.*

Configuración	Error	Sust	Inse	Borr	Acierto
BaseFon	39.0	23.8	7.9	8.1	68.1
LangFon	34.3	21.0	6.6	6.8	72.2
BaseSefo	33.2	20.5	6.3	6.5	73.0
LangSefo	30.4	18.6	5.8	6.0	75.4

Tabla 2.9: *Resultados obtenidos en decodificación acústico fonética independiente del locutor de TIMIT empleando modelos de fonema y semifonema entrenados tanto con Baum-Welch como con EMC. Tanto en el caso de los fonemas, como en el de los semifonemas, el entrenamiento de mínima confusibilidad independiente de la tarea proporciona un beneficio muy importante.*

con máxima verosimilitud **BaseSefo**. Proporcionando el mejor resultado de todas las configuraciones, una tasa de error de fonemas del 30.4%. Este resultado es uno de los más bajos publicados en esta tarea [57, 44, 24, 40], con la dificultad añadida de que la transcripción adoptada en estos experimentos, proporcionando mayor precisión en la caracterización fonética, también hace empeorar *aparentemente* la tasa de decodificación (véase el apartado 1.1.2).

Capítulo 3

Optimización de la Función de Coste: Algoritmo de Búsqueda Adaptativa de Gradiente

Tanto el entrenamiento de máxima verosimilitud como los de mínimo error de clasificación, máxima información mutua y mínima confusibilidad, consisten en la formulación de una función de coste o calidad del sistema y la posterior reestimación de los parámetros de manera que esta función alcance un óptimo, aunque sólo sea local. El proceso de entrenamiento puede considerarse, en todos estos casos, como un problema de optimización de una función de múltiples variables. Ahora bien, mientras en entrenamiento de máxima verosimilitud las características de linealidad y separabilidad de variables posibilitan el empleo de un algoritmo eficiente y seguro de optimización, el de Baum-Welch; en entrenamiento discriminativo la situación es radicalmente distinta. Así, las funciones de coste empleadas acumulan buena parte de las situaciones que dificultan un problema de optimización:

1. Fuerte acoplamiento[1] entre parámetros distintos, debido a que la función de coste se construye como una suma de contribuciones en cada una de las cuales están involucradas dos o más unidades acústicas distintas.

2. No linealidad, necesaria para poder reflejar la regla de decisión empleada en el reconocimiento.

3. Un número muy elevado de parámetros: del orden de 150.000, en la experimentación realizada utilizando modelos acústicos de fonema, y aún más utilizando unidades dependientes del contexto.

La combinación de estas tres situaciones conforma un problema de optimización especialmente difícil de manejar ya que imposibilita, o hace poco fiable, la aplicación de técnicas convencionales de optimización [26, 32, 25, 62]. Por ejemplo, el hessiano de estas funciones no sólo es inabordable desde un punto de vista computacional sino que, además, no permite realizar ninguna presunción fiable acerca de su forma —sabemos que no es diagonal ni constante y, en general, no será ni definido positivo ni negativo—. En casos de este tipo, la

[1]Se entiende como medida de este acoplamiento el valor absoluto de las derivadas cruzadas de la función de coste.

alternativa habitualmente empleada consiste en acudir a algún algoritmo de optimización basado en búsqueda de gradiente. Este tipo de búsqueda garantiza que, si la función a optimizar es continua y diferenciable —y, por construcción, tanto la información mutua, como la función de cómputo de error de clasificación, como la confusibilidad lo son—, el valor de la función de coste disminuye en cada iteración hasta alcanzar un mínimo, al menos local 3.1.

Siendo $\mathcal{L}(X, \Lambda_t)$ la función de coste del sistema Λ_t sobre el conjunto de realizaciones de entrenamiento X, en lo que queda de desarrollo se hará referencia a la función de múltiples variables $\mathcal{G}(\Lambda_t) = \mathcal{L}(X, \Lambda_t)$ y se considerará que el problema de optimización se refiere a su minimización. Con ello se pretende compactar la notación —eliminando la dependencia en el material de entrenamiento, X, que suponemos prefijado—, y poner de manifiesto que no se echa mano de las caraterísticas concretas de las funciones de coste tratadas en esta tesis —las utilizadas en entrenamiento discriminativo de modelos ocultos de Markov—, por lo que las exposiciones y desarrollos que siguen son aplicables a cualquier problema de optimización de funciones de múltiples variables.

3.1 Búsqueda de Gradiente en Optimización de Funciones de Múltiples Variables

La búsqueda de gradiente es uno de los algoritmos más ampliamente empleados en la minimización de funciones de múltiples varibles. Los parámetros del sistema son actualizados según la fórmula:

$$\Lambda_{t+1} = \Lambda_t - \varepsilon_t \nabla_\Lambda \mathcal{G}(\Lambda_t) \tag{3.1}$$

La cual asegura que $\mathcal{G}(\Lambda_{t+1}) \leq \mathcal{G}(\Lambda_t)$, siempre que $\mathcal{G}(\Lambda)$ sea continua y ε_t una cantidad positiva y sufientemente pequeña para garantizar que la aproximación de primer orden es válida. En su realización más sencilla, el paso de aprendizaje, ε_t, se mantiene constante durante el proceso completo de optimización, $\varepsilon_t = \varepsilon_0$. Se hará referencia a este algoritmo con el nombre de *búsqueda de gradiente de paso de aprendizaje fijo* o, simplemente, *búsqueda de gradiente* (en adelante, **GD**, por las siglas en inglés de *gradient descent*).

El principal inconveniente del algoritmo de búsqueda de gradiente radica en su fuerte dependencia del paso de aprendizaje ε_0. Si su valor es excesivamente pequeño, el algoritmo siempre convergerá hacia un punto crítico, pero el número de iteraciones empleadas en alcanzarlo puede ser muy elevado. Si el valor elegido es muy grande, la aproximación de primer orden no será válida y el algoritmo puede diverger. Como consecuencia, cada vez que se plantea un experimento distinto de entrenamiento discriminativo, el valor de ε_0 debe ser optimizado para garantizar consistencia en la convergencia, con el agravante de que el valor idóneo de ε_0 es muy sensible a parámetros tales como el tamaño de la población de entrenamiento, el número de hipótesis considerado, la longitud de los segmentos, etc. Así, si queremos estudiar el comportamiento del sistema frente a variaciones en alguna de estas magnitudes, no sólo deberemos plantear una batería de experimentos que abarque todos sus posibles valores, sino que para cada uno de ellos deberemos realizar una búsqueda adicional sobre el valor del paso de aprendizaje.

Para ilustrar la dependencia de la búsqueda de gradiente respecto a la elección del paso de aprendizaje, se realizó una batería de experimentos en la cual sólo este valor es variado. Cada experimento de la batería consiste en la minimización de la confusibilidad

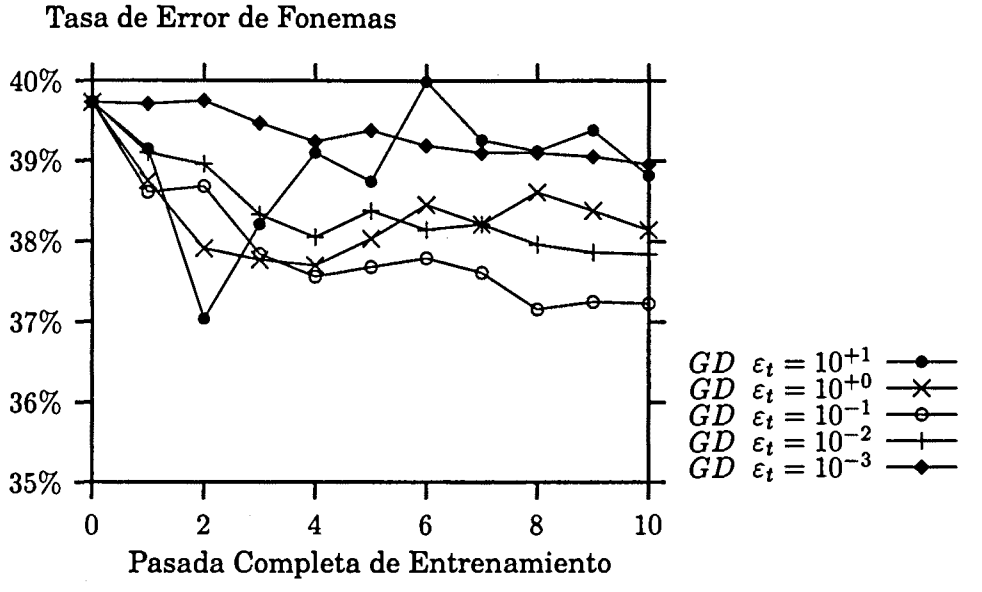

Figura 3.1: *Evolución de la tasa de error en DAF independiente del locutor de TIMIT usando búsqueda de gradiente con distintos pasos de aprendizaje. Se observa que un valor adecuado del paso de aprendizaje,* $\varepsilon_t = \varepsilon_0 \approx 10^{-1}$*, permite alcanzar la convergencia en un número pequeño de iteraciones. También se observan tanto los efectos de un valor excesivo,* $\varepsilon_t = \varepsilon_0 = 10^{+1}$*, apareciendo comportamientos erráticos y llegando a empeorar el resultado inicial; como los de uno insuficiente,* $\varepsilon_t = \varepsilon_0 = 10^{-3}$*, para el cual la velocidad de convergencia es excesivamente lenta.*

independendiente de la tarea utilizando búsqueda de gradiente. Los modelos de Markov son entrenados con la parte masculina del corpus **train** de TIMIT. Los modelos usados como inicialización son los resultantes de aplicar entrenamiento de máxima verosimilitud. Como medida de comparación del resultado se proporciona la tasa de error en decodificación acústico fonética independiente del locutor. La elección de esta medida es discutible ya que sólo refleja la convergencia del algoritmo de manera indirecta. No obstante, la alternativa de ilustrar la convergencia con el valor de la propia función de coste empleada, la confusibilidad, no es realmente interesante debido a que su valor depende de factores diversos y cambiantes a medida que avanza el entrenamiento. Así, y aunque siempre se produce una fuerte disminución de su valor como resultado del proceso de entrenamiento, la comparación entre experimentos distintos no siempre es representativa. Por otro lado, la elección de la tasa de error dependiente del locutor, es decir: reconociendo las frases que participan en el entrenamiento, también se descartó debido a que, aunque también reflejaría mejor el comportamiento del algoritmo de optimización, no es tan representativa de las verdaderas prestaciones del sistema de reconocimiento en independencia del locutor. Además, el número de experimentos realizados es muy elevado y el hecho de reconocer la base de datos de entrenamiento —de tamaño muy superior al corpus de **test**— hubiera resultado excesivamente costoso. Finalmente, el principal problema que hubiera podido aparecer tomando una medida de la convergencia tan alejada de la función optimizada, sería que ésta no reflejara suficientemente las propiedades del algoritmo propuesto, pero, aparentemente, esta situación no se produce (ver las figuras 3.1-3.9).

La figura 3.1 muestra la evolución de la tasa de error conforme avanza el algoritmo de búsqueda de gradiente para cinco valores distintos del paso de aprendizaje. Los cinco

valores abarcan tanto valores cercanos al mejor, $\varepsilon_t = \varepsilon_0 = 10^{-1}$ (determinado previamente mediante tanteo), así como este valor multiplicado y dividido por diez y por cien. De este modo, quedan de manifiesto tanto los efectos de un valor excesivo del paso de aprendizaje, como de uno escaso. A la vista de la gráfica, se comprueba que el comportamiento de la búsqueda de gradiente es el esperado: si el valor de ε_t es excesivo —en el ejemplo, esto ocurre para $\varepsilon_t = 10^{+1}$ y, en menor medida, para $\varepsilon_t = 10^{+0}$—, el algoritmo presenta un comportamiento errático, llegando incluso a empeorar el resultado obtenido con los modelos entrenados aplicando el criterio de máxima verosimilitud. Por contra, si el valor del paso de aprendizaje es escaso —en el ejemplo, para $\varepsilon_t = 10^{-3}$—, aunque el comportamiento del algoritmo es estable, el resultado obtenido al cabo de 10 pasadas por el material de entrenamiento está muy alejado del que se puede conseguir utilizando un valor apropiado. Por otro lado, los dos valores que mejor comportamiento presentan —$\varepsilon_t = 10^{-1}$ y $\varepsilon_t = 10^{-2}$—, dan lugar a resultados claramente distintos entre ellos, con una diferencia cercana a un punto porcentual al final de las 10 pasadas.

3.2 Algoritmo de Búsqueda Adaptativa de Gradiente

Un modo de evitar las dificultades existentes en la elección del paso de aprendizaje consiste en el uso del método *steepest descent* [21, 62]. Este método es uno de los más empleados en la minimización de funciones no lineales de múltiples variables. Su base de funcionamiento consiste en aplicar (3.1) con una sucesión de valores del paso de aprendizaje definida mediante:

$$\varepsilon_t = \arg \min_{0 \leq \varepsilon \leq \infty} \mathcal{G}(\Lambda_t - \varepsilon \nabla \mathcal{G}(\Lambda_t)) \tag{3.2}$$

Esta expresión no sólo elimina la ambigüedad en la elección del paso de aprendizaje, sino que, además, presenta mejores propiedades de convergencia que la búsqueda de gradiente con paso de aprendizaje prefijado. El precio a pagar es la necesidad de determinar el valor de ε_t que satisface (3.2). En principio, se debería realizar una búsqueda lineal sobre ε en cada iteración de entrenamiento, pero el coste computacional de cada búsqueda lineal puede ser tan cara, o más, que el cómputo del gradiente, así que el coste total puede ser muy superior a simplemente optimizar ε_0 en búsqueda de gradiente con paso fijo. Sin embargo, y contrariamente a otros métodos más sofisticados tales que el de Davidon-Fletcher-Powell [26], el *steepest descent* no depende en demasía de la exactitud en la búsqueda lineal de ε_t [62]. Esto se debe a que *steepest descent* es un algoritmo de búsqueda de gradiente *puro*, es decir: en cada instante t, la dirección de actualización de los parámetros es exactamente la del gradiente (estimado), la cual, a su vez, es la dirección que garantiza la máxima mejora en la función de coste con una menor perturbación de los parámetros del sistema. Incluso si no se consigue satisfacer (3.2) en cada iteración, una aproximación suficientemente buena de la exacta conduce a una solución, como mínimo, tan buena como si el paso de aprendizaje fuera tan pequeño como para garantizar el cumplimiento de la aproximación de primer orden en la búsqueda de gradiente. De hecho, es cuestionable la necesidad de una solución exacta a (3.2). Si bien garantizaría buenas propiedades de convergencia para el caso de funciones cuadráticas, las funciones optimizadas en entrenamiento discriminativo no responden a esta forma. Además, la dirección de optimización de los parámetros sólo es una aproximación —más o menos exacta, en función de las restricciones computacionales— del valor real del gradiente.

Esta flexibilidad en la búsqueda lineal sobre ε permite la propuesta de un algoritmo muy sencillo para estimar valores apropiados del paso de aprendizaje: el algoritmo de búsqueda adaptativa de gradiente (**BAG**). Este algoritmo se basa en la adaptación sucesiva de los valores de ε_t suponiendo un comportamiento localmente cuadrático de la función de coste. El objetivo de la aproximación cuadrática no es tan sólo acceder al óptimo en una sola iteración —tal y como plantea *steepest descent*— sino también rectificar el valor utilizado del paso de aprendizaje de manera que, si efectivamente la función tiene un comportamiento cuadrático, se acceda al óptimo en un sólo paso, pero si el comportamiento de la función no lo es, obtengamos como mínimo una buena elección del valor del paso de aprendizaje para la iteración siguiente.

Supongamos que $\mathcal{G}(\Lambda)$ es una función cuadrática con hessiano definido positivo, $\mathcal{H}(\mathcal{G}(\Lambda)) = Q$:

$$\mathcal{G}(\Lambda) = \Lambda^T Q \Lambda - \Lambda^T b + c. \tag{3.3}$$

En este caso, la búsqueda lineal en la dirección del gradiente se reduce a la localización del mínimo de una parábola, dado que la proyección de una forma cuadrática sobre cualquier dirección del espacio lo es. Una manera adecuada de resolver la minimización de una parábola es el método de la *falsa posición* [111, 49, 62]. La base de funcionamiento es el hecho que la derivada de una parábola es una línea recta, correspondiendo la posición del mínimo al punto donde esta línea corta el eje de abcisas. De este modo, el conocimiento del valor de la derivada en dos puntos distintos basta para conocer el punto donde la función se minimiza.

Consideremos ahora, manteniendo la aproximación cuadrática de hessiano definido positivo, cómo se puede alcanzar el valor Λ_t^* que optimiza $\mathcal{G}(\Lambda)$ a lo largo de la línea definida por

$$\Lambda(\mu) = \Lambda_t - \mu \nabla \mathcal{G}(\Lambda_t) \tag{3.4}$$

donde μ es un valor real y postivo. Para ello definimos la función de μ:

$$\mathcal{F}(\mu) = \mathcal{G}(\Lambda - \mu \nabla \mathcal{G}(\Lambda)) \,|_{\Lambda=\Lambda_t} \tag{3.5}$$

Dado el carácter cuadrático de $\mathcal{G}$, $\mathcal{F}(\mu)$ es una función parabólica, por lo tanto su segunda derivada respecto a μ es constante, pudiéndose calcular la posición del único mínimo a partir del conocimiento de la primera derivada en dos puntos cualesquiera, $\mathcal{F}'(\mu_0)$ y $\mathcal{F}'(\mu_1)$:

$$\mu^* = \frac{\mu_0 \mathcal{F}'(\mu_1) - \mu_1 \mathcal{F}'(\mu_0)}{\mathcal{F}'(\mu_1) - \mathcal{F}'(\mu_0)} \tag{3.6}$$

Tomando los puntos $\mu_0 = 0$ y $\mu_1 = \mu_t$ y definiendo $\Lambda_t' = \Lambda_t - \mu_t \nabla \mathcal{G}(\Lambda_t)$, μ^* vale:

$$\mu^* = \mu_t \frac{\delta \mathcal{G}(\Lambda)/\delta \mu \,|_{\Lambda=\Lambda_t}}{\delta \mathcal{G}(\Lambda)/\delta \mu \,|_{\Lambda=\Lambda_t'} - \delta \mathcal{G}(\Lambda)/\delta \mu \,|_{\Lambda=\Lambda_t}} \tag{3.7}$$

Definiendo $\mathbf{g}_t = -\nabla \mathcal{G}(\Lambda)|_{\Lambda=\Lambda_t}$, $\Lambda_t' = \Lambda_t + \mu_t \mathbf{g}_t$, y $\mathbf{g}_t' = -\nabla \mathcal{G}(\Lambda)|_{\Lambda=\Lambda_t'}$,

$$\left.\frac{\delta \mathcal{G}(\Lambda)}{\delta \mu}\right|_{\Lambda=\Lambda_t} = \nabla \mathcal{G}(\Lambda)|_{\Lambda=\Lambda_t} \cdot \frac{\delta \Lambda_t'}{\delta \mu} = -\mathbf{g}_t^T \mathbf{g}_t \tag{3.8}$$

$$\left.\frac{\delta \mathcal{G}(\Lambda)}{\delta \mu}\right|_{\Lambda=\Lambda_t'} = \nabla \mathcal{G}(\Lambda)|_{\Lambda=\Lambda_t'} \cdot \frac{\delta \Lambda_t'}{\delta \mu} = -\mathbf{g}_t'^T \mathbf{g}_t \tag{3.9}$$

(3.10)

Con lo que el valor de μ^* vale,

$$\mu^* = \mu_t \frac{\mathbf{g}_t^T \mathbf{g}_t}{\mathbf{g}_t'^T \mathbf{g}_t - \mathbf{g}_t^T \mathbf{g}_t} \tag{3.11}$$

Alcánzandose el óptimo de la función para Λ_t^*,

$$\begin{aligned} \Lambda_t^* &= \Lambda_t + \mu^* \mathbf{g}_t \\ &= \Lambda_t - \mu_t \frac{\mathbf{g}_t^T \mathbf{g}_t}{\mathbf{g}_t^T \mathbf{g}_t - \mathbf{g}_t'^T \mathbf{g}_t} \mathbf{g}_t \\ &= \Lambda_t - \varepsilon_t^* \mathbf{g}_t \end{aligned} \tag{3.12}$$

$$\varepsilon_t^* = \mu_t \frac{\mathbf{g}_t^T \mathbf{g}_t}{\mathbf{g}_t^T \mathbf{g}_t - \mathbf{g}_t'^T \mathbf{g}_t} \tag{3.13}$$

Donde ε_t^* es el paso de aprendizaje óptimo que conduce a Λ_t^* aplicando la ecuación 3.1. Hay dos características destacables de (3.13):

1. Conduce a Λ_t^* con independencia de μ_t, así que una posible elección para su valor es el del paso de aprendizaje utilizado en la iteración anterior, ε_{t-1}.

2. Debido a la aproximación parabólica, el punto alcanzado, Λ_t^*, es el mismo que si aplicamos (3.1) a Λ_t' sustituyendo el gradiente en este punto, $\mathbf{g}_t'$ por su proyección sobre el gradiente en la iteración anterior, $\mathbf{g}_t$ (para la demostración, intercámbiense los puntos en la ecuación 3.6, haciendo $\mu_0 = \mu_t$ y $\mu_1 = 0$).

Estas dos propiedades permiten un atajo en la aplicación de *steepest descent*: nos movemos desde Λ_t a Λ_{t+1} utilizando el valor de ε_t que hubiera llevado a Λ_{t-1}^* en el caso de haber sido aplicado sólo a la componente del gradiente de Λ_t paralela al gradiente en Λ_{t-1}. No obstante, en lugar de proceder de este modo y acceder a Λ_{t-1}^* —tal y como requeriría *steepest descent*—, aplicamos este paso de aprendizaje tanto a la componente paralela como a la perpendicular. No se puede garantizar el cumplimiento de (3.2) en ningún instante t, pero es equivalente a acceder a Λ_{t-1}^* —y, de este modo, optimizar la búsqueda lineal para el instante anterior— y emplear este valor del paso de aprendizaje como una aproximación del valor óptimo en la componente perpendicular —en cierto modo, es equivalente a realizar una búsqueda lineal adicional; nótese que $\Lambda_t = \Lambda_{t-1}'$ para todo t. En cualquier caso, el algoritmo garantiza que cualquier movimiento realizado durante la búsqueda iterativa será corregido en la siguiente iteración. Como resultado final, el paso de aprendizaje en el instante t, ε_t, es el mismo que el usado en el instante anterior multiplicado por una factor de actualización, $\hat{\varepsilon}_t$:

$$\left.\begin{aligned} \mathbf{g}_t &= -\nabla \mathcal{G}(\Lambda_t) \\ \hat{\varepsilon}_t &= \frac{\mathbf{g}_{t-1}^T \mathbf{g}_{t-1}}{\mathbf{g}_{t-1}^T \mathbf{g}_{t-1} - \mathbf{g}_t^T \mathbf{g}_{t-1}} \\ \varepsilon_t &= \hat{\varepsilon}_t \varepsilon_{t-1} \\ \Lambda_{t+1} &= \Lambda_t - \varepsilon_t \mathbf{g}_t \end{aligned}\right\} \text{ para } t > 0 \tag{3.14}$$

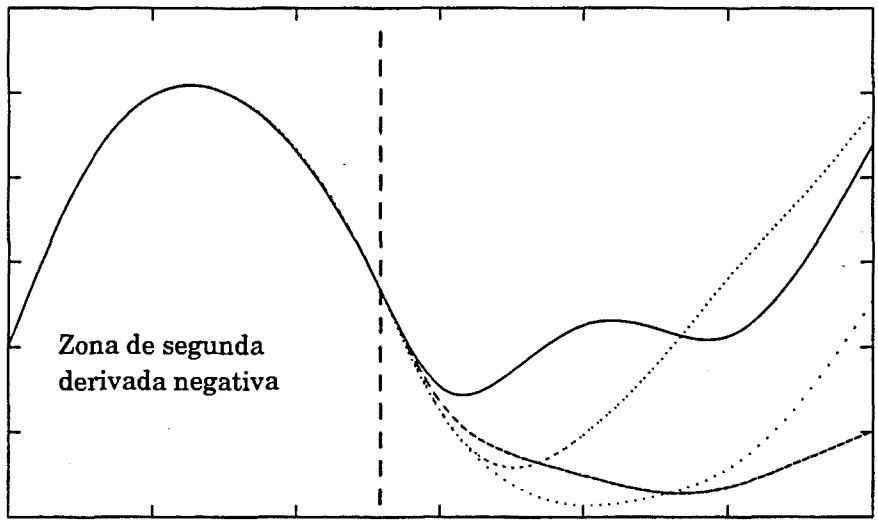

Figura 3.2: *Ejemplo de distintas funciones que comparten valores en una región de segunda derivada negativa. Aunque el conocimiento de la segunda derivada en esa zona permitiría localizar el máximo de la función utilizando la aproximación parabólica, el comportamiento fuera de ella es impredecible, y cada una de las cuatro funciones dibujadas presenta el mínimo (buscado) en un sitio distinto.*

Si la función a minimizar es una forma cuadrática de hessiano igual a una constante multiplicada por la identidad, $\mathcal{H} = kI$, la aplicación de (3.14) es equivalente a la resolución análitica exacta utilizando la inversa del hessiano[2]. Dado que éste no será, en general, el caso, el método tenderá a estimar una aproximación de este tipo, en la cual k es una aproximación del valor que toma la segunda derivada de la función de coste calculada en cada punto en la dirección de su gradiente.

3.2.1 Positividad del hessiano

La ecuación (3.14) sólo es válida para funciones con hessiano definido positivo. Si éste no es el caso, y en general no lo será, entonces la segunda derivada a lo largo de cualquier dirección del espacio —incluyendo, por tanto, la del gradiente— puede ser negativa. En *steepest descent*, la positividad o no del hessiano no es una cuestión de tanta gravedad como en el caso general de optimización de funciones de múltiples variables, ya que sólo es utilizado en una búsqueda lineal. Un valor negativo de la segunda derivada simplemente indica la presencia de un máximo, en lugar del deseado mínimo, en la aproximación de segundo órden. Por tanto, el conocimiento de la segunda derivada —el hessiano, en problemas de múltiples variables— no resulta de ninguna utilidad en problemas de minimización si su valor no es positivo —hessiano definido positivo, en múltiples variables—. La figura 3.2 muestra la evolución de cuatro funciones diferentes, pero que comparten valores en una región en la que la segunda derivada es negativa. Cualquier aproximación parabólica realizada dentro de esta región permite localizar con

[2]Téngase en cuenta que si el hessiano es igual a la identidad por una constante, la segunda derivada de la función en cualquier dirección del espacio igual a esta constante. Así pues, la ecuación 3.14 proporciona la solución analítica exacta en una sola iteración.

precisión la posición del máximo presente en ella. Sin embargo, el comportamiento fuera de la región de segunda derivada negativa es impredecible, y cada una de las funciones presenta el mínimo en un sitio distinto.

La no positividad de la segunda derivada implica que la primera es decreciente. Esto es: $\delta\mathcal{G}(\Lambda)/\delta\mu|_{\Lambda=\Lambda_{t-1}} > \delta\mathcal{G}(\Lambda)/\delta\mu|_{\Lambda=\Lambda_t}$, o, lo que es lo mismo: $\mathbf{g}_t^T\mathbf{g}_{t-1}/\mathbf{g}_{t-1}^T\mathbf{g}_{t-1} > 1$. De hecho, los problemas con el valor de $\mathbf{g}_t^T\mathbf{g}_{t-1}/\mathbf{g}_{t-1}^T\mathbf{g}_{t-1}$ empiezan a partir del momento en que su valor se aproxima a uno. Esto es así porque, mientras para valores negativos, el método de la falsa posición es equivalente a una interpolación, si su valor es positivo, se convierte en una extrapolación. A medida que su valor se acerca a uno, $\hat{\varepsilon}_{t-1}$ crece hasta valer infinito, haciéndose negativo por encima de este valor. Esta situación representa que la aproximación pasa de presentar un mínimo a presentar un máximo. En la figura 3.3 se muestra la evolución de $\hat{\varepsilon}_{t+1}$ en función del valor de $\mathbf{g}_t^T\mathbf{g}_{t-1}/\mathbf{g}_{t-1}^T\mathbf{g}_{t-1}$, así como las tres situaciones que se pueden dar al aplicar la aproximación parabólica propuesta —téngase en cuenta que el movimiento entre Λ_{t-1} y Λ_t es siempre en la dirección de decrecimiento de la función—.

La ecuación (3.14) sólo es válida para funciones con hessiano definido positivo. A pesar de la inutilidad de la segunda derivada si ésta es negativa, es importante tener en cuenta que la dirección del gradiente es siempre la de máximo beneficio. Es decir, en ausencia de más información, lo mejor que se puede hacer es avanzar en la dirección del gradiente, dado que es la que mayor decremento de la función a minimizar garantiza con la menor perturbación de los parámetros del sistema. Un valor positivo de $\mathbf{g}_t^T\mathbf{g}_{t-1}/\mathbf{g}_{t-1}^T\mathbf{g}_{t-1}$ implica que el producto escalar del gradiente calculado en el instante $t-1$ y en el t también es positivo. Por tanto, la componente de $\mathbf{g}_t$ paralela a $\mathbf{g}_{t-1}$ apunta en la misma dirección que este último. Es decir, el movimiento realizado en $t-1$ es reiterado en t. Además, si la función a minimizar no presenta ningún punto crítico en el intervalo,

$$\{\Lambda(\mu) = \Lambda_{t-1} - \mu\nabla\mathcal{G}(\Lambda_{t-1}) \mid 0 \leq \mu \leq \varepsilon_{t-1}\} \tag{3.15}$$

existe, al menos, un punto en la dirección del gradiente en $t-1$, pero más alejado de Λ_{t-1} que Λ_t, tal que el valor de la función de coste es inferior al de esta última. Por tanto, parece razonable que un valor mayor del paso de aprendizaje en el instante $t-1$ podría haber bastado para acceder a ese mismo punto, pero en una iteración menos. Por ejemplo, si dos iteraciones consecutivas son idénticas, multiplicar el paso de aprendizaje por dos equivale a realizar el mismo movimiento en una sola iteración. (Nótese que, en este caso y por contra, la aproximación parabólica obligaría a hacer su valor infinito). Si el movimiento conjunto de dos iteraciones consecutivas es realizado en una única iteración, la convergencia no se ve afectada —al menos de manera negativa, dado que el punto al que accedemos es el mismo— pero la velocidad de convergencia aumenta. Esto conduce a una modificación a (3.14), en la que el valor de ε_t se calcula de manera distinta cuando $\mathbf{g}_t^T\mathbf{g}_{t-1}/\mathbf{g}_{t-1}^T\mathbf{g}_{t-1} \geq a$ siendo $0 \leq a < 1$:

$$\hat{\varepsilon}_t = 1 + (1-a)^{-1}\frac{\mathbf{g}_t^T\mathbf{g}_{t-1}}{\mathbf{g}_{t-1}^T\mathbf{g}_{t-1}} \quad \text{para} \quad \frac{\mathbf{g}_t^T\mathbf{g}_{t-1}}{\mathbf{g}_{t-1}^T\mathbf{g}_{t-1}} \geq a \tag{3.16}$$

Esta expresión garantiza que $\hat{\varepsilon}(\mathbf{g}_t, \mathbf{g}_{t-1})$ es una función continua y de primera derivada igualmente continua (ver la figura 3.4. Si a es igual a cero, el algoritmo descarta cualquier intento de extrapolación utilizando la aproximación parabólica, y actualiza el paso de aprendizaje de manera que el movimiento coincidente de las dos últimas iteraciones se

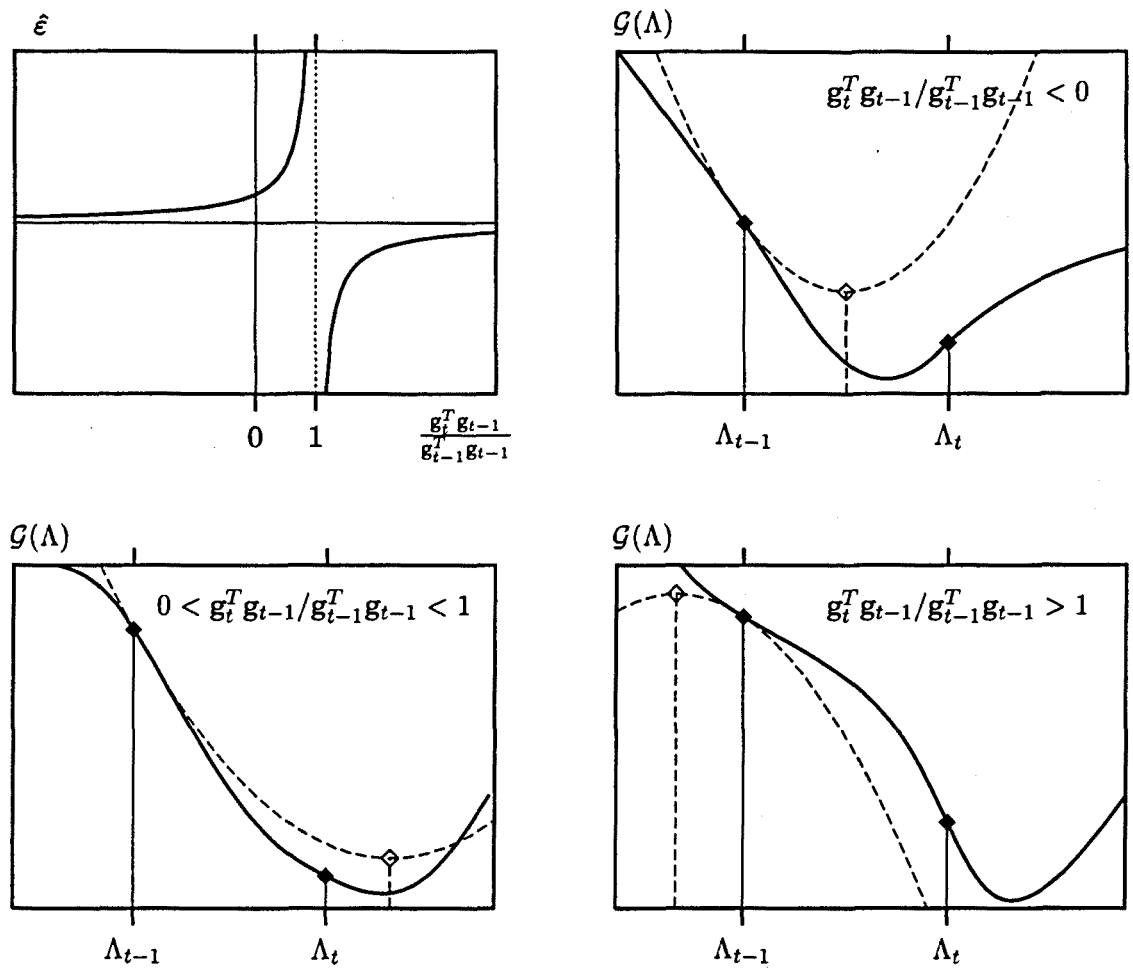

Figura 3.3: *Evolución de $\hat{\epsilon}$ en función del valor de $g_t^T g_{t-1}/g_{t-1}^T g_{t-1}$, y ejemplos de las diferentes situaciones que se pueden dar en la aproximación parabólica. En la esquina superior izquierda se muestra el valor de $\hat{\epsilon}$ en función de ese cociente. A continuación, y siguiendo de izquierda a derecha y de arriba a abajo, se muestran ejemplos en los que la aproximación parabólica —indicada con trazo discontinuo— implica una interpolación a un mínimo, una extrapolación a un mínimo, y una extrapolación a un máximo (la interpolación a un máximo no es posible debido que Λ_t se determina a partir de Λ_{t-1} avanzando en la dirección contraria a su gradiente).*

hubiera realizado en la primera de ellas. Un valor superior de a incluye en la aproximación parabólica situaciones de extrapolación, haciendo que el paso de aprendizaje crezca más rápidamente de lo necesario para realizar *dos iteraciones en una*. Esta mayor velocidad de crecimiento del paso de aprendizaje ha demostrado ser de gran utilidad en los experimentos realizados, habiéndose utilizado valores en torno a 0,75.

3.2.2 Interpretación adaptativa

El algoritmo presentado más arriba admite tres interpretaciones distintas, en absoluto excluyentes: en primer lugar, se realiza una aproximación muy grosera de la inversa del

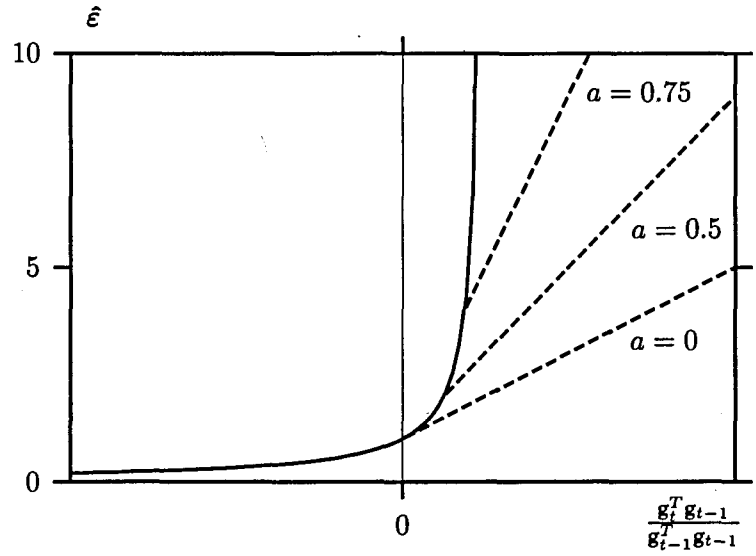

Figura 3.4: *Factor de actualización del paso de aprendizaje $\hat{e}$, en función del cociente $g_t^T g_{t-1}/g_{t-1}^T g_{t-1}$, utilizado en el algoritmo BAG. La función se forma a partir de la concatenación de la linea de trazo continuo —correspondiente a la zona donde se aplica la aproximación parabólica— con la de trazo discontinuo correspondiente al valor de a seleccionado. La función así construida es continua y de primera derivada continua. Compárese esta gráfica con la correspondiente a la estrictamente cuadrática de la esquina superior izquierda de la figura 3.3.*

hessiano; en segundo, equivale a una aplicación del algoritmo *steepest descent* suponiendo forma cuadrática; finalmente, puede ser visto como un método adaptativo de estimar el valor óptimo del paso de aprendizaje en un algoritmo de búsqueda de gradiente. Siguiendo la interpretación adaptativa del algoritmo —la única que permite contemplar los casos en los que la segunda derivada es negativa—, el movimiento realizado en el instante t, $d_t = -\varepsilon_t g_t$ es dividido en dos componentes d_t^P y d_t^N, donde la primera es paralela a la dirección de g_{t-1}, y la segunda, perpendicular. La componente paralela puede ser vista como una corrección al movimiento realizado en $t-1$. Si d_{t-1} es muy pequeño en magnitud —porque el paso de aprendizaje también lo es—, la forma de la función de coste varía poco entre $t-1$ y t, por tanto, sus gradientes en ambos instantes son prácticamente idénticos. En esta situación, la ecuación (3.16) hace que el valor del paso de aprendizaje sea multiplicado por una cantidad, $\hat{\varepsilon} = 1 + (1-a)^{-1}$, mayor que uno —cinco para $a = 0,75$—. Por contra, si el paso de aprendizaje es excesivamente grande, y la función presenta un único mínimo en la recta que une Λ_{t-1} y Λ_t, la proyección del gradiente en t sobre el gradiente en $t-1$ da lugar a un vector de dirección opuesta a este último. El producto escalar es, por tanto, negativo, el valor de $\hat{\varepsilon}$ será menor que uno y el paso de aprendizaje se reducirá[3], tendiéndose a compensar, por tanto, su valor excesivo.

La convergencia del algoritmo depende del número de puntos críticos —máximos o mínimos de la función de coste— en el intervalo comprendido entre Λ_{t-1} y Λ_t definido por (3.15). Teniendo en cuenta que Λ_t se obtiene desplazando Λ_{t-1} en la dirección contraria

[3]Si la derivada de la función de coste en la dirección del gradiente es una parábola, también se cumplirá el criterio de *steepest descent*. Si, además, la función de coste tiene forma cuadrática con hessiano igual a una constante positiva multiplicada por la identidad, el óptimo de la función se alcanza en esta iteración.

a su gradiente, si $\mathcal{G}(\Lambda)$ es una función continua, existirá como mínimo un punto Λ^* tal que, siendo $\mu \leq 1$,

$$\mathcal{G}(\Lambda^*) = \mathcal{G}(\Lambda_{t-1} + \mu(\Lambda_t - \Lambda_{t-1})) \leq \mathcal{G}(\Lambda_{t-1}) \tag{3.17}$$

es decir, como mínimo existe un punto en la dirección tomada en el instante $t-1$ tal que la función de coste en ese punto es inferior a la función de coste en el punto de partida Λ_{t-1}. Por tanto, en el caso de existir uno o más puntos críticos, como mínimo uno de ellos debe ser un mínimo local. Son posibles cuatro situaciones distintas en función de que el número de puntos críticos sea cero, uno o más, y, en este último caso, un número par o impar:

1. Si no hay ningún punto crítico en la recta que une Λ_{t-1} y Λ_t, entonces el mínimo de la función de coste dentro del intervalo se encuentra justamente en Λ_t, y $\mathcal{G}(\Lambda_t) \leq \mathcal{G}(\Lambda_{t-1})$, es decir: el algoritmo converge. En este caso, además, el gradiente en Λ_{t-1} y Λ_t apunta en la misma dirección (a no ser que el propio Λ_t sea un punto crítico), el producto escalar es positivo y el valor del paso de aprendizaje aumentará según la ecuación (3.16).

2. Si hay un punto crítico, y sólo uno, éste será forzosamente un mínimo —siempre que la función de coste sea continua—. Los gradientes en $t-1$ y t apuntarán en direcciones contrarias, su producto escalar será negativo y el paso de aprendizaje se reducirá según la ecuación (3.14). Aunque no se cumpla que $\mathcal{G}(\Lambda_t) \leq \mathcal{G}(\Lambda_{t-1})$, siempre se verificará que $\mathcal{G}(\Lambda_{t+1}) \leq \max(\mathcal{G}(\Lambda_{t-1}), \mathcal{G}(\Lambda_t))$, y el algoritmo convergerá.

3. Si hay un número impar mayor que uno de puntos críticos, los gradientes en $t-1$ y t apuntarán en direcciones contrarias, reduciéndose el paso de aprendizaje. En este caso no se puede garantizar que $\Lambda_{t+1} \leq \Lambda_t$, aunque, si la situación persiste, el paso de aprendizaje se reducirá tanto que acabaremos cayendo en una de las dos primeras situaciones —cero o un punto crítico—.

4. Si hay más de un punto crítico, pero su número es par, los gradientes en $t-1$ y t apuntarán en la misma dirección, provocando el aumento del paso de aprendizaje. Esta es la situación más problemática para el algoritmo de búsqueda adaptativa de gradiente puesto que, no sólo no hay garantía de que el proceso converja — $\mathcal{G}(\Lambda_t)$ puede ser mayor que $\mathcal{G}(\Lambda_{t-1})$—, sino que, además, el paso de aprendizaje aún aumentará más, empeorándose la convergencia del algoritmo.

Considerando como condición de divergencia el hecho que la función de coste presente más de un punto crítico en iteraciones sucesivas, el único de los cuatro casos anteriores que realmente puede provocar que el algoritmo diverja es el cuarto, ya que los dos primeros implican convergencia, y el tercero tiende a corregirse por si solo. En el cuarto caso, cuando el número de puntos críticos es par, ε^* es mayor que uno y el paso de aprendizaje aumenta con independencia de que el movimiento realizado entre Λ_{t-1} y Λ_t fuera beneficioso o no. Puede ocurrir que una cierta dirección aparezca reiteradamente en el gradiente de la función de coste estimada localmente, pero que el comportamiento a mayor escala de la función de coste sea el opuesto. En ese caso el algoritmo de adaptación hará que el paso de aprendizaje aumente aún más de valor, agravándose el alejamiento del óptimo. No obstante, no es previsible que esta situación se dé a menudo. Por ejemplo, en los distintos experimentos realizados para este tesis, nunca se dio una situación de divergencia achacable a un crecimiento incontrolado del paso de

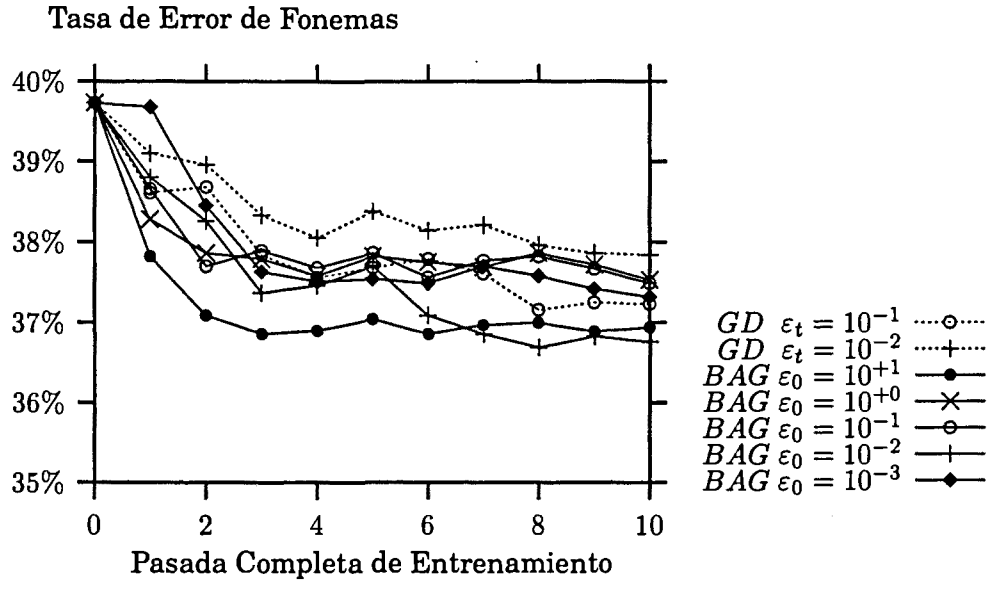

Figura 3.5: *Evolución de la tasa de error en DAF independiente del locutor de TIMIT usando el algoritmo adaptativo de búsqueda de gradiente con distintos valores del paso de aprendizaje inicial. También se muestran, con trazo discontinuo, los dos mejores experimentos usando búsqueda de gradiente con paso de aprendizaje fijo.*

aprendizaje como respuesta al comportamiento contradictorio entre el gradiente local de la función de coste y el comportamiento a gran escala de la misma. Además, esta situación, aunque especialmente grave en el caso de utilizar el algoritmo de adaptación del paso de aprendizaje, es siempre problemática. Así, el algoritmo de búsqueda de gradiente con paso de aprendizaje fijo también fracasa, aunque no tan estrepitosamente, en la minimización de este tipo de funciones.

Suponiendo que, como máximo, hay un punto crítico en la región definida por (3.15), el comportamiento del algoritmo de adaptación es aumentar el valor del paso de aprendizaje, cuando éste es insuficiente para acceder al mínimo en una sola iteración, y disminuirlo, cuando es excesivo para ese cometido. El proceso continua hasta que $\hat{\varepsilon}$ se estabiliza en un valor próximo a uno. En ese momento, los movimientos sucesivos tomados en la búsqueda de gradiente son realizados en direcciones ortogonales. Aunque la convergencia del algoritmo no está asegurada si no se cumple la existencia, como máximo, de un punto crítico en la región definida por (3.15), esta condición puede forzarse estableciendo un valor máximo para ε_t. No obstante, en la experimentación realizada, no se han observado problemas de convergencia sin necesidad de acotar el valor del paso de aprendizaje.

3.2.3 Resultados experimentales del algoritmo de búsqueda adaptativa de gradiente

El algoritmo de búsqueda adaptativa de gradiente ha sido aplicado a la misma tarea en que se aplicó la búsqueda de gradiente con paso de aprendizaje fijo. Para facilitar las comparaciones, los mismos cinco valores que se utilizaron en este último, son tomados ahora como valores iniciales del algoritmo. La figura 3.5 muestra la evolución de la tasa de error durante diez pasadas completas por el material de entrenamiento (cada pasada representa unas 6,5 actualizaciones del sistema y del paso de aprendizaje). También

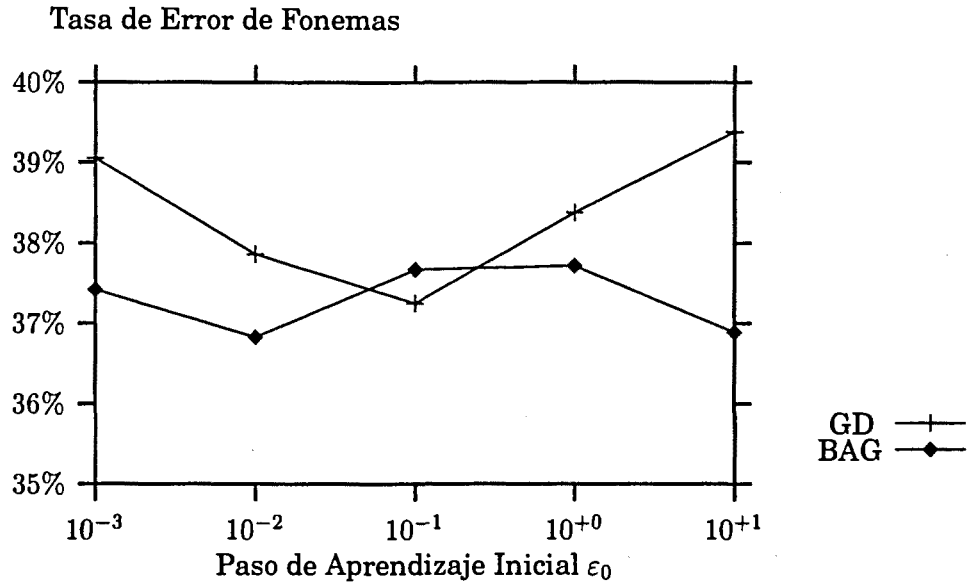

Figura 3.6: *Tasa de error en DAF independiente del locutor de TIMIT después de diez iteraciones utilizando búsqueda de gradiente y el algoritmo BAG con distintos valores del paso de aprendizaje inicial. Se observa que, en tanto la búsqueda de gradiente de paso de aprendizaje fijo presenta la típica forma de* bañera *—mostrando los efectos tanto de un valor excesivo de ε, como de uno escaso—, el algoritmo BAG alcanza un resultado superior sin mostrar ninguna tendencia clara.*

se muestran —con líneas discontinuas— los resultados obtenidos en las dos mejores ejecuciones de la búsqueda de gradiente con paso de aprendizaje fijo.

A la vista de la figura 3.5, puede concluirse que el resultado proporcionado por el algoritmo de búsqueda adaptativa de gradiente alcanza resultados similares al mejor caso de búsqueda de gradiente, con independencia del valor inicial del paso de aprendizaje. Por otro lado, en todos los casos se supera el segundo mejor resultado de búsqueda de gradiente con paso de aprendizaje fijo. Es decir, sin necesidad de determinar el valor óptimo del paso de aprendizaje, obtenemos un resultado como mínimo similar al mejor resultado obtenido utilizado una paso de aprendizaje fijo.

La independencia del algoritmo de la elección del paso de aprendizaje inicial queda aún más de manifiesto si se comparan los resultados en la última de las diez iteraciones. En la figura 3.6 se muestra la tasa de error en función del valor inicial del paso de aprendizaje —inicial y final, si éste no se adapta— tanto para la búsqueda de gradiente con paso de aprendizaje fijo, como para el algoritmo adaptativo. Puede observarse que, en el caso de utilizar paso de aprendizaje fijo, la curva descrita por la tasa de error en función del valor del paso tiene la típica forma de *bañera*: máximos en ambos extremos y un único mínimo más o menos centrado. Esta curva indica claramente la existencia de un único rango válido de valores, fuera del cual el comportamiento del algoritmo no es satisfactorio. Por contra, si el paso de aprendizaje es adaptado, el comportamiento es mucho más *plano*: existen dos mínimos, pero sin marcar una tendencia clara ya que están ubicados en extremos opuestos; y el resto de resultados son muy similares entre si. Además, la dispersión de los resultados es mucho menor cuando se utiliza el algoritmo adaptativo. De hecho, esta dispersión es muy similar al margen de confianza estadístico de los resultados —el número total de fonemas reconocidos es de 4.284, lo cual implica que el margen de confianza del 95% vale $\pm 0,7\%$—

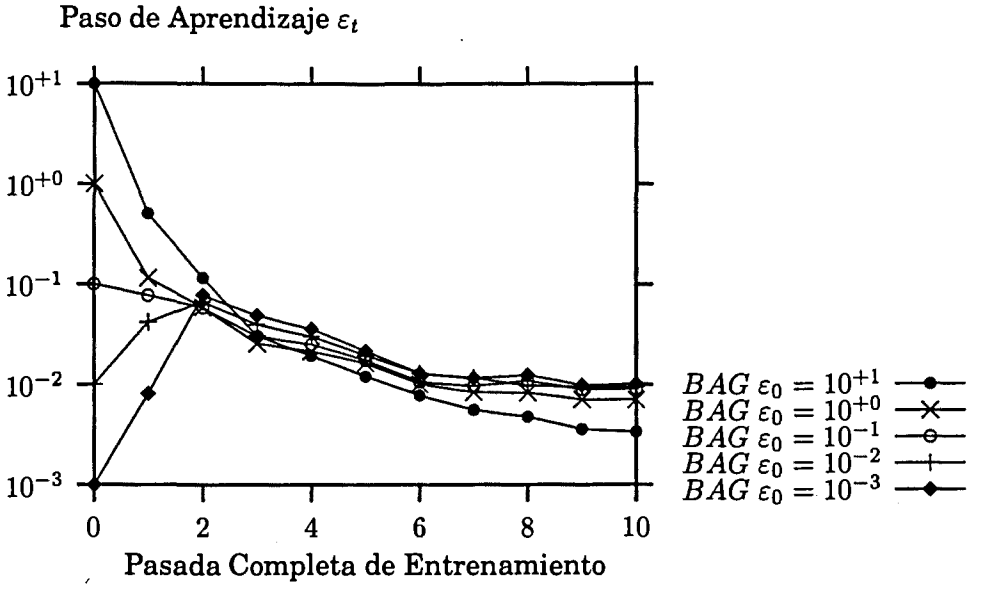

Figura 3.7: *Evolución del paso de aprendizaje, ε_t, en la ejecución del algoritmo BAG para cinco valores iniciales del paso de aprendizaje distintos. Se observa que, tras unas pocas iteraciones dedicadas a eliminar tanto un valor excesivo como uno escaso, la evolución es semejante en todos los casos.*

así que puede que dependa más de las condiciones de la estimación que de la propia elección del paso de aprendizaje inicial.

La figura 3.7 muestra la evolución del paso de aprendizaje a partir de los cinco valores probados. Con independencia del valor inicial, en todos los casos el paso de aprendizaje tiende inicialmente a un valor en torno a 10^{-1} —la elección óptima del paso de aprendizaje en búsqueda de gradiente—. Si el paso de aprendizaje inicial es inferior a ese valor, el algoritmo hace que crezca hasta alcanzarlo, y lo contrario si es superior. A partir de ese momento, en todos los casos se inicia un descenso mantenido del valor del paso de aprendizaje. Este descenso es tanto más acusado cuanto mayor fuera el valor inicial. Así, tras la iteración diez, se observa que la ordenación según el paso de aprendizaje es la opuesta a la inicial. Es decir, la ejecución con mayor paso de aprendizaje inicial es la que tiene un paso de aprendizaje menor en la décima iteración; la segunda mayor en la inicial deviene la segunda menor en la décima; y, así, todas las demás. Dado que el valor al que tiende inicialmente —con independencia de que, para ello, deba crecer o menguar— es apróximadamente igual al valor óptimo en búsqueda de gradiente con paso de aprendizaje fijo, y que este valor se alcanza en las muy primeras iteraciones, se puede afirmar que el algoritmo emplea estas primeras iteraciones en evitar tanto un valor excesivamente grande del paso de aprendizaje, como uno excesivamente pequeño. Una vez evitadas ambas situaciones indeseables, la evolución es siempre semejante: una curva descendente, aparentemente saturada —o, como mínimo, convexa— y que recuerda enormemente a la sucesión de valores del paso de aprendizaje utilizada en el algoritmo GPD [18].

Otro aspecto muy destacable de la figura 3.7 es que, excepto en las primeras iteraciones, en las cuales el objetivo de equilibrar el valor del paso de aprendizaje en un valor cercano al óptimo hace que el valor del paso de aprendizaje varíe rápidamente, una vez alcanzado este valor la evolución es muy lenta. Téngase en cuenta que el eje de abcisas refleja pasadas completas sobre todo el corpus de entrenamiento. Cada una de estas pasadas representa

unas 6,5 actualizaciones de los parámetros del sistema y del valor del paso de aprendizaje. Así pues, entre la pasada dos y la diez, transcurren $8 \times 6,5 = 52$ actualizaciones. En esas cincuenta y tantas actualizaciones el paso de aprendizaje a penas se ve divido por diez. Esto quiere decir que la variación porcentual en cada actualización es sólo de al rededor del 5%. El hecho que el paso de aprendizaje se mantenga casi constante implica que las direcciones del gradiente tomado en instantes consecutivos son ortogonales, verificándose, por tanto, el criterio del *steepest descent*.

3.3 Escalado de las Variables

El algoritmo *steepest descent* es muy sensible a la estructura de valores propios del hessiano de la función de coste. En concreto, la convergencia del *steepest descent* es tanto más lenta cuanto mayor es la relación entre el autovalor máximo y el mínimo. Por contra, si esta relación se mantiene próxima a uno, *steepest descent* accede al mínimo en una sola iteración para formas cuadráticas. Un modo de acelerar la convergencia de *steepest descent* —y, en general, cualquier algoritmo de búsqueda de gradiente— consiste en minimizar la función de coste en un espacio transformado en el cual el hessiano tenga autovalores de igual valor. El cambio de variables habitualmente empleada es una aplicación lineal $\Lambda' = U\Lambda$. Si U es definida positiva, la posición de los mínimos no varía y, optimizando $\mathcal{G}(\Lambda')$ es posible acceder al óptimo de $\mathcal{G}(\Lambda)$. El efecto de introducir este cambio de variables en el algoritmo de optimización se reduce a multiplicar el gradiente por la misma matriz U. Así, siendo U una matriz definida positiva, podemos sustituir (3.1) por la siguiente expresión:

$$\Lambda_{t+1} = \Lambda_t - \varepsilon_t U \nabla_\Lambda \mathcal{G}(X, \Lambda_t) \tag{3.18}$$

sin que el punto alcanzado para $t \to \infty$ se vea alterado —aunque sí lo haga el número de iteraciones empleadas en alcanzarlo—. Si U es diagonal, la transformación del espacio es equivalente a un escalado de las variables lo cual, a su vez, puede ser visto como la aplicación de un paso de aprendizaje distinto para cada variable.

3.3.1 Autoescalado de las variables utilizando el algoritmo de búsqueda adaptativa de gradiente

A la vista de los buenos resultados obtenidos aplicando el algoritmo de búsqueda adaptativa de gradiente, y, sobre todo, del excelente comportamiento de convergencia mostrado, cabe plantearse la posibilidad de su uso en la determinación automática de una matriz U diagonal que proporcione una estructura conveniente de los autovalores del hessiano de la función de coste. Como se ha comentado, la determinación de esta matriz es equivalente a aplicar pasos de aprendizaje distintos para cada variable. Por lo tanto, un mecanismo que permite alcanzar buenos resultados en la determinación del paso de aprendizaje puede resultar, también, conveniente en la determinación automática de la matriz de escalado idónea. La idea consiste en dividir el espacio vectorial de dimensión $|\Lambda|$ de los parámetros del sistema Λ, en P subespacios disjuntos, y calcular independientemente el paso de aprendizaje en cada uno de ellos. Si P es igual a uno, sólo se determina un paso de aprendizaje compartido por todas las variables, siendo equivalente al algoritmo sin escalado visto en el apartado anterior; si P es igual a la dimensión del sistema, $|\Lambda|$, cada variable es escalada independientemente del resto.

Si el hessiano de la función de coste es diagonal por bloques, es posible una estrategia intermedia entre utilizar un único paso de aprendizaje para todas las variables del sistema y uno distinto para cada una. En este caso, el espacio vectorial de las distintas variables del espacio puede ser dividido en subespacios vectoriales formados cada uno de ellos por conjuntos disjuntos de variables tales que el valor de la derivada parcial de la función de coste en la dirección de una variable en concreto sólo depende del valor del resto de variables que pertenecen al mismo subespacio. Por lo tanto, las distintas proyecciones del gradiente de la función de coste sobre cada uno de estos subespacios define un conjunto de direcciones del espacio que son, a un tiempo, ortogonales y conjugadas respecto al hessiano de la función de coste. Además, su suma[4] proporciona la dirección del gradiente. Forman, por tanto, un conjunto de direcciones idóneo para ser empleado en el método de los gradientes conjugados [62]. La ventaja de este método es que permite reducir la dimensionalidad del problema de optimización, disminuyendo el efecto de la relación entre los autovalores máximo y mínimo. Suponiendo diagonalidad por bloques, la posición del mínimo de la función de coste en la dirección de cada uno de estos gradientes conjugados no depende del resto, y acceder al mínimo en cada uno de ellos independientemente es equivalente a hacerlo secuencialmente —recalculando el gradiente después de modificar cada uno de los subespacios por separado—. Es posible, por tanto, dividir el conjunto de variables del sistema en S subconjuntos disjuntos; calcular el gradiente de la función de coste, g_t; su proyección sobre cada uno de los subespacios, g_t^s; aplicar la ecuación (3.16) independientemente en cada uno; y, finalmente, actualizar los parámetros del sistema, con la suma de los movimientos calculados, $\Lambda_{t+1} = \Lambda_t - \sum_S \varepsilon_t^s g_t^s$. Siendo equivalente, todo el proceso anterior, a la aplicación del método de los gradientes conjugados, utilizando S gradientes y las mismas hipótesis utilizadas en el algoritmo de búsqueda adaptativa de gradiente.

Si el hessiano de la función a optimizar no es diagonal por bloques, el gradiente de la función en un subespacio cualquiera dependerá del valor del resto de variables del sistema. El efecto de esta dependencia es que la posición del mínimo dentro del subespacio varía al modificar el valor de las variables en el subespacio complementario. En *steepest descent*, esta alteración de la posición del mínimo puede llegar a tener consecuencias perniciosas, si el algoritmo es aplicado separada e independientemente a ambos subespacios. En el caso del algoritmo de búsqueda adaptativa de gradiente, la situación es distinta, ya que el efecto de la alteración del resto de variables queda reflejado en la regla de adaptación del paso de aprendizaje. Así, la alteración del subespacio complementario puede tener dos repercusiones diferentes sobre la evolución del valor del gradiente entre los instantes $t-1$ y t: o bien el gradiente en $t-1$ es reiterado en t, con independencia del grado de aproximación alcanzado en la línea marcada por el gradiente en ese subespacio; o bien es corregido. El primer caso implica que, aunque el mínimo en el subespacio se alcanzara con el paso de aprendizaje actual, éste debería aumentar para hacer innecesarios ulteriores movimientos en esa misma dirección provocados por el efecto de la modificación del resto de variables. Del mismo modo, si el movimiento es corregido, el punto final accedido dentro del subespacio puede ser alcanzado en menos iteraciones si el paso de aprendizaje es disminuido. En última instancia, el caso que peores consecuencias tiene en *steepest*

[4]El término *suma* puede llevar a confusión, ya que se trata de subespacios vectoriales disjuntos. El sentido en que se emplea en el texto es el de *movimiento compuesto*, equivalente a realizar el movimiento en cada subespacio de manera independiente. Alternativamente, el término *suma* puede ser válido si se considera que la dimensión de cada subespacio es la misma que la del espacio completo, pero de manera que las variables no pertenecientes al subespacio valen cero.

descent —cuando el gradiente en el subespacio sufre alteraciones drásticas (con cambio de sentido), por culpa de la modificación del resto de variables— lleva al algoritmo de búsqueda adaptativa de gradiente a disminuir continuamente el paso de aprendizaje para ese subespacio, llegando a anular la búsqueda en él si su comportamiento depende más del valor del resto de variables que del de las que lo forman.

Aunque el autor carece de una garantía de convergencia del algoritmo fuera del caso en que la descomposición en subespacios desacoplados es posible, si ésta no es posible, el comportamiento del algoritmo puede no ser el idóneo, pero tampoco fatal —en el sentido de llegar a diverger—. Esto es así por que el algoritmo fuerza a que no haya exceso de entrenamiento en cada una de las direcciones en las cuales se descompone el gradiente, lo cual es mucho más restrictivo que sólo evitarlo en la dirección de éste. Así, el punto de equilibrio en la adaptación de un único paso de aprendizaje se alcanza cuando gradientes sucesivos son ortogonales. Para que gradientes sucesivos sean ortogonales es necesario que, o bien afectan a subconjuntos disjuntos de variables del sistema, o bien existe un subconjunto de las mismas cuyas derivadas parciales en iteraciones sucesivas son de signo contrario. Es decir, existirán subespacios vectoriales en los cuales el algoritmo tenderá a corregir en cada iteración el movimiento realizado en la anterior. Aún y existir cambios de signo en el gradiente, el valor del paso de aprendizaje se mantiene constante porque la contribución negativa al producto escalar de este subconjunto se compensa con la contribución positiva debida al subconjunto de variables cuyas derivadas parciales en iteraciones sucesivas es coincidente. Es decir, el método tiende a mantener un cierto grado de exceso de aprendizaje en algunas de las variables en aras a aprovechar al máximo la capacidad de aprendizaje de las que más pueden mejorar. En el caso de utilizar un paso de aprendizaje distinto para cada subespacio, por contra, esta compensación debe producirse en todos ellos. Si en alguno de ellos, por ejemplo, se da una alternancia continua en la dirección del gradiente en iteraciones sucesivas, el valor del paso de aprendizaje se reducirá hasta valer cero, sin necesidad de haber alcanzado un cero de la derivada parcial de la función de coste. Ahora bien, el hecho que el paso de aprendizaje se haga cero puede no ser una situación idónea, pero tampoco es fatal: no modificar el valor de los parámetros sin duda no puede mejorarlos, pero tampoco puede empeorarlos. Además, en tanto no se alcance un punto crítico, existirá al menos un subespacio para el que esta situación no se produzca, y el paso de aprendizaje para éste se mantendrá por encima de cero.

3.3.2 Elección de las direcciones conjugadas para funciones de coste en entrenamiento discriminativo de modelos acústicos

Los hessianos de las funciones de coste usualmente utilizadas en entrenamiento discriminativo distan mucho de ser diagonales o diagonales por bloques. Las funciones utilizadas suelen estar formadas por la suma de contribuciones correspondientes a cada uno de los posibles errores que se pueden cometer en el reconocimiento de una base de datos. Dado que cada una de estas contribuciones depende a un tiempo de más de una unidad, podemos esperar que el valor del gradiente para los parámetros del modelo de la palabra correcta dependerá del valor de los de la incorrecta, y viceversa. Por ejemplo: si modificamos los parámetros del modelo de la palabra incorrecta de manera que se imposibilite totalmente la comisión de la confusión, la contribución de este error al gradiente de los parámetros del modelo de la palabra a correcta se anulará totalmente. Es difícil, por tanto, encontrar conjuntos de variables perfectamente desacoplados. No obstante, existen dos posibles

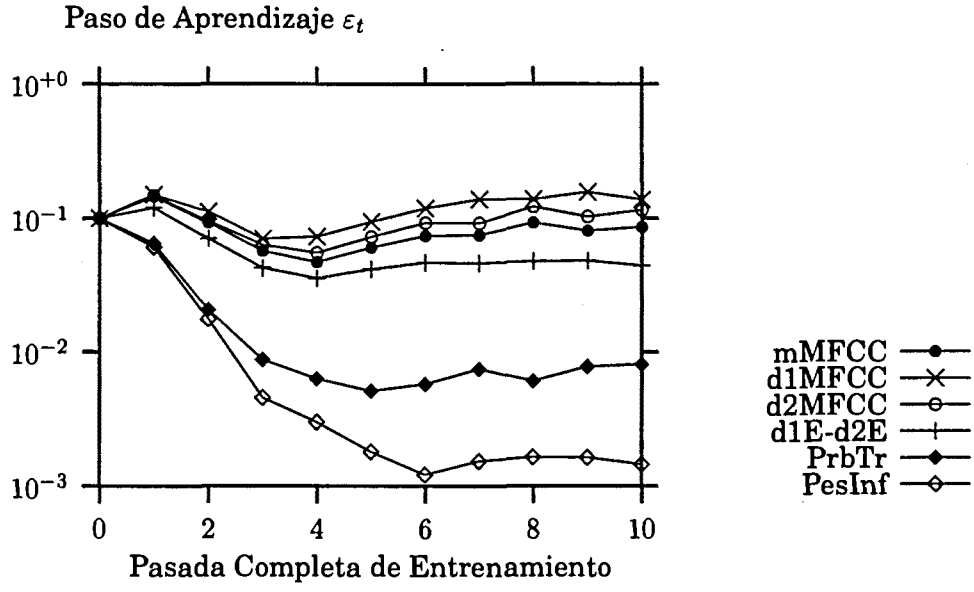

Figura 3.8: *Evolución del paso de aprendizaje, ε_t, en la ejecución del algoritmo BAG con autoescalado de las variables para cada tipo de información. Hay diferencias de más de un orden de magnitud entre los valores del paso de aprendizaje de parámetros distintos. Nuevamente, el algoritmo dedica las primeras iteraciones a evitar tanto el exceso como el defecto de entrenamiento, tras lo cual el valor alcanzado se mantiene aproximadamente constante.*

descomposiciones tales que sí podemos suponer que el valor de la derivada parcial dependa más del valor de variables del mismo subconjunto que del resto: la descomposición según tipo de parámetro y la según la unidad fonética. En el caso del tipo de parámetro —probabilidad de transición, de emisión de símbolo, peso dado a cada información, etc.—, es previsible una fuerte dependencia del propio tipo. Así podemos encontrarnos con que el valor óptimo del paso de aprendizaje para las probabilidades de emisión de símbolo sea varios órdenes de magnitud distinto que el óptimo para la probabilidad de transición o el peso de cada información (ver la figura 3.8. El motivo de esta discrepancia puede ser achacado a la propia naturaleza del tipo de cada parámetro, más que a las relaciones existentes entre parámetros de distinto tipo. Por otro lado, el valor de la derivada parcial de la función de coste con respecto a un parámetro del modelo de una cierta unidad fonética dependerá, entre otros factores, de lo frecuentemente que esta unidad está involucrada en errores, y de cuales son los errores más habituales en su reconocimiento. Esta es una característica que podemos suponer común a todos los parámetros que definen el modelo de la unidad y que, en principio, guardará menor relación con los parámetros de otras unidades que con el simple hecho de pertenecer a una u otra unidad fonética. Parece razonable, por tanto, dividir el conjunto de variables en subconjuntos formados por los diferentes tipos de variable de cada uno de los modelos de las diferentes unidades, y calcular un valor distinto del paso de aprendizaje para cada uno de estos subconjuntos. El número de parámetros a calcular se mantiene razonable —en los experimentos realizados, del órden de 300—; si se cumple la aproximación diagonal por bloques del hessiano en función del tipo de parámetro y de la unidad fonética, el método se beneficiará de las mayores prestaciones del método de los gradientes conjugados; por último, y aunque no se cumpla la aproximación diagonal, el algoritmo adaptativo tenderá a un valor *seguro* en

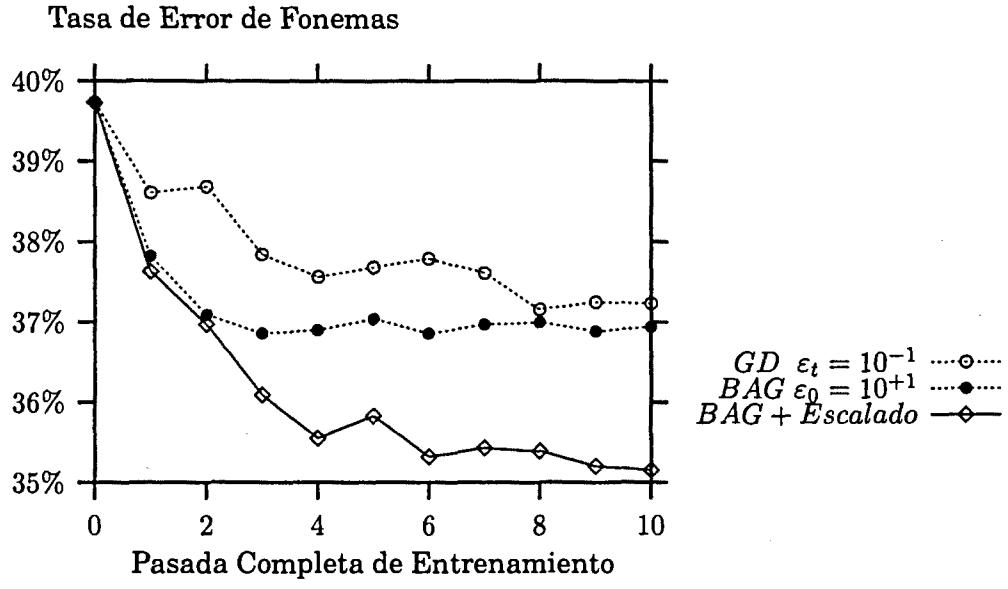

Figura 3.9: *Evolución de la tasa de error en DAF independiente del locutor de TIMIT usando el algoritmo BAG con auto escalado de las variables. También se muestran, con trazo discontinuo, los resultados del mejor experimento utilizando el algoritmo sin auto escalado y la búsqueda de gradiente de paso de aprendizaje fijo.*

cada subespacio de manera que sea imposible la divergencia separadamente. Ese valor, a medida que la aproximación diagonal por bloques es más exacta, tiende a nivelar la estructura de valores propios del hessiano, acelerando la velocidad de convergencia del *steepest descent*.

3.3.3 Resultados experimentales del algoritmo de búsqueda adaptativa de gradiente con autoescalado de las variables

En la experimentación realizada se optó por una estrategia combinada que, a un tiempo, relaja la aproximación de diagonalidad por bloques y permite estimar un valor distinto del paso de aprendizaje para cada tipo de parámetro y unidad fonética: en las iteraciones impares se supone que el gradiente de la función de coste sólo depende del tipo de parámetro; en las pares, de la unidad fonética. La figura 3.9 muestra el resultado de aplicar esta estrategia, así como los dos mejores obtenidos utilizando búsqueda de gradiente tanto con paso de aprendizaje fijo como adaptativo. La mejoría es notoria no sólo en cuanto a velocidad de convergencia sino, también, en cuanto al valor final alcanzado —casi dos puntos porcentuales mejor, duplicando el beneficio obtenido por la búsqueda de gradiente sin escalado de las variables—. Esta conclusión ya había sido apuntada usando el método GPD [41].

3.4 Otros Algoritmos de Optimización

Aunque el objetivo de esta tesis no es el estudio de los posibles algoritmos de aprendizaje utilizados en la aplicación de entrenamiento discriminativo, sino simplemente utilizarlos, el hecho de haber propuesto el uso de un algoritmo nunca antes empleado en este cometido hace necesario, como mínimo, referir otras alternativas más extendidas. Dejando

de lado el ya comentado algoritmo de búsqueda de gradiente con paso de aprendizaje fijo, existen otros dos algoritmos que han sido ampliamente utilizados en aplicaciones de entrenamiento discriminativo: el GPD y la fórmula de reestimación debida a Gopalakrishnan *et al.* La experimentación inicial llevada a cabo con el primero de ellos fue lo que impulsó al autor a probar por otros derroteros, y desarrollar el algoritmo finalmente empleado. Dado que, utilizando este otro, se alcanzó una solución satisfactoria, ya no se efectuó ninguna prueba más, ni con GPD, ni con el método de Gopalakrishnan —cuyo estudio hubiera requerido importantes modificaciones en los algoritmos utilizados—.

3.4.1 El algoritmo de búsqueda de gradiente estocástico, GPD

Una modificación al algoritmo de la ecuación (3.1) de gran popularidad en su aplicación al entrenamiento discriminativo es el algoritmo de búsqueda de gradiente estocástico, GPD [18]. En este algoritmo los parámetros del sistema son actualizados para cada elocución de entrenamiento de acuerdo con el gradiente instantáneo de la función para ella, $\nabla_\Lambda \mathcal{G}(X_t, \Lambda_t)$, con una elección de los sucesivos ϵ_t tal que éstos conforman una sucesión de valores decreciente que cumple $\sum_{t=0}^{\infty} \epsilon_t \to \infty$ y $\sum_{t=0}^{\infty} {\epsilon_t}^2 < \infty$.

La primera intención del autor de esta tesis fue la de utilizar GPD en la minimización de la confusibilidad. Prueba de esta firme voluntad inicial es el nombre dado al programa de entrenamiento: **BreoVGPD**, donde **Breo**[5] es el prefijo usado en los programas de entrenamiento de RAMSES, la **V** indica que el algoritmo seguido en la determinación de las N hipótesis es el de Viterbi, y **GPD**, el algoritmo supuestamente empleado en la optimización. Para garactizar las condiciones expuestas para el paso de aprendizaje se utilizaron secuencias de valores del tipo $\epsilon_t = \epsilon_0 D/(D + t)$, donde D controla la velocidad de decaimiento de la sucesión: un valor muy bajo implica un decaimiento muy rápido, en tanto que para $D \to \infty$, el paso de aprendizaje se mantiene constante. En la experimentación realizada con esta estrategia de optimización aparecieron innumerables problemas de convergencia. Así, si bien para elecciones adecuadas tanto de ϵ_0 como D, el algoritmo mostraba buena convergencia, pequeñas variaciones de estos valores, conducían a resultados muy pobres —en ocasiones, incluso peores que los iniciales—. Finalmente se optó por renunciar a la utilización de GPD por tres motivos:

1. La actualización de los parámetros agrupando las frases de entrenamiento en grupos de unas cien frases cada uno presentaba un comportamiento más estable que la actualización frase a frase.

2. Independientemente de la elección de D, el algoritmo es muy sensible al valor de ϵ_0, siendo necesaria una búsqueda sobre este valor para garantizar convergencia.

3. Una vez obtenido un valor adecuado de ϵ_0, el valor óptimo de D era siempre muy elevado, siendo válido un valor cercano a infinito tal que $\epsilon_t = \epsilon_0$.

En definitivas cuentas, la mejor elección de los parámetros del algoritmo GPD conducía, de manera sistemática, a algoritmos prácticamente idénticos a lo que se ha denominado búsqueda de gradiente. Así que se descartó el GPD en favor de esta última.

De todos modos, es de destacar que los resultados obtenidos utilizando el algoritmo de búsqueda adaptativa de gradiente vienen a avalar, en cierto sentido, al GPD. Así, de

[5] De Breogán, caudillo celta originario de Galicia que colonizó Irlanda. Padre de la nación gallega según el poeta romántico y autor de la letra de su himno, E. Pondal.

la observación de la figura 3.7 se deduce que, con independencia del valor del paso de aprendizaje inicial, existe una ley de decaimiento —cuando menos cuando $\epsilon_0 \geq 10^{-1}$— que sigue una forma semejante a la propuesta en [18] y alcanza el mismo resultado que el algoritmo adaptativo. Por tanto, la elección de los parámetros del GPD no requiere una búsqueda en dos dimensiones —valor inicial y decaimiento— sino que uno de los dos puede ser fijado de antemano a un valor arbitrario, y sólo ajustar el otro. No obstante, la comparación seguiría beneficiando al algoritmo adaptativo por dos razones: ni tan siquiera requiere del ajuste de un sólo parámetro; y proporciona un mecanismo válido para realizar el escalado de las variables de manera automático.

3.4.2 Fórmula de reestimación tipo Baum-Welch debida a Gopalakrishnan *et al.*

Una alternativa a los métodos basados en búsqueda de gradiente aparece en [33] para el caso de funciones de coste racionales, y es extendida en [84] a otras situaciones como las que aparecen en la optimización de modelos continuos de Markov. El mayor interés de esta solución radica en que proporciona una formulación similar a la del algoritmo de Baum-Welch, aunque sin heredar de éste una de sus características más notorias: la independencia de parámetros ajustables. Así, en este algoritmo también pueden llegar a usuarse dos parámetros ajustables distintos: uno que controla la velocidad de convergencia, y otro que maneja las situaciones en que el hessiano no es definido positivo. Aunque no se llegó a experimentar este método —más que nada, por que era más sencillo implementar el algoritmo de búsqueda adaptativa de gradiente a partir del algoritmo preexistente de GPD, y el resultado fue suficientemente satisfactorio—, los resultados publicados hasta el momento no muestran una diferencia sustancial entre éste y uno de búsqueda de gradiente [106]. No obstante, hay que destacar que el algoritmo de Gopalakrishnan *et al.* realiza un escalado intrínseco de las variables —agrupadas también en tipo de parámetro y unidad fonética— que probablemente resulte tan beneficioso como el realizado en la sección anterior, con el algoritmo de búsqueda adaptativa de gradiente, o en [41], usando GPD (aunque, en este último caso, el escalado no se realiza de manera automática, sino determinada heurísticamente).

Capítulo 4

Conclusiones

Aportaciones

La principal aportación de esta tesis es la de proporcionar un mecanismo útil para el entrenamiento discriminativo de unidades subléxicas en su aplicación a tareas de reconocimiento del habla continua, utilizando bases de datos independientes de la tarea: el entrenamiento de mínima confusibilidad en segmentos acústicos de longitud limitada. Con la experimentación presentada en este trabajo queda fuera de toda duda el importante beneficio que se puede obtener aplicando esta estrategia en el entrenamiento de modelos acústicos de fonemas independientes del contexto. En el caso de los semifonemas dependientes del contexto también se ha detectado un beneficio, aunque de menor magnitud que el alcanzado con los fonemas.

Como un subproducto del desarrollo teórico que se ha utilizado en la propuesta del entrenamiento de mínima confusibilidad sobre segmentos acústicos de longitud limitada, figura la adaptación a la tarea usando bases de datos de propósito general. Aunque el interés de este tipo de sistema es menor que los sistemas absolutamente independientes de la tarea, este esquema de adaptación es novedoso y permite mejorar aún más los resultados obtenidos con el entrenamiento independiente.

Finalmente, en la confección de la experimentación presentada se ha empleado un algoritmo, el de búsqueda adaptativa de gradiente, que permite tanto la independencia respecto a la elección del paso de aprendizaje en búsqueda de gradiente, como el escalado automático de las variables acelerando el proceso de optimización y permitiendo alcanzar prestaciones muy superiores al caso de utilizar un mismo paso de aprendizaje para todo el sistema.

Trabajo Futuro

Entre los aspectos tratados en esta tesis y que han quedado incompletos o pendientes de realización se pueden destacar:

- El tratamiento dado a la relevancia en esta tesis es muy rudimentario. En concreto, la suposición de que cualquier segmento puede ser confundido con cualquier otro tendría que ser revisada. Un tratamiento más sofisticado de la relevancia podría ser de interés no sólo en adaptación a la tarea, sino también en entrenamiento independiente.

- El entrenamiento de mínima confusibilidad está siendo adaptado, en la actualidad, a sistemas de reconocimiento de palabras clave *word-spotting* y de verificación de hipótesis.
- El algoritmo de búsqueda adaptativa de gradiente debería ser estudiado en mayor profundidad.
- Debería realizarse una experimentación más completa, incluyendo tareas de reconocimiento de vocabularios medios y grandes (por encima de las 2.000 palabras).

Apéndice A

Aspectos Prácticos de la Experimentación Presentada

A lo largo de esta tesis se ha ilustrado la utilidad de las distintas propuestas realizadas con los resultados obtenidos en el reconocimiento de una tarea concreta de reconocimiento del habla continua: el de las cadenas de dígitos de TIDIGITS utilizando modelos acústicos de unidades subléxicas entrenados a partir de las frases de la base de datos TIMIT. Los detalles fundamentales de esta tarea fueron expuestos en el capítulo introductorio, y los fundamentos del entrenamiento de mínima confusibilidad, en el siguiente. Quedan por exponer, no obstante, dos aspectos de la realización práctica que han resultado fundamentales en el éxito de la experimentación realizada. En primer lugar figura la acotación de la modificación de los parámetros en cada iteración del algoritmo de reestimación. En segundo, el modelado del lenguaje empleado en el cómputo de la relevancia y la función de confusibilidad. Ambas cuestiones son tratadas en este apéndice. También en él se comparan los resultados obtenidos en la serie de experimentos que sustentan esta tesis, analizándose los motivos de las discrepancias existentes entre ellos.

A.1 Acotación de la Modificación de los Parámetros

Desde la realización de los primeros experimentos destinados a la confección de esta tesis —utilizando el algoritmo GPD para minimizar el error de clasificación— se detectó la presencia de errores concretos que contribuían de manera exagerada a la fórmula de reestimación de los parámetros. La presencia de estos errores obligaba a utilizar un paso de aprendizaje muy pequeño, con el consiguiente aumento en el número de iteraciones necesario para alcanzar la convergencia. Este efecto era tan marcado que la reestimación de los parámetros involucrados en la comisión de otros errores puede llegar a hacerse despreciable frente a la de éstos. Sin embargo, son mucho más fiables los errores que dan lugar a modificaciones pequeñas de los parámetros que aquellos que implican modificaciones importantes. Muchas veces, estos últimos, son producidos por situaciones excepcionales. Por ejemplo, que la sucesión auténtica de modelos sea de imposible, o muy difícil, cumplimiento por la secuencia de entrenamiento. Dado que el proceso de optimización es iterativo, requiriendo decenas de iteraciones para llegar a término, es posible imponer una cota superior a la modificación introducida en cada parámetro en cada iteración. Si la modificación impuesta por el gradiente de la función de coste y el valor

del paso de aprendizaje no supera esa cota máxima, la modificación es realizada siguiendo fielmente el algoritmo de búsqueda de gradiente; por contra, si la modificación es superior a la cota impuesta, el parámetro sólo se modifica en la magnitud indicada por ésta. En el peor de los casos, que realmente sea necesaria la modificación del parámetro indicada por los valores del gradiente y el paso de aprendizaje, esta dirección aparecerá en el gradiente de la función de coste en las siguientes iteraciones, pudiéndose acceder igualmente al mínimo, aunque empleando más tiempo. En todos los experimentos realizados para esta tesis, la modificación de los parámetros en cada actualización de los mismos está limitada, mediante una función hiperbólica, a incrementarse o decrementarse en un 50%. Aunque no se ha hecho un estudio extenso del comportamiento de los distintos algoritmos de optimización probados frente a este parámetro, el valor indicado pareció apropiado tanto para la busqueda de gradiente con paso de aprendizaje fijo o adaptativo, como para el GPD. Ahora bien, mientras en el caso del algoritmo de búsqueda adaptativa, el beneficio de su inclusión era pequeño, y podía prescindirse totalmente de la acotación. En los otros dos casos, con paso de aprendizaje fijado de antemano, la acotación resultaba fundamental, puesto que permitía utilizar valores del paso de aprendizaje muy elevados sin que se produjera la variación exagerada de ninguna de los parámetros del sistema.

A.2 Modelado del Lenguaje en Entrenamiento de Mínima Confusibilidad

El esquema de entrenamiento discriminativo de unidades subléxicas para su aplicación al reconocimiento automático del habla continua presentado en esta tesis, la minimización de la función de confusibilidad dependiente, o no, de la tarea, se basa en otorgar una relevancia distinta a cada posible error entre segmentos acústicos en función de la frecuencia de aparición de este error en el conjunto de errores cometidos al confundir cada frase de la tarea con cada una de las demás. El cómputo exacto de esta cantidad requiere determinar la probabilidad de que se dé cada uno de los posibles alineados entre las secuencias que definen cada uno de los errores léxicos permitidos en la tarea. Estas cantidades dependerán de múltiples factores, como el parecido entre las secuencias a confundir, o el contexto en que se encuentran tanto en la frase correcta como en el error. Así, en condiciones reales de funcionamiento, existen errores que no se pueden dar, aún siendo secuencias acústicamente semejantes, porque implicarían la presencia de otros errores de muy dificil comisión. En otros casos, aún cuando es fácil distinguir la secuencia correcta del error, es importante realizar esta distición con el máximo margen posible ya que los contextos en que se encuentran son fácilmente confundibles.

A pesar del posible interés teórico del mejor modelado de las características de la tarea a reconocer, en la experimentación realizada para esta tesis sólo se ha contemplado una versión muy simplificada en la que se ha supuesto que tanto la secuencia correcta como la reconocida son independientes entre sí y de su contexto. Es decir, siempre que cometamos un error entre secuencias acústicas tendremos un error léxico, y sólo uno. La frecuencia de aparición de cada error entre secuencias acústicas se aproxima, simplemente, con el producto de las frecuencias de aparición de la secuencia acústica correcta y de la incorrecta en las frases de la tarea. El cálculo de esta frecuencia se realiza a partir del bigrama de unidades subléxicas. Es decir, la frecuencia de una cadena es la frecuencia de la primera unidad que la forma multiplicada por la probabilidad, condicionada a la unidad anterior, de

cada una de las que la siguen. Esta aproximación se realiza en el cómputo de la frecuencia de aparición tanto de las secuencias de entrenamiento en la base de datos —el bigrama del entrenamiento—, como de las hipótesis erróneas en la tarea —el bigrama de la tarea—.

En lo que resta de sección se repasan dos cuestiones distintas referidas al cómputo de la relevancia tal y como se ha realizado en la experimentación que acompaña esta tesis: el suavizado de los bigramas empleados, y la compensación de la relación entre inserciones y borrados.

A.2.1 Suavizado de la frecuencia de aparición de los segmentos de entrenamiento

Aunque la aproximación basada en el bigrama es muy burda y permite la inclusión de errores que realmente son muy improbables en condiciones reales de reconocimiento, la introducción de restricciones gramáticales, tanto en las frases que participan en el entrenamiento, como en los posibles errores considerados, provoca una reducción sustancial del material de entrenamiento realmente aprovechado. En concreto, el reconocimiento de las cadenas de dígitos en inglés incorpora un número muy pequeño de las transiciones entre fonemas presentes en la base de datos TIMIT. Dado que cualquier segmento que incorpore una transición entre fonemas imposible en las cadenas de dígitos no participa del entrenamiento, sólo una porción muy pequeña de TIMIT es realmente utilizada. Este problema sólo afecta al caso de entrenamiento adaptado a la tarea, ya que en el entrenamiento independiente de la tarea se considera como material de la tarea a optimizar al perteneciente a la propia base de entrenamiento y, por tanto, todos los segmentos que la forman son posibles.

A.2.1.1 Incorporación de transiciones frecuentes en el entrenamiento

Una posible solución al desaprovechamiento del material de entrenamiento debido a la introducción de restricciones gramaticales, consiste en aumentar la tarea con segmentos de aparición muy frecuente en la base de datos de entrenamiento. Por ejemplo, podemos utilizar una gramática ampliada donde el 90% de las frases esta formado por cadenas de dígitos, y el 10% restante, por la gramática de las frases que forman el propio entrenamiento. De este modo se consigue que todas las secuencias de entrenamiento participen realmente en él, aunque la relevancia de las secuencias con mayor número de transiciones presentes en la gramática de las cadenas de dígitos será mayor que la de aquellas secuencias que no pueden aparecer en esta gramática.

El objetivo de la introducción de estos segmentos es aprovechar el mayor número posible de secuencias de entrenamiento con el mínimo incremento de tamaño de la gramática considerada. Usando bigramas de unidades, la ampliación de la gramática es equivalente a permitir transiciones que no aparecen en la tarea. Por este motivo, en los experimentos realizados, se ha considerado cuatro situaciones distintas en función de las unidades involucradas en la transición:

1. Ambas unidades están presentes en la tarea, y su transición está permitida. En este caso no es necesario modificar el bigrama de la tarea aunque, si la frecuencia de aparición de la transición es muy superior en el entrenamiento que en la tarea, probablemente interese reflejarlo en la gramática utilizada de manera que se combata más aquellos errores que más aparecen en el entrenamiento —aunque

sean poco frecuentes en la tarea, tendremos una estimación muy buena de su comportamiento—.

2. Ambas unidades están presentes en la tarea; la transición entre ellas es muy habitual en la base de entrenamiento, pero no en la tarea. En este caso, incorporar la transición a la gramática extendida es muy interesante porque aumenta la probabilidad de considerar unidades que aparecen en la tarea.

3. Una de las unidades está en la tarea pero la otra no, y la transición es muy frecuente en el entrenamiento. Nuevamente interesa incorpora la transición a la gramática. De este modo, se incorporan unidades extrañas a la tarea, pero sólo acompañando a unidades que sí pueden aparecer.

4. Ninguna de las dos unidades participa en la tarea. No tiene sentido incorporar la transición. No incorporar transiciones entre unidades imposibles en la tarea garantiza que en cada uno de los segmentos que participan en el entrenamiento hay, como mínimo, la mitad del número de unidades del segmento menos uno unidades que pueden aparecer en la tarea. Si sí se incorporan estas transiciones, entonces puede participar, incluso, segmentos en los que ninguna de las unidades sea posible en la tarea.

Teniendo en cuenta estas consideraciones se optó por utilizar, en todos los casos, una gramática de la tarea formada por la suma ponderada de la gramática original de la tarea más la gramática formada por todas las transiciones desde o hasta una unidad presente en la tarea original de la gramática de menor tamaño que incorpora el 80% de las transiciones entre fonemas de TIMIT. Esta gramática se obtiene considerando las N transiciones más frecuentes tales que la suma de sus frecuencias de aparición supere el umbral prefijado del total de transiciones de la base de datos —el 80%, en este caso—. El factor de ponderación empleado es tal que la gramática original de la tarea representa el 90% de las transiciones de la extendida.

La utilización de esta gramática extendida demostró ser sumamente útil en todos los casos sin excepción. Así, en el entrenamiento de semifonemas, la mayor parte de las cadenas de cinco fonemas que se pueden obtener a partir del material de entrenamiento de TIMIT incorpora unidades o transiciones entre ellas que nunca pueden aparecer en la tarea del reconocimiento de cadenas de dígitos. En el caso de utilizar semifonemas dependientes del contexto, la situación resulta tan crítica que, aunque el algoritmo de entrenamiento es capaz de disminuir muy considerablemente la confusibilidad del material de entrenamiento, esta reducción depende tanto del *muy escaso* material realmente empleado que, no sólo no permite mejorar el reconocimiento de las cadenas de TIDIGITS, sino que, incluso, lo empeora ligeramente. En el caso del entrenamiento de fonemas independientes del contexto, la situación no es tan problemática: aunque el sistema funciona mejor si se entrena utilizando la gramática ampliada descrita en el párrafo anterior, en caso de no hacerlo, el resultado sigue siendo claramente mejor que el de referencia.

A.2.1.2 Suavizado de los modelos del lenguaje

Los modelos del lenguaje tanto del material de entrenamiento como de la tarea a reconocer tienen dos objetivos distintos: por un lado, el modelo del entrenamiento es utilizado para

normalizar la función de coste de la secuencia en función de la frecuencia de aparición de la secuencia en el entrenamiento. De este modo conseguimos que las contribuciones al gradiente de la función de coste debidas a secuencias distintas no dependan del número de veces que la secuencia aparece en el entrenamiento. Por otro lado, el modelo del lenguaje de la aplicación es utilizado para determinar lo frecuente que es un determinado error, cometido a nivel de secuencia de unidades subléxicas, en la tarea a reconocer.

El primero de los cometidos es el más problemático: normalizando por la frecuencia de aparición igualamos la contribución neta de cada posible secuencia de entrenamiento, con independencia de si esa contribución neta ha sido estimada con más o menos representantes. Si la cantidad de entrenamiento es muy elevada, todas las posibles secuencias del entrenamiento aparecerán lo suficiente como para que la estimación del gradiente sea acertada. En ese caso la normalización es necesaria para evitar que la estructura de la base de datos de entrenamiento influya en la optimización de la confusibilidad dependiente de la tarea. Lamentablemente, no siempre es así, y las transiciones entre fonemas más raras pueden aparecer un número muy pequeño de veces en la base de entrenamiento —media docena, o así, en TIMIT—. Dar a estas transiciones una importancia semejante a aquellas que aparecen miles de veces es poco realista atendiendo, únicamente, a la capacidad de generalizar el comportamiento del sistema en un caso y otro.

Para evitar que la normalización de la frecuencia de aparición de los segmentos de entrenamiento afecte a la fiabilidad del entrenamiento, se ha utilizado una versión aplanada de la gramática proporcionada por el bigrama. Esta versión consiste en utilizar la probabilidad elevada a una cierta potencia $0 \leq a \leq 1$ tal que el margen dinámico de sus valores se reduzca. Si se escoge $a = 0$, el sistema asigna la misma relevancia a todos los errores; si $a = 1$, la relevancia asignada es igual a la teórica. En los experimentos realizados para esta tesis, el valor habitualmente utilizado fue $a = 0,5$. En el caso del entrenamiento de mínima confusibilidad independiente de la tarea, se utilizó también este valor para el bigrama de la tarea. Su inclusión en este punto esta justificada en el hecho de que una gramática real de habla continua —sobre todo si el léxico es muy limitado, formado por palabras de corta longitud y todas ellas pueden ir detrás de cualquier otra; como es el caso de las cadenas de dígitos— suele tener una forma del tipo *todo o nada*. Es decir, o bien está permitida una cierta secuencia de unidades subléxicas, o no lo está. Las secuencias que sí pueden aparecer, además, suelen tener una frecuencia de aparición más o menos parecida. Cuando menos, la transición más común en una aplicación de habla continua de tamaño limitado no suele ser varios órdenes de magnitud más frecuente que la más rara, como sí es el caso en la gramática que representa al entrenamiento. Por otro lado, si el material de entrenamiento es escaso, el muestreo de los errores será muy pobre. En esas condiciones —que, utilizando TIMIT para el reconocimiento de las cadenas de dígitos, se dan con claridad— es más conveniente relajar el cómputo de la relevancia introduciendo un factor exponencial semejante al usado en el factor de normalización de la frecuencia de aparición en el entrenamiento. Este aplanado de la gramática de la tarea sólo es necesario cuando se aproxima utilizando la gramática del entrenamiento. Si se dispone de la verdadera gramática de la tarea, es mejor usarla tal cual, aunque sí se aplique el factor $a = 0,5$ en la gramática del entrenamiento.

En los experimentos realizados se utilizó siempre la gramática del entrenamiento elevado a un medio; y lo mismo se hizo con la de la tarea, pero sólo en entrenamiento independiente de la misma. En este último caso, el hecho de utilizar la misma gramática

para modelar tanto el material de entrenamiento como la tarea, y elevada a la misma potencia, hace que uno de los términos de la relevancia se anule exactamente con el término de la frecuencia de aparición en el entrenamiento.

A.2.2 Compensación de la relación entre inserciones y borrados en el cómputo de la relevancia

Durante los primeros experimentos realizados utilizando entrenamiento de mínima confusibilidad se detectó que el comportamiento del algoritmo respecto a los errores de inserción y borrado dependía de la incorporación o no de la relevancia de los errores en la función de confusibilidad. Así, en la experimentación del apartado 2.2.2, y presentada en [81], aún no se echa mano de la relevancia de cada error en el cómputo de la confusibilidad por dos motivos: porque el sistema de entrenamiento aún no estaba preparado para ello; y para poder comparar en igualdad de condiciones el entrenamiento de mínima confusibilidad con el de mínimo error de clasificación. En su lugar, se utilizó el valor uno como la relevancia de cualquier error visto en el entrenamiento —equivalente, por ejemplo, al trabajo de Lee [53]—. En estos experimentos, el entrenamiento discriminativo disminuye efectivamente la tasa de error del sistema, pero a base, fundamentalmente, de reducir la tasa de inserción. Esta situación se dio tanto para entrenamiento de mínima confusibilidad, como para el de mínimo error de clasificación, y tanto si el entrenamiento se realizaba en segmentos acústicos de corta duración, como si se utilizaba la frase entera. Por contra, al introducir la estimación de la relevancia, calculada a partir de los bigramas correspondientes, en el cómputo de la confusibilidad, el efecto era justamente el opuesto: la tasa de error también disminuía, pero a pesar de un incremento importante de la tasa de inserción.

Suponiendo que el comportamiento ideal fuera el atacar por igual inserciones y borrados, el motivo del alejamiento de este comportamiento en uno y otro caso es distinto, pero, en última instancia, depende de las hipótesis erróneas consideradas para cada segmento. Consideremos inicialmente el caso en que sí utilizamos la estimación de la relevancia de los errores. La aplicación estricta del algoritmo de mínima confusibilidad requiere calcular el gradiente de la función de posibilidad de comisión para cada posible confusión, multiplicar este gradiente por la relevancia del error y acumular el resultado en el gradiente de la confusibilidad global del sistema. En la aproximación utilizada de las gramáticas del entrenamiento y la tarea, el bigrama, no se impone ninguna restricción en la longitud de la cadena, o sea que incluir todas las posibles confusiones implica incluir cadenas de cualquier longitud —infinito, incluso, si no fuera por la longitud en tramas finita de la secuencia y las restricciones topológicas de los modelos de Markov—. De hecho, en la aproximación proporcionada por gramáticas estocásticas sin restricciones en la duración, la probabilidad del conjunto de cadenas de una longitud determinada es la misma e igual a uno, con independencia de la longitud considerada. Como consecuencia, la suma de las relevancias de los errores de una longitud determinada debería ser siempre la misma, cualquiera que sea la longitud considerada.

No obstante, la aplicación práctica del entrenamiento discriminativo no permite, en general, considerar todas las posibles hipótesis erróneas, ya que sería un número inabordable. En lugar de esto, sólo está en nuestras manos realizar un reconocimiento de las N hipótesis erróneas más probables y utilizar estas N hipótesis como una generalización del resto. Ahora bien, la representación de errores de inserción y de borrado en este conjunto de N errores no tiene porque estar equilibrada en ningún sentido y

dependerá, en gran medida, del número de hipótesis consideradas. Esto es así debido a que el número de hipótesis en las que se cometen borrados es inferior al de las que incluyen inserciones: sólo podemos dejar de reconocer aquellas unidades que efectivamente aparecen en la secuencia correcta; por contra, cualquier unidad del vocabulario puede dar lugar a una inserción. Considérese, por ejemplo, el caso de las posibles confusiones con un único error de una cadena de cinco fonemas sobre un total de 50. El número de hipótesis con un único error de borrado es cinco. Sin embargo, una cadena de cinco fonemas deja seis posiciones disponibles para insertar cualquier otra cosa, ante lo cual, el número de hipótesis con un error de inserción será $6 * 50 = 300$. Si a esto añadimos las hipótesis en las que sólo hay un error de sustitución, $5 * 49 \approx 250$, tenemos que para poder considerar todas las posibles secuencias erróneas en una sola unidad, debemos realizar un reconocimiento de las 600 hipótesis más probables, como mínimo. Evidentemente, este número es inabordable y, ni tan siquiera, garantizaría que la relevancia de los errores de cuatro, cinco y seis unidades fuera la misma, ya que no podríamos evitar la presencia de hipótesis con más de un error.

Si el número de hipótesis utilizado en entrenamiento de mínima confusibilidad es reducido, sólo una pequeña parte de los errores de inserción aparecerá entre ellas. Por contra, es relativamente fácil encontrar una parte importante de los posibles errores de borrado[1] —en el ejemplo de las cadenas de cinco unidades, un solo borrado representa el 20% del total de posibles borrados, para cubrir el 20% de hipótesis con una sólo inserción necesitamos 75 hipótesis—. En ese caso, la suma de las relevancias dadas a los errores de borrado superará a la de las dadas a los de inserción, y el sistema, tal y como se había detectado, tenderá a eliminar borrados, aún a costa de añadir inserciones. Ahora bien, si no se utiliza gramática alguna, el efecto es considerar la probabilidad de cualquier cadena idéntica e igual a uno; la relevancia dada a cada error con inserciones será igual a la dada a cualquier error con borrados; y, dado que el número de los primeros es muy superior al del los segundos, el sistema tenderá a eliminar inserciones, aún a costa de aumentar el número de borrados, tal y como también se había detectado.

No está claro cual puede ser el mecanismo de compensación del efecto de considerar un número escaso de hipótesis en el equilibrio entre la relevancia dada a errores de inserción o borrado. En principio, debería tenderse a una situación en la que la suma de las relevancias dada a hipótesis de la misma longitud no dependa de ésta pero, dado que la representación de los errores de longitud muy distinta de la real será pobre, este objetivo será de muy dificil cumplimiento. En su lugar, se ha optado por considerar que la mayor parte de los errores que aparecerán en la lista de las N hipótesis más probables serán de longitud igual a la secuencia correcta o, a lo sumo, una unidad más o menos. En ese caso, la relevancia dada a errores de longitud menor o igual a la de la secuencia correcta debería ser la misma que la dada a errores de longitud mayor o igual. Ello se consigue multiplicando la relevancia de cada error por un factor de compensación $C > 0$ elevado al número de unidades de la hipótesis. Si este factor de compensación es mayor que uno, el efecto de su introducción es aumentar la relevancia de las hipótesis más largas frente a las más cortas y, por tanto, aumentar la relevancia relativa de las inserciones frente a los borrados. Si es menor que uno, el efecto es el contrario y se aumenta la relevancia relativa de los borrados. El proceso se mantiene hasta que la relevancia dada a errores de longitud menor o igual de la longitud correcta es la misma que la de errores de longitud mayor o igual. Dado que el objetivo de la

[1]Podemos suponer que el sistema siempre tendrá errores de borrado en primera hipótesis. Si ese no fuera el caso, y el sistema no cometiera ningún error de borrado, éstos dejarían de ser un problema

inclusión de este factor es garantizar que la relevancia dada a un tipo y otro de error es la misma, y teniendo en cuenta que durante la ejecución del algoritmo es posible conocer ambas relevancias, un posible mecanismo para su determinación automática consiste en adaptar su valor a los errores considerados en la reestimación. Si el cómputo de la relevancia dada al conjunto de errores con borrados es superior a la de los que presentan errores de inserción, el factor C se incrementa, decrementándose en caso contrario. El proceso continua hasta que las relevancias se equilibran.

En los experimentos realizados para esta tesis, la compensación de la relevancia de inserciones y borrados ha demostrado ser una herramienta muy útil para conseguir que la aplicación de entrenamiento discriminativo no presentara especial predilección por uno u otro tipo de error, tanto si realmente se utiliza una estimación de la relevancia calculada a partir del modelo del lenguaje del error —situación que, *a priori*, favorece la erradicación de los errores de inserción— como si la relevancia dada a cualquier error es igual a uno —situación que, por el contrario, favorece la erradicación de borrados—.

A.3 Acerca de los Resultados Obtenidos en el Reconocimiento de TIDIGITS

En la serie de resultados experimentales presentada en esta tesis se ha intentado mantener unas condiciones de trabajo lo más homogéneas posible, centradas en una tarea concreta de reconocimiento del habla continua: el reconocimiento de las cadenas de TIDIGITS utilizando modelos acústicos entrenados a partir de TIMIT. El fruto de los distintos experimentos fue publicado en una serie de artículos presentados en 1998 y 1999, [81, 83, 80, 82]. Lamentablemente, el mantenimiento estricto de la coherencia entre los distintos resultados hubiera representado un sacrificio en las prestaciones del sistema. Por este motivo, y aunque cada uno de los artículos sí mantiene escrupulosamente la coherencia de sus resultados, existen diferencias importantes entre la experimentación realizada en cada uno de ellos por separado. Estas diferencias han sido siempre en un mismo sentido: mejorar las prestaciones del sistema en el reconocimiento de esta tarea; y han afectado a partes distintas del mismo. Presentadas en orden cronológico, las diferencias fundamentales en la experimentación de los cuatro artículos son tres:

1. En los experimentos del apartado 2.2.2.2 (presentados en [81]) no se compensan las tasas de inserción y borrado, ni en el entrenamiento de mínima confusibilidad (ver el apartado A.2.2), ni en el reconocimiento de la tarea. El hecho de introducir esta compensación en el reconocimiento resulta en una reducción muy importante de la tasa de error del sistema de referencia.

2. En los experimentos efectuados con semifonemas del apartado 2.4 (presentados en [82]) fue necesario incorporar el corpus de test de TIMIT en el entrenamiento para evitar el exceso de adaptación al material de entrenamiento debido a su escasez. No obstante, la condición de independencia del locutor se mantiene dado que los conjuntos de locutores de TIMIT y TIDIGITS son disjuntos.

3. También en los experimentos efectuados con semifonemas, fue necesario modificar la estructura de los modelos de Markov de manera que fueran equivalentes los modelos de Markov correspondientes a la concatenación de dos semifonemas independientes del contexto y el correspondiente a un fonema igualmente independiente del contexto.

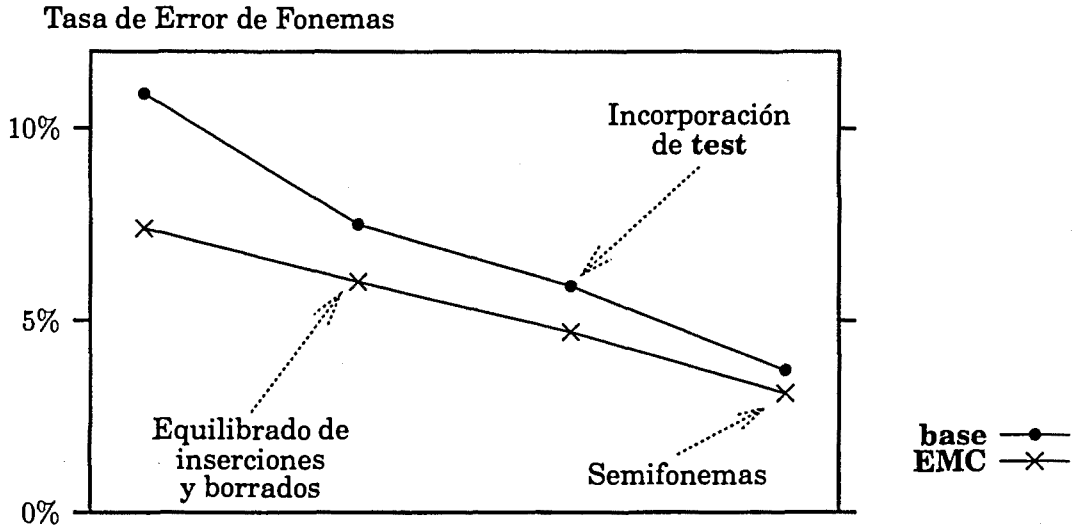

Figura A.1: *Evolución de la tasa de error de cadenas en el reconocimiento de TIDIGITS utilizando el sistema de referencia entrenado con Baum-Welch (**base**), y entrenamiento de mínima confusibilidad independiente de la tarea (**EMC**). A pesar de que las prestaciones del sistema de referencia han aumentado continuamente, la aplicación de entrenamiento discriminativo siempre ha resultado en mejoras importantes. No obstante, la tendencia es que el beneficio es cada vez menor.*

En los dos primeros casos, las modificaciones introducidas producen un notorio descenso en la tasa de error respecto al sistema de referencia (modelos acústicos de fonema entrenados según el criterio de máxima verosimilitud); el tercero, la modificación de la estructura de los modelos, no afecta significativamente al resultado. Así, mientras el experimento de referencia en el apartado 2.2.2.2 proporciona una tasa de error del 10,9%, en el apartado 2.4 esta tasa se reduce al 5,9%, gracias a la mayor disponibilidad de material de entrenamiento y a la compensación de las tasas de inserción y borrado. Ahora bien, en ambos casos se consigue una reducción similar, e importante, de la tasa de error de cadenas de dígitos aplicando entrenamiento de mínima confusibilidad —un 38% en el primer caso, un 34% en el segundo—, con lo que, aún cuando la comparación entre artículos distintos es difícil, el resultado proporcionado por cada uno separadamente sí permite visualizar la utilidad de este tipo de entrenamiento. Otras modificaciones de menor calado afectaron a la transcripción de la tarea, los algoritmos de optimización y los sistemas de entrenamiento y reconocimiento de RAMSES —siendo un sistema *vivo*, está sometido continuamente a revisión y mejora—. No obstante, otra vez, ninguna de estas modificaciones fue introducida durante la realización de la experimentación de ninguno de los artículos y, por tanto, los resultados en ellos contenidos pueden considerarse como realizados en igualdad de condiciones.

La figura A.1 muestra la evolución en la tasa de error sufrida por el experimento de referencia a lo largo de la tesis. Como punto final de esta evolución se ha incorporado también el resultado obtenido con semifonemas dependientes del contexto, el cual puede ser considerado punto de partida de los trabajos futuros. También se muestra el resultado obtenido en cada caso aplicando entrenamiento de mínima confusibilidad sobre segmentos acústicos de longitud limitada independiente de la tarea.

Apéndice B

Experimentación Realizada con SpeechDat en Castellano

A lo largo de esta tesis se ha ilustrado las propuestas realizadas —el entrenamiento de mínima confusibilidad y el algoritmo de búsqueda adaptativa de gradiente— con los resultados alcanzados en el reconocimiento de las cadenas de TIDIGITS a partir de modelos acústicos entrenados con TIMIT. La elección de esta tarea se debe, fundamentalmente, a cuestiones de índole práctica: son dos bases de datos de dominio público y de tamaño razonable. Además, la tarea reconocida es muy sencilla, con lo que el esfuerzo dedicado al reconocimiento es mucho menor. No obstante, también presenta inconvenientes: están grabadas en condiciones muy distintas, el vocabulario es de tamaño muy reducido y la tarea es poco representativa de lo que se ha dado en llamar *sistemas de reconocimiento de grandes vocabularios en habla continua*. Por este motivo, y aunque en el desarrollo de los algoritmos y presentación de esta tesis sólo se utilizara esta tarea, en todo momento se consideró la necesidad de corroborar los resultados obtenidos con otras tareas más significativas. En conreto, se han realizado varias baterías de experimentos utilizando la base de datos SpeechDat [75].

SpeechDat está formada por elocuciones en castellano correspondientes a 5.000 locutores provinientes de las cuatro grandes zonas dialectales del castellano hablado en España —centro, sur, norte y este—. De los 5.000 locutores, 4.000 son empleados en el entrenamiento, y los 1.000 restantes para evaluación. Las señales fueron grabadas telefónicamente, muestreadas a 8KHz y cuantificadas a 8 bits utilizando la codificación *Ley-A*. Cada locutor pronuncia cuarenta y tantas frases, de las que 9 son textos balanceados fonéticamente, y el resto frases correspondientes a distintas tareas de reconocimiento (fechas, cantidades monetarias, palabras clave, etc.). De las 36.000 frases útiles para realizar el entrenamiento de los modelos (las 9 balanceadas fonéticamente de los 4.000 locutores de entrenamiento) sólo se utilizan 20.000 para reducir los requisitos de cálculo. Se distinguen 31 alófonos del castellano, siguiendo las reglas de transcripción propuesta por Llisterri *et al.* [61]. Las tareas reconocidas son las siguientes:

Palabras Palabras aisladas ricas fonéticamente.

Horas G Horas del día utilizando una gramática regular.

Horas N Horas del día sin gramática.

Experimento	Palabras TEP	Horas N TEP	Horas G TEP	Horas G FRS
BaseFon	41,2	52,0	6,04	29,3
LangFon	19,2	39,7	3,51	21,1
BaseSefo	19,0	30,3	1,52	12,9
LangSefo	12,8	26,4	1,70	12,1

Tabla B.1: *Resultados obtenidos en el reconocimiento de la tarea de las palabras ricas fonéticamente y de las horas utilizando modelos acústicos entrenados con SpeechDat. Se muestran resultados de la tasa de error de palabras (**TEP**) y de frases (**FRS**). Se observa que el entrenamiento de mínima confusibilidad reporta un beneficio muy importante a todas las tareas cuando la unidad acústica empleada es el fonema independiente del contexto. En el caso del semifonema, sólo las tareas que no utilizan gramática se benefician de manera apreciable.*

La tabla B.1 muestra el resultado del reconocimiento de las tres tareas utilizando cada una de las cuatro configuraciones de entrenamiento. Sólo se indican las tasas de error por palabras (**TEP**) y por frases (**FRS**).

Apéndice C

Formulación de los Algoritmos Empleados en la Reestimación

Utilizando los algoritmos propuestos en esta tesis, la reestimación de los parámetros del sistema de reconocimiento, Λ, aplicando el criterio de mínima confusibilidad se realiza según la fórmula:

$$\Lambda' = \Lambda - \varepsilon \nabla_{\Lambda} l(\Lambda) \tag{C.1}$$

Donde ε se determina según lo indicado en el capítulo 3. Para poder aplicar esta fórmula, es necesario conocer el gradiente de la función de confusibilidad o, lo que es lo mismo, la derivada parcial de ésta con respecto a cada una de las variables λ^i.

$$\lambda'_i = \lambda_i - \varepsilon \frac{\delta l(\Lambda)}{\delta \lambda_i} \tag{C.2}$$

En este apéndice se detallan las fórmulas de reestimación correspondientes a cada uno de los parámetros optimizados en el proceso —probabilidades de transición entre estados, de emisión de símbolo y peso dado a cada información—.

Dada la base de datos de entrenamiento X formada por N elocuciones o segmentos de elocución, x_n^i —donde el superíndice indica que se trata de una realización de la palabra w_i—, la función de confusibilidad del sistema vale:

$$C(X, \Lambda, W) = \sum_{x_n^i \in X} \sum_{j \neq i} \frac{\mathcal{R}(w_i, w_j, W)}{|X^i|} \mathcal{E}_{ij}(x_n^i, \lambda_i, \lambda_j) \tag{C.3}$$

Donde la relevancia $\mathcal{R}(w_i, w_j, W)$ y el término $|X^i|$ sólo dependen de la tarea a reconocer y la estructura del material de reconocimiento, respectivamente. Aplicando la aproximación de la ecuación 2.5, la posibilidad de comisión de error, $\mathcal{E}_{ij}(x_n^i, \lambda_i, \lambda_j)$, vale

$$\mathcal{E}_{ij}(x_n^i, \lambda_i, \lambda_j) = \frac{1}{2}\left(1 + \tanh d_j(x_n^i)\right)\Big|_{j \neq i} = \frac{1}{2}\left(1 + \tanh \frac{g_j(x_n^i) - g_i(x_n^i)}{G_0}\right)\Big|_{j \neq i} \tag{C.4}$$

Donde $g_j(x_n^i)$ es el logaritmo de la probabilidad de generación de x_n^i por el modelo de la palabra w_j. Si el proceso de optimización se lleva a cabo únicamente sobre la sucesión de estados (o alineado) de máxima verosimilitud a lo largo del modelo de la palabra w_j, $Q_n^j = \{q_1^j q_2^j \ldots q_T^j\}$, y siendo $a_{q_{t-1}^j q_t^j}$ la probabilidad de transitar entre el estado q_{t-1}^j y el q_t^j,

γ_f el peso dado a la información f, y $B^f_{q^j_t}(x(t))$ la probabilidad de que el estado q_t del modelo λ_j produzca la parte de la trama $x^i_n(t)$ correspondiente a la información f, tenemos:

$$g_j(x^i, \lambda_j) = \sum_t \left[\log a_{q^j_{t-1} q^j_t} + \sum_f \gamma_f \log B^f_{q^j_t}(x^i_n(t)) \right] = \sum_t P_{q^j_t}(x^i_n(t)) \tag{C.5}$$

Donde $P_{q^j_t}(x^i(t)) = \log \left[a_{q^j_{t-1} q^j_t} \prod_f B^f_{q^j_t}(x^i_n(t))^{\gamma_f} \right]$ es el logaritmo de la contribución global de la trama t a la probabilidad de generación de la elocución.

Aplicando la regla de la cadena para determinar las derivadas parciales de la función se tiene:

$$\begin{aligned}
\frac{\delta C(x^i_n, \Lambda, W)}{\delta \lambda_k} &= \frac{\delta}{\delta \lambda_k} \sum_{x^i_n \in X} \sum_{j \neq i} \frac{\mathcal{R}(w_i, w_j, W)}{|X^i|} \frac{1}{2} \left(1 + \tanh d_j(x^i_n)\right) \Bigg|_{j \neq i} = \\
&= \sum_{x^i_n \in X} \sum_{j \neq i} \frac{\mathcal{R}(w_i, w_j, W)}{2|X^i| G_0 \cosh^2 d_j(x^i_n)} \frac{\delta}{\delta \lambda_k} \left[g_j(x^i_n) - g_i(x^i_n) \right] \Bigg|_{j \neq i} = \\
&= \sum_{x^i_n \in X} \sum_{j \neq i} \frac{\mathcal{R}(w_i, w_j, W)}{2|X^i| G_0 \cosh^2 d_j(x^i_n)} \begin{cases} -\sum_t \frac{\delta P_{q^i_t}(x^i_n(t))}{\delta \lambda_k} & k = i \neq j \\ \sum_t \frac{\delta P_{q^j_t}(x^i_n(t))}{\delta \lambda_k} & k = j \neq i \\ 0 & k \neq j,\ k \neq i \end{cases}
\end{aligned} \tag{C.6}$$

Es decir, cada posible confusión entre dos palabras, o segmentos acústicos, w_i y w_j, sólo contribuye al gradiente de la confusibilidad para los parámetros de los modelos acústicos de ambas palabras. Además, gracias a considerar únicamente el camino de máxima verosimilitud, cada trama de señal $x^i_n(t)$ sólo participa en las derivadas parciales de los parámetros del estado —en el modelo correcto y el incorrecto, $\lambda_{q^i_t}$ y $\lambda_{q^j_t}$—, a que es asignada según ese camino. De este modo es posible expresar la ecuación C.6 como el sumatorio, para todas las tramas de señal presentes en el entrenamiento, de un término igual para todas las tramas de una misma elocución, multiplicado por la derivada parcial del logaritmo de la probabilidad con que se genera la trama en el estado a que es asignada, $P_{q^k_t}(x^i_n(t))$.

$$\frac{\delta C(x^i_n, \Lambda, W)}{\delta \lambda_{q^k_t}} = \sum_{x^i_n(t) \in X} \sum_{j \neq i} \frac{\mathcal{R}(w_i, w_j, W)}{2|X^i| G_0 \cosh^2 d_j(x^i_n)} \begin{cases} -\frac{\delta P_{q^i_t}(x^i_n(t))}{\delta \lambda_{q^k_t}} & k = i \neq j \\ \frac{\delta P_{q^j_t}(x^i_n(t))}{\delta \lambda_{q^k_t}} & k = j \neq i \\ 0 & k \neq j,\ k \neq i \end{cases} \tag{C.7}$$

Donde la forma de $\delta P_{q^j_t}(x^i_n(t))/\delta \lambda_{q^k_t}$ depende del tipo de parámetro considerado.

Formula de reestimación para las probabilidades de transición entre estados. En este caso sólo se impone la condición de que todas las probabilidades de transición sean positivas. Para ello se introduce el siguiente cambio de variable:

$$\bar{a}_{q_{t-1}^k q_t^k} = \log a_{q_{t-1}^k q_t^k} \tag{C.8}$$

$$a_{q_{t-1}^k q_t^k} = e^{\bar{a}_{q_{t-1}^k q_t^k}} \tag{C.9}$$

Con este cambio de variable se garantiza que cualquier proceso de optimización que dé lugar a un valor real de $\bar{a}^k_{q^k_{t-1}q_t}$, implica un valor de $a_{q^k_{t-1}q^k_t}$ real y positivo.

No se impone, por el contrario, la restricción estocástica. Esto es, las probabilidades de transición reestimadas no tienen por qué sumar uno. Se ha optado por no imponer esta restricción debido a que el efecto de una suma de probabilidades distinta de uno es un mecanismo adecuado para compensar un exceso de omisiones o falsas alarmas para cada unidad.

Con el cambio de variables anterior, la derivada parcial de $P_{q_t^k}(x_n^i(t)$ respecto a $\bar{a}_{q^k_{t-1}q^k_t}$ queda:

$$\begin{aligned}
\frac{\delta P_{q_t^k}(x_n^i(t))}{\delta \bar{a}_{q^k_{t-1}q^k_t}} &= \frac{\delta}{\delta \bar{a}_{q^k_{t-1}q^k_t}} \log\left[a_{q^k_{t-1}q^k_t} \prod_f B^f_{q_t^k}(x_n^i(t))^{\gamma_f}\right] = \\
&= \frac{\delta}{\delta \bar{a}_{q^k_{t-1}q^k_t}} \left[\bar{a}_{q^k_{t-1}q^k_t} + \log \prod_f B^f_{q_t^k}(x_n^i(t))^{\gamma_f}\right] = \\
&= 1
\end{aligned} \tag{C.10}$$

Y la formula de reestimación de las probabilidades de transición transformadas es:

$$\bar{a'}_{q^k_{t-1}q^k_t} = \bar{a}_{q^k_{t-1}q^k_t} - \epsilon \sum_{x_n^i(t)\in X} \sum_{j\neq i} \frac{\mathcal{R}(w_i, w_j, W)}{2|X^i|G_0 \cosh^2 d_j(x_n^i)} \begin{cases} -1 & k = i \neq j \\ 1 & k = j \neq i \\ 0 & k \neq j,\ k \neq i \end{cases} \tag{C.11}$$

Formula de reestimación para las probabilidades de emisión de símbolo. En los modelos semicontinuos utilizados en esta tesis, la probabilidad de emisión de símbolo de la información f de la trama $x_n^i(t)$ en el estado q_t^k vale:

$$B^f_{q_t^k}(x_n^i(t)) = \sum_c w_n^c(t) b^{cf}_{q_t^k} \tag{C.12}$$

Donde $w_n^{cf}(t)$ da el valor de la cuantificación de $x_n^i(t)$ para el c-ésimo centroide del cuantificador de la f-ésima información, y $b^{cf}_{q_t^k}$ es la probabilidad de ese centroide en el estado q_t^k. En este caso, se impone tanto la condición de positividad, como de que la suma de las probabilidades de emisión de símbolo para cada estado e información sea uno. Para ello se introduce el siguiente cambio de variable:

$$\bar{b}^{cf}_{q_t^k} = \log b^{cf}_{q_t^k} \tag{C.13}$$

$$b^{cf}_{q_t^k} = \frac{e^{\bar{b}^{cf}_{q_t^k}}}{\sum_{c'} e^{\bar{b}^{c'f}_{q_t^k}}} \tag{C.14}$$

Este cambio de variables garantiza que cualquier proceso de optimización que dé lugar a un valor real de $\bar{b}^{cf}_{q^k_t}$, implica un valor de $b^{cf}_{q^k_t}$ real, positivo y tal que se cumplen las restricciones estocásticas.

Con este cambio de variables, la derivada parcial de $P_{q^k_t}(x^i_n(t)$ respecto a $\bar{b}^{cf}_{q^k_t}$ queda:

$$\begin{aligned}
\frac{\delta P_{q^k_t}(x^i_n(t))}{\delta \bar{b}^{cf}_{q^k_t}} &= \frac{\delta}{\delta \bar{b}^{cf}_{q^k_t}} \log\left[a_{q^k_{t-1}q^k_t}\prod_f B^f_{q^k_t}(x^i_n(t))^{\gamma_f}\right] = \\
&= \frac{\delta}{\delta \bar{b}^{cf}_{q^k_t}}\left[\log a_{q^k_{t-1}q^k_t} + \sum_{f'}\gamma_{f'}\log B^{f'}_{q^k_t}(x^i_n(t))\right] = \\
&= \gamma_f \frac{1}{B^f_{q^k_t}(x^i_n(t))}\frac{\delta}{\delta \bar{b}^{cf}_{q^k_t}}\sum_{c'} w^{c'}_n(t) b^{c'f}_{q^k_t} \\
&= \gamma_f \frac{1}{B^f_{q^k_t}(x^i_n(t))}\frac{\delta}{\delta \bar{b}^{cf}_{q^k_t}}\sum_{c'} w^{c'}_n(t)\frac{e^{\bar{b}^{c'f}_{q^k_t}}}{\sum_{c''} e^{\bar{b}^{c''f}_{q^k_t}}} \\
&= \gamma_f b^{cf}_{q^k_t}\left(\frac{w^c_n(t)}{B^f_{q^k_t}(x^i_n(t))} - 1\right)
\end{aligned} \tag{C.15}$$

Y la formula de reestimación de las probabilidades de emisión transformadas es:

$$\begin{aligned}
\bar{b}'^{cf}_{q^k_t} = \bar{b}^{cf}_{q^k_t} - \varepsilon \sum_{x^i_n(t)\in X}\sum_{j\neq i}\frac{\mathcal{R}(w_i,w_j,W)}{2|X^i|G_0\cosh^2 d_j(x^i_n)} \\
\gamma_f b^{cf}_{q^k_t}\left(\frac{w^c_n(t)}{B^f_{q^k_t}(x^i_n(t))} - 1\right)\begin{cases} -1 & k = i \neq j \\ 1 & k = j \neq i \\ 0 & k \neq j,\ k \neq i \end{cases}
\end{aligned} \tag{C.16}$$

Formula de reestimación para los pesos de las distintas informaciones. En este caso, también se impone la condición de positividad. Así mismo, se fuerza a que la suma de los pesos dados a las distintas informaciones de un estado se mantenga constante. Dado que esta suma es, inicialmente, igual al número de informaciones F, éste es el valor que mantendrá a lo largo del proceso de reestimación. El cambio de variable introducido es:

$$\bar{\gamma}^f_{q^k_t} = \log \gamma^f_{q^k_t} \tag{C.17}$$

$$\gamma^f_{q^k_t} = F\frac{e^{\bar{\gamma}^f_{q^k_t}}}{\sum_{f'} e^{\bar{\gamma}^{f'}_{q^k_t}}} \tag{C.18}$$

Con este cambio de variables, la derivada parcial de $P_{q_t^k}(x_n^i(t)$ respecto a $\bar{\gamma}_{q_t^k}^f$ queda:

$$\begin{aligned}
\frac{\delta P_{q_t^k}(x_n^i(t))}{\delta \bar{\gamma}_{q_t^k}^f} &= \frac{\delta}{\delta \bar{\gamma}_{q_t^k}^f} \log \left[a_{q_{t-1}^k q_t^k} \prod_f B_{q_t^k}^f (x_n^i(t))^{\gamma_{q_t^k}^f} \right] = \\
&= \frac{\delta}{\delta \bar{\gamma}_{q_t^k}^f} \left[\log a_{q_{t-1}^k q_t^k} + \sum_{f'} \gamma_{q_t^k}^{f'} \log B_{q_t^k}^{f'} (x_n^i(t)) \right] = \\
&= \log B_{q_t^k}^f (x_n^i(t)) - \frac{\sum_{f'} \gamma_{q_t^k}^{f'} \log B_{q_t^k}^{f'} (x_n^i(t))}{F}
\end{aligned} \tag{C.19}$$

Y la formula de reestimación de los pesos transformados de las informaciones vale:

$$\begin{aligned}
\bar{\gamma}'^{cf}_{q_t^k} &= \bar{\gamma}^{cf}_{q_t^k} - \varepsilon \sum_{x_n^i(t) \in X} \sum_{j \neq i} \frac{\mathcal{R}(w_i, w_j, W)}{2|X^i| G_0 \cosh^2 d_j(x_n^i)} \\
&\left(\log B_{q_t^k}^f (x_n^i(t)) - \frac{\sum_{f'} \gamma_{q_t^k}^{f'} \log B_{q_t^k}^{f'} (x_n^i(t))}{F} \right) \begin{cases} -1 & k = i \neq j \\ 1 & k = j \neq i \\ 0 & k \neq j,\ k \neq i \end{cases}
\end{aligned} \tag{C.20}$$

Bibliografía

[1] O. Alavedra. *Ús de Sub-bandes en la Parametrització del Senyal de Veu per al Reconeixement de la Parla.* Proyecto final de carrera, Universitat Politècnica de Catalunya, 1995.

[2] L. Bahl, P. Brown, P. de Souza and R. Mercer. Estimating Hidden Markov Model Parameters so as to Maximize Speech Recognition Accuracy. *IEEE Trans. on Speech and Audio Proc.*, pp. 77–83, Jan. 1993.

[3] L. Bahl, P.F. Brown, P.V. de Souza and R.L. Mercer. Maximum Mutual Information Estimation of Hidden Markov Parameters for Speech Recognition. In *Proc. of ICASSP'86*, pp. 49–52. 1986.

[4] L. Bahl, P.F. Brown, P.V. de Souza and R.L. Mercer. A New Algorithm for the Estimation of Hidden Markov Model Parameters. In *Proc. of ICASSP'88*, pp. 493–496. 1988.

[5] J.K. Baker. The Dragon System—An Overview. *IEEE Trans. on ASSP*, vol. 23 (1): pp. 24–29, Feb. 1975.

[6] L.E. Baum and J.A. Egon. An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and a Model for Ecology. *Bull. Amer. Meteorol. Soc.*, vol. 73: pp. 360–363, 1967.

[7] L.E. Baum and T. Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Bull. Amer. Meteorol. Soc.*, vol. 73: pp. 360–363, 1967.

[8] A. Bonafonte. *Comprensión del Habla en Tareas Semánticamente Restringidas.* Tesis doctoral, Universitat Politècnica de Catalunya, 1995.

[9] A. Bonafonte, R. Estany and E. Vives. Study of Subword Units for Spanish Speech Recognition. In *Proc. of EUROSPEECH'95*, pp. 1607–1610. Sep. 1995.

[10] A. Bonafonte and J.B. Mariño. Using X-Gram for Efficient Speech Recognition. In *Proc. of ICSLP'98*, pp. CD–ROM. Sep. 1998.

[11] A. Bonafonte, J.B. Mariño and A. Nogueiras. SETHOS: The UPC Speech Understanding System. In *Proc. of ICSLP'96*, pp. 2151–2154. Oct. 1996.

[12] A. Bonafonte, J.B. Mariño, A. Nogueiras and J.A. Rodríguez-Fonollosa. RAMSES: El Sistema de Reconocimiento del Habla Continua y Gran Vocabulario Desarrollado por la UPC. In *VIII Jornadas de I+D en Telecomunicaciones, Madrid.* Oct. 1998.

[13] A. Bonafonte, A. Nogueiras and A. Rodríguez. Explicit Segmentation of Speech Using Gaussian Models. In *Proc. of ICSLP'96*, pp. 1269–1272. Oct. 1996.

[14] J.L. Borges. *Narraciones*, chap. La Biblioteca de Babel, pp. 101–110. Cátedra, 1980.

[15] F. Casacuberta, R. García, R. Llisterri, C. Nadeu, J.M. Pardo and J.M. Rubio. Development of Spanish Corpora for Speech Recognition Research. In *Proc. of Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assesment Methods*. 1991.

[16] P.C. Chang and B.-H. Juang. Discriminative Training of Dynamic Programming Based Speech Recognizers. In *Proc. of ICASSP'92*. Mar. 1992.

[17] C. Chesta, A. Girardi and P. Laface. Discriminative Training of Hidden Markov Models Using a Classification Measure Criterion. In *Proc. of ICASSP'98*, pp. 449–452. May 1998.

[18] W. Chou, B.-H. Juang and C.-H. Lee. Segmental GPD Training of HMM based Speech Recognizer. In *Proc. of ICASSP'92*, pp. 473–476. 1992.

[19] W. Chou, C.-H. Lee and B.-H. Juang. Minimum Error Rate Training of Inter-Word Context Dependent Acoustic Model Units in Speech Recognition. In *Proc. of ICSLP'94*, pp. 439–442. 1994.

[20] Y.L. Chow, M.O. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, J. Makhoul, S. Roucos and R.M. Schwartz. BYBLOS: The BBN Continuous Speech Recognition System. In *Proc. of ICASSP'87*, pp. 89–92. 1987.

[21] H. Curry. The Method of Steepest Descent for Nonlinear Minimization Problems. *Quart. Appl. Math.*, vol. 2: pp. 258–261, 1944.

[22] K.H. Davis, B. Biddulph and S. Balashek. Automatic Recognition of Spoken Digits. *Journal of the Acoustic Society of America*, vol. 24 (6): pp. 637–642, 1952.

[23] R. de Cordoba and J.M. Pardo. Different Strategies of Distribution Clustering Using Discrete, Semicontinuous and Continuous HMM's in CSR. In *Proc. of ICSLP'96*, pp. 1101–1104. Oct. 1996.

[24] R. De Mori, M. Galler and F. Brugnara. Search and Learning Strategies for Improving Hidden Markov Models. *Computer Speech and Language*, vol. 9: pp. 107–121, Apr. 1995.

[25] R.S. Dembo, S.C. Eisenstat and T. Steinhaug. Inexact Newton Methods. *SIAM J. Numer. Anal.*, vol. 19 (2): pp. 400–408, Apr. 1982.

[26] R. Fletcher and M.J.D. Powell. A Rapidly Convergent Descent Method for Minimization. *Computer J.*, vol. 6: pp. 163–168, 1963.

[27] J.W. Forgie and C.D. Forgie. Results Obtained from a Vowel Recognition Computer Program. *Journal of the Acoustic Society of America*, vol. 31 (11): pp. 1480–1489, 1959.

[28] G.D. Forney. The Viterbi Algorithm. *Proc. of IEEE*, vol. 61: pp. 268–278, Mar. 1973.

[29] S. Furui. Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Trans. on ASSP*, vol. 34 (1): pp. 52–59, 1986.

[30] P.L. Galindo. A Competitive Algorithm for Training HMM for Speech Recognition. In *Proc. of EUROSPEECH'95*, pp. 2187–2190. Sep. 1995.

[31] M.B. Gandhi and J. Jacob. Natural Number Recognition Using MCE Trained Inter-Word Context Dependent Acoustic Models. In *Proc. of ICASSP'98*, pp. 457–460. May 1998.

[32] P.E. Gill and W. Murray. Quasi- Newton Methods for Unconstrained Optimization. *J. Inst. Maths. Applics.*, vol. 9: pp. 91–108, 1972.

[33] P. Gopalakrishnan, D. Kanevsky, A. Nadas and D. Nahamoo. An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems. *IEEE Trans. on Information Theory*, pp. 107–113, Jul. 1991.

[34] S. Herman and R. Sukkar. Joint MCE Estimation of VQ and HMM Parameters for Gaussian Mixture Selection. In *Proc. of ICASSP'98*, pp. 485–488. May 1998.

[35] J. Hernando. Maximum Likelihood Weighting of Dynamic Speech Features for CDHMM Speech Recognition. In *Proc. of ICASSP'97*, pp. 1267–1270. May 1997.

[36] X.D. Huang and M.A. Jack. Semi-Continuous Hidden Markov Models for Speech Signals. *Computer Speech and Language*, vol. 3: pp. 239–251, 1989.

[37] X.D. Huang and M.A. Jack. Unified Techniques for Vector Quantisation and Hidden Markov Modeling using Semi-Continuous Models. In *Proc. of ICASSP'89*, pp. 639–642. 1989.

[38] F. Jelinek. Continuous Speech Recognition by Statistical Methods. *Proc. of IEEE*, vol. 64: pp. 532–536, Apr. 1976.

[39] F. Jelinek. The Development of an Experimental Discrete Dictation Recognizer. *Proc. of IEEE*, vol. 73: pp. 1616–1624, Nov. 1985.

[40] F.T. Johansen. A Comparison of Hybrid HMM Architectures Using Global Discriminative Training. In *Proc. of ICSLP'96*, pp. 498–501. Oct. 1996.

[41] B.-H. Juang and S. Katagiri. Discriminative Learning for Minimum Error Classification. *IEEE Trans. on Signal Proc.*, pp. 3043–3054, Dec. 1992.

[42] B.-H. Juang and L.R. Rabiner. Mixture Autoregressive Hidden Markov Models for Speech Signals. *IEEE Trans. on ASSP*, vol. 33 (6): pp. 1404–1413, Dec. 1985.

[43] B.-H. Juang and L.R. Rabiner. The Segmental K-Means Algorithm for Estimating Parameters of Hidden Markov Models. *IEEE Trans. on ASSP*, vol. 38 (9): pp. 1639–1641, Sep. 1990.

[44] S. Kapadia, V. Valtchev and S.J. Young. MMI Training for Continuous Phoneme Recognition on the TIMIT Database. In *Proc. of ICASSP'93*, pp. 491–494. Apr. 1993.

[45] S. Katagiri and C.-H. Lee. A New Hybrid Algorithm for Speech Recognition Based on HMM Segmentation and Learning Vector Quantization. *IEEE Trans. on Speech and Audio Proc.*, vol. 1 (4), Oct. 1993.

[46] A. Kellner. Initial Language Model for Spoken Dialogue Systems. In *Proc. of ICASSP'98*, pp. 185–188. May 1998.

[47] T. Kohonen. The Self-Organizing Map. *Proc. of the IEEE*, vol. 78 (9), Sep. 1993.

[48] T. Kohonen, G. Barna and R. Chrisley. Statistical Pattern Recognition with Neural Networks: Benchmarking Studies. In *Proc. of IEEE Conference on Neural Networks*, pp. 61–68. 1988.

[49] J. Kowalik and M.R. Osborne. *Methods of Unconstrained Optimization Problems.* Elsevier, 1968.

[50] M. Kurimo. Comparison Results for Segmental Training Algorithms for Mixture Density HMM's. In *Proc. of EUROSPEECH'97*, pp. 87–90. Sep. 1997.

[51] L.F. Lamel, R.H. Kessel and S. Seneff. Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. In *Proc. of the Speech Recognition Workshop (DARPA).* 1986.

[52] C.-H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini and A.E. Rosenberg. Improved Acoustic Modeling for Large Vocabulary Speech Recognition. *Computer Speech and Language*, vol. 6: pp. 103–127, 1992.

[53] C.-H. Lee, B.-H. Juang, W. Chou and J.J. Molina. A Study on Task-Independent Subword Selection and Modeling for Speech Recognition. In *Proc. of ICLSP'96*, pp. 1820–1823. 1996.

[54] C.-H. Lee, L.R. Rabiner, R. Pieraccini and J.G. Wilpon. Acoustic Modeling for Large Vocabulary Speech Recognition. *Computer Speech and Language*, vol. 4: pp. 1237–1265, Jan. 1990.

[55] K.F. Lee. *Automatic Speech Recognition—The Development of the SPHINX System.* Kluwer Academic Publishers, 1989.

[56] K.F. Lee. Context Dependent Phonetic Hidden Markov Models for Speaker Independent Continuous Speech Recognition. *IEEE Trans. on ASSP*, vol. 38 (4): pp. 599–609, 1990.

[57] K.F. Lee and H.W. Hon. Speaker Independent Phone Recognition Using Hidden Markov Models. *IEEE Trans. on ASSP*, vol. 37 (11): pp. 1641–1648, 1989.

[58] K.F. Lee, H.W. Hon and D.R. Reddy. An Overview of the SPHINX Speech Recognition System. *IEEE Trans. on ASSP*, vol. 38: pp. 600–610, 1990.

[59] K.F. Lee and S. Mahajan. Corrective and Reinforcement Learning for Speaker-Independent Continuous Speech Recognition. In *Proc. of EUROSPEECH'89*, pp. 490–493. Sep. 1989.

[60] R.G. Leonard. A Database for Speaker-Independent Digit Recognition. In *Proc. of ICASSP'84*. 1984.

[61] J. Llisterri and J.B. Mariño. Spanish Adaptation of SAMPA and Automatic Phonetic Transcription. Report SAM-A/UPC/001/V1, Feb. 1993.

[62] D.G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, 2 edn., 1984.

[63] J.B. Mariño and E. Monte. Generation of Multiple Hypotheses in Connected Phonetic-Unit Recognition by a Modified One-Stage Dynamic Programming Algorithm. In *Proc. of EUROSPEECH'89*, pp. 408–411. Sep. 1989.

[64] J.B. Mariño, C. Nadeu and E. Lleida. Finite State Grammar Inference for Connected Word Recognition. In *Proc. of EUSIPCO'88*, pp. 1035–1038. Sep. 1988.

[65] J.B. Mariño, C. Nadeu, A. Moreno, E. Lleida and E. Monte. Recognition of Numbers and Strings of Numbers by Using Demisyllables: One Speaker Experiment. In *Proc. of EUROSPEECH'89*, pp. 102–105. Sep. 1989.

[66] J.B. Mariño and A. Nogueiras. Top-Down Bottom-Up Hybrid Clustering Algorithm for Acoustic-Phonetic Modeling of Speech. In *Proc. of EUROSPEECH'99*, pp. 1343–1346. Sep. 1999.

[67] J.B. Mariño, A. Nogueiras and A. Bonafonte. The Demiphone: an Efficient Subword Unit for Continuous Speech Recognition. In *Proc. of EUROSPEECH'97*, pp. 1215–1218. Sep. 1997.

[68] J.B. Mariño, A. Nogueiras, P. Pachès-Leal and A. Bonafonte. The Demiphone: a New Contextual Subword Unit for Continuous Speech Recognition. *Speech Communication*, Pendiente de publicación.

[69] J.B. Mariño, P. Pachès-Leal and A. Nogueiras. The Demiphone versus the Triphone in a Decision-Tree State-Tying Framework. In *Proc. of ICSLP'98*, pp. CD–ROM. Sep. 1998.

[70] J.B. Mariño, C. Nadeu, A. Moreno, E. Lleida, E. Monte and J. Salavedra. RAMSES: A Spanish Demisylable Based Continuous Speech Recognition System. *Speech Recognition and Understanding, NATO ASI Series F*, vol. 75: pp. 113–118, 1989.

[71] T.B. Martin, A.L. Nelson and H.J. Zadell. Speech Recognition by Feature Abstraction Techniques. Tech. rep., Tech. Report AL–TDR–64–176, Air Force Avionics Lab, 1964.

[72] M. Meteer and J.R. Rohlicek. Statistical Language Modelling Combining N-Gram and Context Free Grammars. In *Proc. of ICASSP'93*, pp. 37–40. Apr. 1993.

[73] E. Monte, J.B. Mariño and E. Lleida. The Back Propagation Using the Conjugate Gradient Method. In *EUSIPCO'90*, pp. 1615–1618. Apr. 1990.

[74] R. Moore. The Challenge of Domain-Independent Speech Understanding. In *Proc. of ICASSP'98*, pp. 1045–1048. May 1998.

[75] A. Moreno and R. Winsky. Spanish Fixed Network Speech Corpus. SpeechDat Project LRE-63314, 1997.

[76] C.S. Myers and L.R. Rabiner. A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition. *IEEE Trans. on ASSP*, vol. 29 (2): pp. 284–297, Apr. 1981.

[77] H. Ney, U. Essen and R. Kneser. On Structuring Probabilistic Dependencies in Stochastic Language Modelling. *Computer Speech and Language*, vol. 8: pp. 1–38, 1994.

[78] A. Nogueiras. *Entrenamiento Conectado de Modelos Ocultos de Markov para el Reconocimiento del Habla Continua.* Proyecto final de carrera, Universitat Politècnica de Catalunya, 1990.

[79] A. Nogueiras and J.B. Mariño. Maximum Likelihood Based Discriminative Training of Acoustic Models. In *Proc. of EUROSPEECH'95*, pp. 85–88. Sep. 1995.

[80] A. Nogueiras and J.B. Mariño. Task Adaptation of Sub-Lexical Unit Models using the Minimum Confusibility Criterion on Task Independent Databases. In *Proc. of ICSLP'98*, pp. CD–ROM. Sep. 1998.

[81] A. Nogueiras and J.B. Mariño. Task Independent Minimum Cofusibility Training for Continuous Speech Recognition. In *Proc. of ICASSP'98*, pp. 477–480. May 1998.

[82] A. Nogueiras and J.B. Mariño. Minimum Confusibility Training of Context Dependent Demiphones. In *Proc. of EUROSPEECH'99*, pp. 2741–2744. Sep. 1999.

[83] A. Nogueiras, J.B. Mariño and E. Monte. An Adaptative Gradient Search based Algorithm for Discriminative Training of HMM's. In *Proc. of ICSLP'98*, pp. CD–ROM. Sep. 1998.

[84] Y. Normadin, R. Cardin and R.D. Mori. High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation. *IEEE Trans. on Speech and Audio Proc.*, pp. 299–311, Apr. 1994.

[85] H.F. Olson and H. Belar. Phonetic Typewriter. *Journal of the Acoustic Society of America*, vol. 28 (6): pp. 1072–1081, 1956.

[86] P. O'Neill, S. Vaseghi, B. Doherty, W.-H. Tan and P. McCourt. Multi-Phone Strings as Subword Units for Speech Recognition. In *Proc. of ICSLP'98*, pp. 2523–2526. 1998.

[87] P. Pachès-Leal and C. Nadeu. On Parameter Filtering in Continuous Subword-Unit-Based Speech Recognition. In *Proc. of ICSLP'96*, pp. 1065–1068. Oct. 1996.

[88] A.M. Peinado, J.C. Segura, A.J. Rubio, P. García and J.L. Pérez. Discriminative Codebook Design Using Multiple Vector Quantization in HMM-Based Speech Recognizers. *IEEE Trans. on Speech and Audio Proc.*, vol. 4 (2), Mar. 1996.

[89] D. Povey and P.C. Woodland. Frame Discrimination Training of HMM's for Large Vocabulary Speech Recognition. In *Proc. of ICASSP'99*, pp. CD–ROM. 1999.

[90] P. Price, W.M. Fischer, J. Bernstein and D.S. Pallett. The DARPA 1,000-Word Resource Management Database for Continuous Speech Recognition. In *Proc. of ICASSP'88*, pp. 651–654. 1988.

[91] L.R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. of IEEE*, vol. 77 (2): pp. 257–286, Feb. 1989.

[92] L.R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[93] L.R. Rabiner, B.-H. Juang, S.E. Levinson and M.M. Sodhi. Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities. *AT&T Tech. Journal*, vol. 64 (6): pp. 1211–1234, 1985.

[94] L.R. Rabiner, S.E. Levinson, A.E. Rosenberg and J.G. Wilpon. Speaker Independent Recognition of Isolated Words Using Clustering Techniques. *IEEE Trans. on ASSP*, vol. 27: pp. 336–349, Aug. 1979.

[95] M. Rahim, Y. Bengio and Y. LeCun. Discriminative Feature and Model Design for Automatic Speech Recognition. In *Proc. of EUROSPEECH'97*, pp. 75–78. Sep. 1997.

[96] M.G. Rahim and C.H. Lee. Simultaneous ANN Feature and HMM Recognizer Design Using String-Based Minimu Clasification Error (MCE) Training. In *Proc. of ICSLP'96*, pp. 1824–1827. Oct. 1996.

[97] W. Reichl, S. Harengel, F. Wolfertstetter and G. Ruske. Neural Networks for Nonlinear Discriminat Analysis in Continuous Speech Recognition. In *Proc. of EUROSPEECH'95*, pp. 2163–2166. Sep. 1995.

[98] W. Reichl and G. Ruske. Discriminative Training for Continuous Speech Recognition. In *Proc. of EUROSPEECH'95*, pp. 537–540. Sep. 1995.

[99] L. Rigazio, J.-C. Junqua and M. Galler. Multi-Level Discriminative Training for Spelled Word Recognition. In *Proc. of ICASSP'98*, pp. 489–492. May 1998.

[100] G. Rigoll, C. Neukirchen and J. Rottland. A New Hybrid System Based on MMI-Neural Networks for the RM Speech Recognition Task. In *Proc. of ICASSP'96*, pp. 613–616. May 1996.

[101] R. Rosenfeld. Optimizing Lexical and N-Gram Coverage via Judicious Use of Linguistic Data. In *Proc. of EUROSPEECH'95*, pp. 1763–1766. Sep. 1995.

[102] J. Rottland, A. Lüdecke and G. Rigoll. Efficient Computation of MMI-Neural Networks for Large Vocabulary Speech Recognition Systems. In *Proc. of ICSLP'98*, pp. CD–ROM. Sep. 1998.

[103] J. Rottland, C. Neukirchen, D. Willett and G. Rigoll. Large Vocabulary Speech Recognition with Context Dependent MMI-Connectionist/HMM Systems Using the WSJ Database. In *Proc. of EUROSPEECH'97*, pp. 79–82. Sep. 1997.

[104] H. Sakoe. Two Level DP Matching—A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition. *IEEE Trans. on ASSP*, vol. 27: pp. 588–595, Dec. 1979.

[105] H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. on ASSP*, vol. 26 (1): pp. 43–49, Feb. 1978.

[106] R. Schlüter and W. Macherey. Comparison of Discriminative Training Criteria. In *Proc. of ICASSP'98*, pp. 493–496. May 1998.

[107] R. Schlüter, W. Macherey, S. Kanthak, H. Ney and L. Welling. Comparison of Optimization Methods for Discriminative Training Criteria. In *Proc. of EUROSPEECH'97*, pp. 15–18. Sep. 1997.

[108] R. Schwartz, Y. Chow, S. Roucos, M. Krasner and J. Makhoul. Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition. In *Proc. of ICASSP'84*, pp. 35.6.1–35.6.4. 1984.

[109] K. Seymore and R. Rosenfold. Scalable Backoff Language Models. In *Proc. of ICSLP'96*, pp. 232–235. Oct. 1996.

[110] J.E. Shoup. *Phonological Aspects of Speech Recognition*, chap. 6, pp. 125–138. Prentice-Hall, 1980.

[111] J.F. Traub. *Iterative Methods for the Solution of Equations*. Prentice-Hall, 1964.

[112] T. Vaich and A. Cohen. Comparison of Continuous-Density and Semi-Continuous HMM in Isolated Words Recognition Systems. In *Proc. of EUROSPEECH'99*, pp. 1515–1518. Sep. 1999.

[113] V. Valtchev, J.J. Odell, P.C. Woodland and S.J. Young. Lattice-Based Discriminative Training for Large Vocabulary Recognition. In *Proc. of ICASSP'96*, pp. 865–868. May 1996.

[114] V. Valtchev, J.J. Odell, P.C. Woodland and S.J. Young. MMIE Training of Large Vocabulary Speech Recognition Systems. *Speech Communication*, vol. 22: pp. 303–314, 1997.

[115] V. Valtchev, P.C. Woodland and S.J. Young. Discriminative Optimisation for Large Vocabulary Recognition Systems. In *Proc. of ICSLP'96*, pp. 18–21. Oct. 1996.

[116] A.J. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm. *IEEE Trans. on Information Theory*, vol. 13: pp. 260–269, Apr. 1967.

[117] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev and S.J. Young. The 1994 HTK Large Vocabulary Speech Recognition System. In *Proc. of ICASSP'95*, pp. 73–76. May 1995.

[118] G. Zavaliagkos, J. McDonough, D. Miller, A. El-Jaroudi, J. Billa, F. Richardson, K. Ma, M. Shiu and H. Gish. The BBN Byblos 1997 Large Vocabulary Conversational Speech Recognition System. In *Proc. of ICASSP'98*, pp. 905–908. May 1998.

[119] V. Zue, J. Glass, M. Phillips and S. Seneff. The MIT Summit Speech Recognition System: A Progress Report. In *Proc. of DARPA Speech and Natural Language Workshop*, pp. 179–189. 1989.

Índice de Figuras

Índice de Tablas

Glosario de Abreviaturas

BAG Búsqueda adaptativa de gradiente. Algoritmo de optimización de funciones de múltiples variables, utilizado en la experimentación presentada en esta tesis.

Base En las tablas de resultados de reconocimiento, representa el experimento de referencia, esto es: modelos acústicos de fonema entrenados con Baum-Welch.

Borr En las tablas de resultados de reconocimiento, indica el tanto por ciento de unidades acústicas omitidas, o borradas, en la cadena reconocida.

Clas En las tablas de resultados del reconocimiento, indica el tanto por ciento de unidades acústicas reconocidas correctamente conociendo con anterioridad los límites de cada una —clasificación de unidades subléxicas, o reconocimiento de dígitos aislados—.

Corr En las tablas de resultados del reconocimiento, indica el tanto por ciento de frases reconocidas correctamente.

DAF Decodificación acústico fonética. Consiste en el reconocimiento libre de gramática, o utilizando una estocástica o fonotáctica, de la cadena de fonemas que forman la frase.

ED Entrenamiento discriminativo.

EMC Entrenamiento de mínima confusibilidad. Criterio de entrenamiento discriminativo basado en la minimización de una medida suavizada del número esperado de posibles confusiones.

Error En las tablas de resultados de reconocimiento, indica la suma de las tasas de sustitución (Sust), inserción (Inse) y borrado (Borr).

GD *Gradiente descent.* Algoritmo básico de búsqueda de gradiente, en el que el paso de aprendizaje se determina a priori y se mantiene constante a lo largo del proceso de optimización.

Acierto En las tablas de resultados de reconocimiento, indica el tanto por ciento de unidades acústicas correctamente reconocidas.

GPD *Gradient probabilistic descent.* Algoritmo de optimización de funciones de múltiples variables, muy empleado en entrenamiento discriminativo.

HMM *Hidden Markov model.* Modelo oculto de Markov.

Inse En las tablas de resultados de reconocimiento, indica el tanto por ciento de unidades acústicas insertadas en la cadena reconocida.

LVQ *Learning vector quantization*. Algoritmo de entrenamiento discriminativo utilizado en la construcción de cuantificadores vectoriales óptimos.

MCE *Minimum clasification error*. Criterio de entrenamiento discriminativo basado en la minimización de una medida suavizada del número esperado de frases reconocidas incorrectamente.

MMIE *Maximum mutual information estimation*. Criterio de entrenamiento discriminativo basado en la maximización de la información mutua, medida derivada de la teoría de la información y representativa de la ambigüedad en la transcripción de la frase.

RAHC Reconocimiento automático del habla continua.

SALL Segmentos acústicos de longitud limitada.