UNIVERSIDAD POLITECNICA DE CATALUÑA

Departamento de Teoria de la señal y comunicaciones

TECNICAS DE PROCESADO Y REPRESENTACION DE LA SEÑAL DE VOZ PARA EL RECONOCIMIENTO DEL HABLA EN AMBIENTES RUIDOSOS

Autor: Francisco Javier Hernando Pericas
Director: Climent Nadeu i Camprubi

Barcelona, mayo 1993

Capítulo 2

RECONOCIMIENTO DEL HABLA EN ENTORNOS ADVERSOS

El desarrollo de la electrónica y la tecnología de los ordenadores está causando un crecimiento enorme del uso de máquinas para procesar información. En la mayoría de los casos esta información proviene de un ser humano y finalmente también es usada por un ser humano. Por tanto, son necesarios métodos efectivos de transferencia de información entre hombres y máquinas en ambas direcciones.

El habla es el medio más espontáneo y natural de comunicación entre los hombres. Sin embargo, hasta el presente se puede afirmar que en su comunicación con las máquinas el hombre ha hecho uso exclusivo del lenguaje escrito. Resulta natural, por tanto, extender la capacidad de comunicación hombre-máquina al mensaje oral.

Además de la naturalidad y espontaneidad aludidas, la comunicación oral hombre-máquina presenta importantes ventajas en gran cantidad de aplicaciones, como el diálogo interactivo o la entrada de grandes cantidades de datos en la máquina.

Una de estas ventajas es que en la comunicación oral las manos y la vista del usuario quedan liberadas, pudiendo dedicarse a una tarea simultánea a la comunicación. Ello ofrece posibilidades muy interesantes en el gobierno de sistemas de gran complejidad en que la atención visual sea muy importante.

Una segunda ventaja importante proviene del hecho de la universalidad de la red telefónica. Aunque ésta puede ser aprovechada para la transferencia de información sin

acudir al habla, la comunicación oral, al no requerir otro equipo que el teléfono, ofrece una ventaja sustancial: cualquier aparato telefónico se convierte en un enlace potencial con el ordenador y de este modo los acceso a bases de datos, las reservas y ventas de billetes de viaje, las operaciones bancarias, etc. podrían realizarse desde cualquier punto.

También pueden citarse como ventajas el hecho de que es factible en la oscuridad, facilita la comunicación simultánea con hombres y máquinas, puede ser más rápida que otros medios de comunicación, etc.

Una comunicación oral hombre-máquina debe reproducir el modelo que rige en el proceso de comunicación cotidiana entre humanos. Debemos, por tanto, facultar al ordenador para hablar y entender lo que se le dice. La capacidad de entendimiento constituye hoy en día un horizonte lejano.

Si se considera que entender el habla implica reconocer las distintas palabras del mensaje oral e interpretar los contenidos sintácticos y semánticos, se ha de reconocer que hoy en día solamente se han obtenido resultados parciales en el reconocimiento de la voz y se requieren avances fundamentales en gran número de disciplinas para acceder al entendimiento del habla.

Este capítulo está dedicado al estado del arte en el reconocimiento automático del habla en entornos adversos. En el apartado 2.1 se estudia el problema del reconocimiento del habla, en general. En primer lugar, se discuten las dificultades que conlleva el reconocimiento automático del habla en general derivadas de su carácter multiinteractivo, variable, continuo y redundante (apartado 2.1.1), que obligan a imponer restricciones al problema global del reconocimiento en cuanto al tipo de habla, la talla del léxico, la gramática del lenguaje, el número de locutores y las condiciones ambientales (apartado 2.1.2). Seguidamente se describen las principales técnicas que se aplican al reconocimiento automático del habla: comparación de patrones, modelos ocultos de Markov, redes neuronales y métodos basados en el conocimiento (apartado 2.1.3). En el apartado 2.2 se aborda el problema concreto del reconocimiento del habla en entornos adversos. Después de unas ideas sobre la naturaleza y dimensión del problema, se revisan los principales fenómenos físicos que provocan entornos adversos para el reconocimiento automático del habla (apartado 2.2.1) y las principales técnicas de reconocimiento robusto que se han propuesto en la literatura (apartado 2.2.2): transductores especiales, nuevas representaciones de la

señal, preprocesado de mejora de la voz, medidas de distorsión robustas, enmascaramiento, modelos adaptativos, etc.

2.1. SISTEMAS DE RECONOCIMIENTO AUTOMATICO DEL HABLA

Los orígenes del reconocimiento automático del habla hemos de buscarlos en la aparición de las primeras versiones del espectrógrafo en la década 1930-40, que permitieron vislumbrar por vez primera la posibilidad de realización de dispositivos automáticos capaces de reconocer la voz humana. Poco después, Davis, Bidulph y Balashek (laboratorios Bell, 1952) idearon el primer sistema, totalmente electrónico, capaz de discriminar con cierta precisión los dígitos ingleses pronunciados de forma aislada por un mismo locutor.

Los primeros trabajos que hacen uso de la tecnología informática comienzan a aparecer en los años 60. En estas fechas se produce una explosión de trabajos, principalmente de reconocimiento de palabras aisladas, con la impresión optimista de poder extrapolar los resultados y llegar en poco tiempo a sistemas capaces de reconocer cualquier frase pronunciada por cualquier locutor de manera continua.

Con este objetivo se emprenden posteriormente grandes proyectos de investigación, en los que se pretende llegar a las menores restricciones léxicas y gramaticales posibles. Concretamente, en 1971 el Departamento de Defensa de los EE.UU. lanza el mayor proyecto conocido de la historia del reconocimiento del habla, el ARPA-SUR (*Advanced Research Projects Agency - Speech Understanding System*) [New73].

Aunque los ambiciosos objetivos emprendidos por este y otros proyectos no llegaron a alcanzarse, las aportaciones derivadas de ellos contribuyeron de forma muy notable a un mejor conocimiento de los mecanismos del habla y de los problemas y limitaciones relacionados con el reconocimiento automático del habla, derivados de la complejidad de estos mecanismos. Esta toma de conciencia sobre la verdadera magnitud del problema planteó la necesidad de una mayor investigación fundamental.

A partir de entonces, los trabajos en este dominio han proseguido de forma más o menos continua. En la actualidad, el reconocimiento del habla es un campo de investigación con objetivos, métodos y aplicaciones bien definidos, en el que hay mucho trabajo a realizar a distintos niveles (teórico-práctico) y en distintas materias

(Procesamiento de la Señal, Acústica, Fonética, Reconocimiento de Formas, Inteligencia Artificial, etc.).

2.1.1. PROBLEMAS ASOCIADOS AL RECONOCIMIENTO DEL HABLA

El habla es una de las principales manifestaciones de la inteligencia humana y en la actualidad se es consciente de la enorme dificultad que entraña la concepción de sistemas que intenten aproximarse a sus prestaciones. La dificultad de automatizar los procesos de percepción y comprensión del habla reside en la complejidad de los mismos. Ninguno de estos dos procesos es suficientemente conocido como para ser incorporado a una máquina en forma de algoritmo.

En este apartado se revisarán someramente los principales problemas que dificultan el reconocimiento automático del habla.

Multiinteractividad

Existen varios niveles de percepción y/o comprensión, que interaccionan dinámicamente entre sí y en combinación con otros sistemas perceptivos (visual, por ejemplo) y motores (interacción entre aparato fonador y auditivo, producción de gestos, etc.). Cada uno de estos niveles utiliza los conocimientos sobre el lenguaje que le son propios, para extraer su parte correspondiente de la información total necesaria para la comprensión del lenguaje. La división de niveles más comúnmente aceptada es la siguiente:

1) Nivel acústico: en el que se analizan las características físicas de la señal de voz (energía, frecuencia fundamental, formantes, transiciones, etc.).

2) Nivel fonético: en el que se extraen los objetos sonoros elementales (fonemas, ruidos simples, etc.).

3) Nivel léxico: donde empieza la abstracción y se determinan las estructuras simbólicas primarias (palabras o morfemas).

4) Nivel sintáctico: en el que se aplican reglas para analizar la sucesión de palabras y comprobar su adecuación a la gramática del lenguaje, lo cual impone una determinada relación entre ellas.

5) Nivel semántico-pragmático: donde se llega a la comprensión del significado del lenguaje, eliminando las posibles interpretaciones absurdas y comprobando la coherencia del mensaje recibido con el conocimiento previo que de la realidad se dispone, así como del contexto en que discurre el diálogo.

En la actualidad, todavía no existe un formalismo que permita la integración e interpretación de las informaciones correspondientes a los diferentes niveles, haciendo compleja la tarea de reconocimiento del lenguaje natural.

Continuidad

A pesar de que se tenga la impresión contraria, en el habla natural ni los fonemas, ni las sílabas, ni siquiera las palabras, constituyen elementos discretos que se puedan separar fácilmente de forma automática. No existen pausas entre los elementos y además se influyen unos a otros debido a lo progresivo de los movimientos articulatorios del aparato fonador (coarticulación). La separación de estos elementos la realiza el ser humano gracias a sus conocimientos previos de la lengua, siendo esta una de las principales características distintivas entre el reconocimiento automático del habla y el tratamiento de textos ortográficos.

Variabilidad

El habla presenta una gran variabilidad: es imposible que un locutor (y con más razón varios locutores) pronuncie dos veces exactamente igual una misma sílaba, palabra o frase.

En cuanto a la variabilidad intralocutor, esta se produce incluso en la lectura cuidadosa de locutores entrenados. Aparte de variaciones circunstanciales de entonación, amplitud, etc., se producen variaciones temporales no lineales en los elementos constitutivos del habla, todas ellas dependientes del contexto, así como alteraciones producidas por el estado de ánimo del locutor y sus condiciones físicas (cansancio, catarro, etc.) o por el modo de pronunciación (susurrar, cantar, gritar, etc.).

Por lo que respecta a la variabilidad interlocutor, ésta es debida principalmente a las diferencias físicas de los aparatos fonadores de los locutores, que dependen en gran medida del sexo y la edad y dan lugar a variaciones en la escala de frecuencias. También son importantes los hábitos de habla diferentes, según la procedencia geográfica, el entronque social, etc.

Por último, una importante fuente adicional de variabilidad la constituyen el entorno y el canal de transmisión: ruidos, interferencias, reverberaciones de la propia sala, tipo de micrófono, características frecuenciales de la línea de transmisión, modos de articulación dependientes del entorno,... Este tema será objeto de un estudio detallado en el apartado 2.2 y, como ya se ha mencionado, constituye la motivación de este trabajo.

Debido a esta variabilidad, es necesario observar una gran cantidad de datos relativos a los diferentes elementos constitutivos del habla, a fin de extraer las características esenciales de estos con independencia del entorno, contexto y locutor.

Redundancia

La mayor parte de la información contenida en la voz es redundante. Se puede demostrar que unos 50 bits/segundo son suficientes para transmitir el mensaje lingüístico contenido en ella, mientras que para una transmisión sonora completa se requieren del orden de 100.000 bits/segundo (8 KHz de frecuencia de muestreo y 12 bits por muestra, por ejemplo).

Este suplemento de información contiene los datos que identifican al locutor, su estado de ánimo, su entorno,... y los que hacen posible la comprensión a pesar de la variabilidad y en entornos plagados de ruido. Es por esto que un sistema de reconocimiento ha de enfocar su atención en la extracción de los parámetros que caractericen el tipo de información útil para el proceso de reconocimiento.

2.1.2. RESTRICCIONES DE LOS SISTEMAS DE RECONOCIMIENTO

Dada la gran complejidad del proceso general de producción de la voz, en el desarrollo de los sistemas de reconocimiento automático del habla se hace necesario, pues, introducir restricciones más o menos severas con el objeto de simplificar el problema general hasta llevarlo a planteamientos abordables.

La primera restricción suele consistir en prescindir de la interacción con otros sistemas perceptivos (con la consiguiente pérdida de importantes aportaciones de información). Además se suele simplificar el esquema de niveles de percepción, así como reducir el contexto semántico y muchas de las variabilidades de la señal vocal. Aun con estas restricciones, el problema es todavía de una complejidad enorme y en la concepción de sistemas reales siempre se enfatiza en mayor o menor grado algunas de

las simplificaciones, con lo que los sistemas resultantes quedan especializados en determinados aspectos del habla.

A continuación se revisarán los cinco aspectos fundamentales en que se suelen realizar las simplificaciones mencionadas.

Tipo de habla

Este tipo de restricción está relacionada con la forma en que el locutor ha de pronunciar las palabras. La restricción más fuerte corresponde a los sistemas denominados de palabras aisladas, en los que se condiciona al locutor a pronunciar las palabras con una cierta separación temporal (200 ms) entre ellas. Un nivel inferior de condicionamiento corresponde a los sistemas de palabras conectadas, en los que el locutor puede pronunciar las palabras de forma fluida, pero cuidadosa. Una aproximación similar es la denominada *word-spotting* [Lle92], en la que el objetivo es la identificación de palabras correspondientes a un determinado vocabulario inmersas en frases en las que pueden aparecer otras palabras ajenas al mismo. Naturalmente, el interés final de la investigación en reconocimiento del habla hay que situarlo en el diseño de sistemas capaces de interpretar el mensaje oral tal como es producido por un interlocutor humano cuando se comunica con sus semejantes (habla continua), con la misma naturalidad y rapidez.

El reconocimiento del habla continua es significativamente más difícil que el de palabras aisladas. Su complejidad es el resultado de tres propiedades del habla continua. En primer lugar, en el habla continua los límites de cada palabra no son claros y son difíciles de encontrar, mientras que en el reconocimiento de palabras aisladas estos límites son conocidos y pueden usarse para mejora la tasa de acierto y limitar la búsqueda. En segundo lugar, en el habla continua los efectos coarticulatorios entre sonidos son más fuertes que en palabras aisladas y además aparecen efectos coarticulatorios entre palabras, que son más difíciles de predecir. Finalmente, en el habla continua las palabras significativas (nombres, verbos, adjetivos, etc.) suelen enfatizarse, mientras que el resto (artículos, preposiciones, pronombres, conjunciones, etc.) se articula de manera más pobre. Como resultado, las tasas de error se incrementan drásticamente al pasar de palabras aisladas a habla continua.

Talla del léxico

Los reconocedores se pueden clasificar atendiendo al número de palabras de su vocabulario en pequeños, medianos y grandes, según tengan decenas, centenas o más de mil palabras, respectivamente.

Cuando el número de palabras del vocabulario aumenta, el primer problema que aparece es el de la propia confusión entre palabras, que incrementa la tasa de error del sistema. Aunque puede haber vocabularios pequeños o medianos de alta confusión, se considera que el número de palabras fácilmente confundibles crece notablemente cuando el tamaño del vocabulario alcanza alrededor de 1000 palabras.

Por otro lado, en el caso de pequeños vocabularios cada palabra puede modelarse individualmente, ya que es razonable esperar suficientes datos para entrenar cada palabra y es posible almacenar los parámetros de cada modelo de palabra separadamente. Sin embargo, cuando el tamaño del vocabulario aumenta ya no es posible modelar cada palabra explícitamente. En su lugar, se han de usar unidades de decisión inferiores a la palabra (fonemas, sílabas, etc.), que conducen a una degradación del comportamiento del sistema porque son difíciles de detectar y sólo pueden capturar parcialmente los efectos de articulación, que en gran medida quedan absorbidos en los modelos de palabras.

Otra dificultad es la complejidad de la búsqueda. Para pequeños vocabularios, es posible realizar búsquedas exhaustivas. Sin embargo, en grandes vocabularios una búsqueda exhaustiva es inabordable por el excesivo tiempo de cálculo que supone y es necesario recurrir a restricciones, que pueden conducir a errores de reconocimiento.

Restricciones gramaticales

Incertidumbres procedentes de errores acústico-fonéticos pueden resolverse a menudo usando un análisis sintáctico, semántico o pragmático. Cada tipo de conocimiento de alto nivel define restricciones adicionales que las frases deben satisfacer. Si son explotadas convenientemente, estas restricciones pueden sugerir hipótesis plausibles o eliminar interpretaciones improbables en el proceso de reconocimiento.

La dificultad de una gramática, es decir, la cantidad de restricciones impuestas por una gramática, se suele medir mediante la perplejidad: una medida de la incertidumbre media en cada punto de decisión. Las gramáticas de perplejidad baja dan

lugar a mejores tasas de reconocimiento, pero imponen al usuario una sintaxis rígida que dista mucho de la estructura flexible requerida en un proceso de comunicación oral hombre-máquina.

Aumentar la perplejidad de la gramática redunda normalmente en una disminución de la tasa de reconocimiento. Sin embargo, aceptar gramáticas de alta perplejidad es una aspiración importante en reconocimiento del habla, ya que sólo estas gramáticas son altamente versátiles. Para aplicaciones como el dictado automático, por ejemplo, es imposible trabajar con gramáticas de baja perplejidad.

No obstante, algunos formalismos que se han mostrado eficientes en tratamiento del lenguaje natural no son fácilmente aplicables al reconocimiento automático del habla. El problema es que la mayoría de estos formalismos requieren una entrada determinista, mientras que en el caso del reconocimiento del habla la secuencia de palabras es ambigua, a nivel de decodificación acústica y segmentación en palabras, y el proceso de comprensión interviene en la resolución de dichas ambigüedades.

Variabilidad interlocutor

Respecto al grado de variabilidad en los locutores aceptables por el sistema, se distingue entre sistemas monolocutor, si reconocen la voz de un único locutor; multilocutor, si admiten voz de un conjunto limitado de locutores; e independientes del locutor, si admiten voz perteneciente a cualquier locutor.

La diferencia esencial entre los sistemas de reconocimiento multilocutor e independientes del locutor estriba en el conocimiento previo que el sistema posee de los locutores. En un sistema multilocutor, es posible entrenar al sistema con las características de los locutores que componen el conjunto restringido antes citado, de forma que se posee cierta información a priori de los mismos. En un sistema independiente del locutor, en cambio, no es posible incorporar este tipo de información previa sobre los locutores y, normalmente, estos sistemas son evaluados utilizando un conjunto de locutores diferentes de los utilizados para el entrenamiento del mismo.

Los sistemas monolocutor y multilocutor consiguen tasas mejores de reconocimiento que los sistemas independientes del locutor, pero requieren un molesto período de aprendizaje o adaptación para cada nuevo locutor. Los sistemas independientes del locutor no tienen este inconveniente, pero sus tasas de acierto son menores debido a que la mayoría de las representaciones paramétricas de la señal de

voz son altamente dependientes del locutor. En cualquier caso hay aplicaciones específicas para cada tipo de sistema.

Entornos adversos

La inmensa mayoría de los actuales sistemas de reconocimiento del habla se diseñan suponiendo que las condiciones ambientales en que van a funcionar no van a afectar sustancialmente la señal de voz, lo cual supone una simplificación del problema general de reconocimiento. Suponiendo locutores cooperativos, estas variaciones constituyen los mayores factores de degradación del comportamiento de los sistemas de reconocimiento del habla cuando se usan en la práctica.

Por esta razón, el reconocimiento en entornos adversos ha atraído la atención de muchos investigadores en los últimos años y se ha propuesto en la literatura un gran número de técnicas en la dirección de desarrollar sistemas que operen siempre robusta y fiablemente como si hubieran sido entrenados en las mismas condiciones en que se realiza el reconocimiento.Este tema será objeto de un estudio detallado en el apartado 2.2.

2.1.3. DIFERENTES APROXIMACIONES AL RECONOCIMIENTO

La primera cuestión que surge al diseñar un sistema de reconocimiento automático del habla es cómo representar o codificar la misma señal de voz antes de que se realice el reconocimiento propiamente dicho. Esta primera etapa consiste siempre en la conversión de la señal de voz en una sucesión de vectores de parámetros acústicos, regularmente espaciados en el eje temporal, con la información espectral suficiente para poder identificar los sonidos en las siguientes etapas del sistema de reconocimiento. Esta etapa de parametrización de la señal de voz se estudia en detalle en el capítulo 3 de la memoria.

Esta secuencia de vectores es la entrada a la etapa de reconocimiento acústico-fonético, que suele estar basado en técnicas de comparación de patrones, modelos ocultos de Markov, redes neuronales o en métodos "basados en el conocimiento". Sea cual sea la aproximación elegida, previamente a la fase de reconocimiento propiamente dicha es necesaria una fase de entrenamiento en que se aprenden los parámetros de las referencias del vocabulario a reconocer. Todas estas técnicas serán objeto de revisión en este apartado.

Naturalmente, debido a las ambigüedades existentes en el habla, esta etapa de reconocimiento acústico-fonético debe complementarse en la mayoría de las aplicaciones con un modelado del lenguaje. Como se verá, algunas de las anteriores aproximaciones inicialmente ideadas para el reconocimiento acústico-fonético del habla también se han aplicado al modelado de lenguaje. Sin embargo, se escapa al interés de esta memoria la descripción de los escasos sistemas avanzados que intentan integrar el reconocimiento del habla y el tratamiento del lenguaje natural [You88] [Mat89].

2.1.3.1. COMPARACION DE PATRONES

La primera aproximación utilizada en reconocimiento automático del habla fue la de comparación de patrones (*pattern matching*, en la literatura inglesa). A continuación, se revisarán las técnicas básicas de comparación de patrones usadas inicialmente para el reconocimiento monolocutor, de palabras aisladas y con vocabularios pequeños. Finalmente, se describirán las modificaciones que han permitido progresivamente relajar estas restricciones.

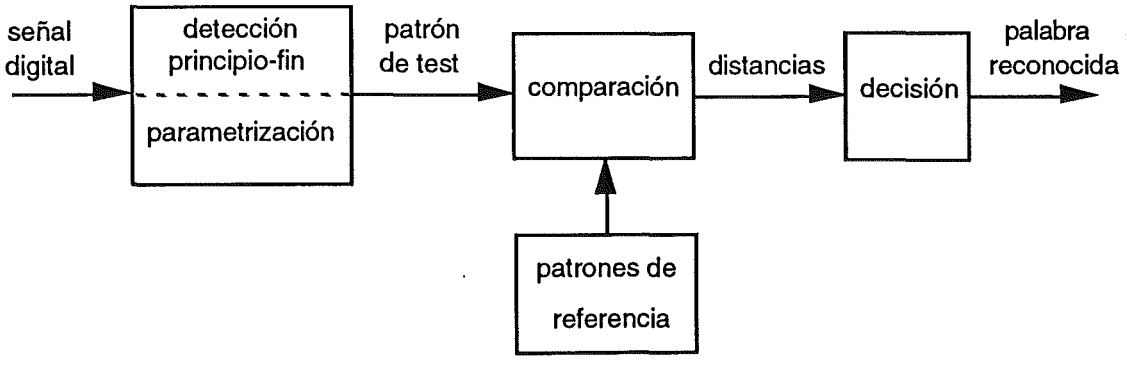

Fig. 2.1. Sistema de reconocimiento de palabras aisladas por comparación de patrones

En la fase de entrenamiento del sistema de la aproximación básica, el usuario pronuncia cada palabra del vocabulario una sola vez. La señal correspondiente es parametrizada a nivel acústico y la secuencia de vectores extraída es almacenada en memoria y etiquetada como patrón de referencia de cada palabra.

En la fase de reconocimiento, la palabra de test pronunciada por el locutor es parametrizada del mismo modo y el correspondiente patrón es comparado con todos los patrones de referencia previamente almacenados en memoria usando una medida de similitud espectral definida usualmente en base a una medida de distancia entre parejas de vectores. Sobre el tipo de distancia entre vectores empleado versa el capítulo 5 de esta memoria.

Debido a la variabilidad intralocutor de la señal, hay diferencias no lineales en la duración de los sonidos incluso cuando el patrón de test es comparado con el patrón de referencia correcto. Por tanto, es necesario realizar un alineamiento temporal de los patrones de test y de referencia de forma que, cuando se realice la comparación con el patrón de referencia correcto, los vectores correspondientes al mismo sonido se correspondan. Este alineamiento se consigue minimizando la distancia total entre los dos patrones, que se calcula como suma de las distancias entre los vectores alineados. Para realizar este alineamiento óptimo se usa el algoritmo de programación dinámica, conocido en el el contexto del reconocimiento del habla como DTW (alineamiento temporal dinámico, *Dynamic Time Warping* en la literatura inglesa), que fue presentado por R. Bellman [Bell57] y aplicado por primera vez al reconocimiento del habla por T. Vintsjuk [Vin68] y G. Slutsker [Slu68] a finales de los 60.

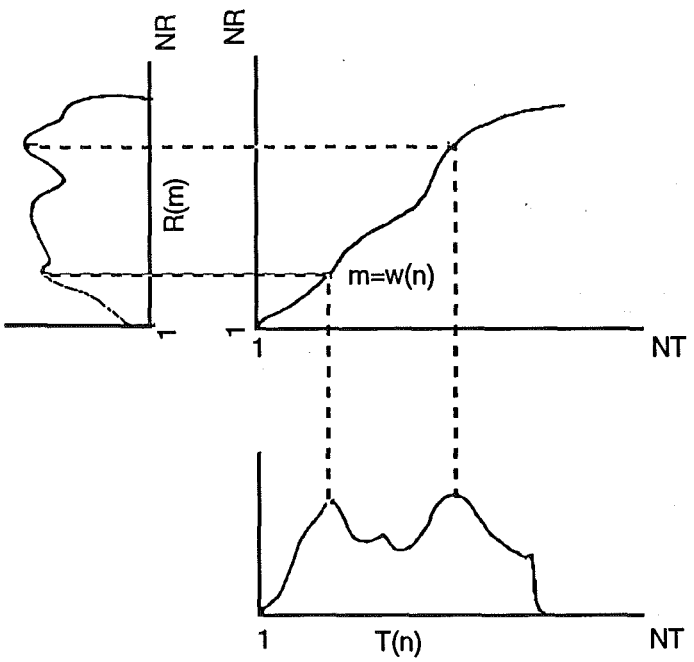

Fig. 2.2. Alineamiento temporal entre patrones de test T(n) y referencia R(m)

El criterio de decisión utilizado se basa en determinar el patrón más similar a la palabra de test, es decir, el que ha proporcionado una menor distancia en la etapa de comparación. Si la distancia menor es demasiado elevada, respecto a un nivel de referencia predefinido, el sistema puede emitir un mensaje de rechazo.

Extensión a varios locutores

La técnica básica de comparación de patrones se ha extendido a sistemas multilocutor e independientes del locutor mediante el uso de varias referencias por palabra. Para ello, cada palabra del vocabulario es pronunciada por varios locutores, con diferentes características propias; después de la etapa de parametrización, se calcula la distancia entre todos las pronunciaciones de una misma palabra usando DTW; finalmente, se aplica un algoritmo de agrupamiento (*clustering*, en la literatura inglesa), tal como el K-medias [Wil85], para determinar los grupos correspondientes a cada tipo de pronunciación para cada palabra y se selecciona el centroide de cada grupo como patrón de referencia del tipo de pronunciación correspondiente. La etapa de reconocimiento se realiza del mismo modo que en el caso monolocutor, opcionalmente con un criterio de decisión más sofisticado, tal como el KNN (K vecinos más próximos, *K Nearest Neighbours* en la literatura inglesa).

Extensión a palabras conectadas

La extensión a reconocimiento de palabras conectadas se ha realizado con métodos que determinan a partir de una secuencia de palabras de test el número y la identidad de cada palabra, así como los límites entre ellas, generalizando el algoritmo DTW de palabras aisladas. Ejemplos de estos métodos son el *Two-Level DP Matching* [Sak79], el *Level Building* [Mye81] y el *One Pass DP* [Bri82]. Para combatir el problema de la falta de contexto que se produce al realizar entrenamiento aislado y evitar el problema de la segmentación, se ha propuesto el entrenamiento conectado [Rab82]. Para mejorar el entrenamiento, también se han propuesto técnicas de entrenamiento multirreferencia para capturar la variabilidad intralocutor semejantes a las descritas en el párrafo anterior para capturar la variabilidad interlocutor.

Extensión a grandes vocabularios

Como ya se ha mencionado, el incremento del tamaño del vocabulario necesita grandes requerimientos de memoria y de cálculo, ya que se precisa almacenar uno o

más patrones de referencia por cada palabra del vocabulario y comparar cada palabra de test con cada referencia. Además, si un locutor ha de entrenar el sistema pronunciando todas las palabras, la tarea rápidamente se hace tediosa. Para combatir estos problemas se han propuesto varios métodos, como la cuantificación vectorial, el uso de unidades subléxicas, la compresión temporal y el reconocimiento en dos pasos.

El método basado en la cuantificación vectorial [Gra84] [Mak85] comienza por aplicar un algoritmo de agrupamiento sobre los vectores de parámetros de una razonable cantidad de voz de un locutor para determinar los grupos de vectores correspondiente a cada tipo de vector, que se representa por su centroide, llamado palabra-código (*codeword*, en la literatura inglesa). El conjunto de palabras-código recibe el nombre de diccionario (*codebook*, en la literatura inglesa). En la fase de entrenamiento, después del procesado acústico de cada palabra, cada vector se identifica con una palabra-código del diccionario. Por tanto, cada palabra, en lugar de estar representada por una secuencia de vectores, está representada por una secuencia de números enteros, también llamados etiquetas, correspondientes a las palabras-código. Esto supone un gran ahorro de memoria. Por otro lado, también se disminuye el coste computacional durante el reconocimiento en el caso de vocabularios grandes, ya que por cada vector de entrada del patrón de test sólo han de calcularse tantas distancias como palabras-código, en lugar de calcular todas las distancias con todos los vectores de todos los patrones de referencia. Además, las distancias entre palabras-código pueden ser calculadas después del entrenamiento y almacenadas en una matriz de distancias.

Estos diccionarios pueden incorporar no sólo información espectral, sino también información de energía o de variación de espectro o de energía con el tiempo. Todo ello puede ser representado por un simple diccionario con supervectores construidos incluyendo los distintos tipos de información [Shi86a]. También puede ser representado por un diccionario diferente para cada tipo de información [Gup87].

Los diccionarios también pueden construirse a partir de voz de varios locutores (diccionarios independientes del locutor) [Lee88a]. Por otro lado, se puede conseguir la adaptación de las referencias de un locutor a un nuevo locutor a través de sus respectivos diccionarios. Para ello, se construyen los diccionarios con las mismas frases pronunciadas por los dos locutores y se crea la correspondencia entre los dos diccionarios realizando un alineamiento temporal de los dos conjuntos de frases [Shi86a] [Bon87].

Otra forma de reducir los requerimientos de memoria es usar unidades de decisión más pequeñas que la palabra, unidades subléxicas. Las palabras se reconocen como concatenación de dichas unidades usando un algoritmo de DTW de "palabras" conectadas. Estas unidades deben escogerse de forma que no estén muy afectadas por los problema de la coarticulación y de los límites, pero no deben ser muy numerosas. Ejemplos de tales unidades son los fonemas [Sug83], difonemas [Mar81], sílabas [Hun80], semisílabas [Rus81] y disílabas [Sin 88]. Otros métodos usan unidades sin afiliación lingüística; por ejemplo, segmentos obtenidos mediante algoritmos como la Cuantificación de Segmentos [Rou82] o la de Matrices [Bur85]. Por otro lado, ha habido intentos de representar las referencias mediante redes de segmentos [Kop85].

La compresión temporal también puede reducir la cantidad de información [Kuh81]. La idea es comprimir los estados estacionarios, que pueden tener longitudes muy diferentes dependiendo de la velocidad del habla, manteniendo todos los vectores durante las transiciones.

Por último, se ha propuesto para reducir el coste computacional el reconocimiento en dos pasos, en el primero de los cuales se utiliza un método poco preciso pero rápido para eliminar los candidatos más improbables. Esta eliminación puede realizarse sumando las distancias de la diagonal de la matriz de distancias usada en el algoritmo DTW [Gau86] o sin ningún alineamiento temporal [Mar87].

Muchos de los progresos relacionados con la aproximación de comparación de patrones se han obtenido para una sola de las extensiones mencionadas simultáneamente. Para superar este problema, se ha debido recurrir a otras aproximaciones.

2.1.3.2. MODELOS OCULTOS DE MARKOV

Los modelos ocultos de Markov, abreviadamente HMM (*Hidden Markov Models,* en la literatura inglesa) representan un nivel superior de abstracción con respecto al método de comparación de plantillas [Rab86a]. Fueron utilizados por primera vez en reconocimiento del habla en CMU [Bak75] e IBM [Jel76]. Desde entonces se han usado con gran profusión en todo tipo de reconocedores (palabras aisladas/habla continua, dependiente/independiente del locutor, vocabulario grande/pequeño, etc.).

En esta aproximación cada referencia se representa por un modelo estocástico. Si consideramos una señal acústica A, el proceso de reconocimiento en una

aproximación estocástica consiste en calcular la probabilidad P(WIA) de que la secuencia de palabras o frase W corresponda a la señal acústica A, y encontrar la secuencia de palabras con mayor probabilidad. Usando la regla de Bayes, esta probabilidad puede escribirse como

$$P(W|A) = \frac{P(W)\,P(A|W)}{P(A)} \qquad (2.1)$$

donde P(W) es la probabilidad de la secuencia de palabras W, P(AIW) es la probabilidad de la señal acústica A dada una secuencia de palabras W, y P(A) es la probabilidad de la señal acústica. Por tanto, es necesario tener en cuenta P(AIW), que es el modelo acústico, y P(W), que es el modelo de lenguaje. Ambos modelos pueden representarse mediante modelos de Markov.

Se considerará en primer lugar el modelado acústico, en el cual se distinguirán dos aproximaciones básicas, la discreta y la continua, y se considerarán algunos problemas relacionados con el entrenamiento, el modelado temporal, la elección de la unidad de decisión y la adaptación al locutor. Por último, se abordará el tema del modelado del lenguaje.

Aproximación discreta

En la aproximación discreta, cada unidad de decisión está representada por un autómata de estados finitos compuesto por un conjunto de estados y un conjunto de arcos que los unen. Asociada al arco que va del estado i al estado j existe una probabilidad de transición a_{ij} que representa la probabilidad de que esta transición tenga lugar. También existe una probabilidad de observación $b_i(k)$ de que un símbolo k de un alfabeto finito pueda ser emitido en el estado i. En algunas variantes, esta probabilidad de observación está asociada a la transición, en lugar de al estado.

Cuando se usa cuantificación vectorial esta distribución de probabilidad de observación, también llamada función de densidad de probabilidad de observación (pdf), es la distribución de probabilidad de las palabras- código.

En un modelo oculto de Markov de primer orden, se supone que la probabilidad de que la cadena de Markov esté en un estado particular en un instante t depende sólo del estado donde estaba en el instante t-1, y que la probabilidad de observación en el instante t sólo depende del estado en que se encuentra en ese instante.

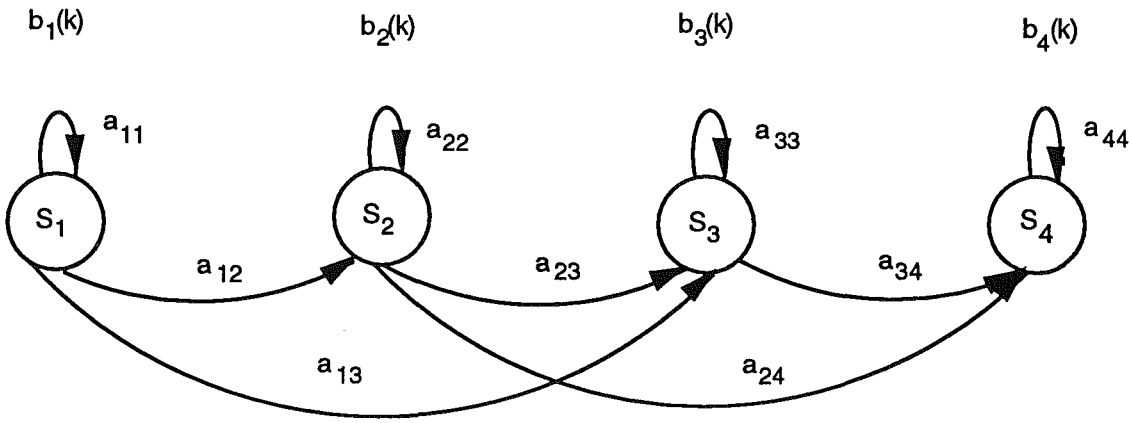

Fig. 2.3. Representación gráfica de un HMM discreto

Aproximación continua

En esta aproximación, la distribución de probabilidad de observación discreta se reemplaza por un modelo continuo. Un modelo usual es la función de densidad gaussiana multivariada [Pau86] o una combinación lineal o mezcla de ellas [Rab86a] [Jua86]. También se ha propuesto el uso de una mezcla de laplacianas [Ney88].

Se han hecho varios intentos de comparar los modelos discretos con los continuos. Parece ser que sólo modelos continuos complejos consiguen mejores resultados que los discretos.

Entre estas dos aproximaciones básicas existen aproximaciones intermedias, como los HMM semicontinuos [Hua89] y el múltiple etiquetado [Nis87], que intentan unir las ventajas de las dos aproximaciones básicas (ver capítulo 5).

Sea cual sea la aproximación utilizada, el número de estados y las transiciones permitidas entre ellos, así como las ligaduras entre estados y arcos y la posible existencia de estados sin observaciones, son elegidos al diseñar el sistema. Una vez elegida la estructura del modelo, hay tres problemas por resolver:

- Evaluación (¿cuál es la probabilidad de que una secuencia de observaciones haya sido producida por un modelo dado?). Se puede resolver usando el algoritmo *Forward* [Bau67], que proporciona la estimación de máxima verosimilitud.

- Decodificación (¿qué secuencia de estados ha producido la secuencia de observaciones?). Puede resolverse con el algoritmo de Viterbi, muy similar al DTW [Vit67].

- Aprendizaje o Entrenamiento (¿cómo obtener los parámetros del modelo dada una secuencia de observaciones?). Puede resolverse con el algoritmo de Baum-Welch [Bau72] cuando el entrenamiento está basado en máxima verosimilitud.

Entrenamiento

Antes de comenzar el entrenamiento, se ha de llevar a cabo la inicialización de los parámetros del modelo. Si existen suficientes datos de entrenamiento, la distribución uniforme es suficiente para unidades de decisión homogéneas, como modelos de sonidos, en el caso de HMM discretos [Lee88a]. Para modelos de palabras, se han de usar técnicas más sofisticadas [Pau88].

La estimación de máxima verosimilitud fue el principio usado inicialmente para el entrenamiento. Esta estimación sería óptima si el proceso de producción de la voz correspondiera realmente con el HMM. Para mejorar el poder discriminativo de los modelos, se han propuesto dos alternativas:

- Entrenamiento correctivo. En este método, en primer lugar los modelos son estimados usando parte de los datos de entrenamiento mediante el criterio de máxima verosimilitud. Luego, estos modelos son usados para reconocer los datos de entrenamiento. Cuando hay un error de reconocimiento, o incluso si un candidato erróneo se acerca al correcto, el modelo inicial es modificado para bajar la probabilidad de los símbolos responsables. El proceso se repite con los parámetros modificados y se para cuando no se observan más modificaciones [Bah88].

- Información mutua máxima. El objetivo en este método, más formalizado que el anterior, es determinar los parámetros del modelo que maximizan la probabilidad de generar los datos acústicos dada la secuencia de palabras correcta, pero al mismo tiempo minimizando la probabilidad de generar cualquier secuencia de palabras errónea [Mer88].

Por otro lado, para obtener buenos resultados, un modelo de Markov necesita muchos datos de entrenamiento. En el caso de los modelos discretos, si un símbolo no aparece nunca en un estado durante el entrenamiento, se dará valor nulo a la probabilidad de la distribución correspondiente a dicho estado; y, si aparece en el reconocimiento, esta probabilidad cero puede ser atribuida a toda la unidad de decisión. La solución a este problema son las técnicas de suavizado. Un método simple de suavizado es dar un valor muy bajo de probabilidad a todas las probabilidades que sean nulas, *floor smoothing* [Lev83]. Otro método es el suavizado de coocurrencias [Lee88b], que suaviza las probabilidades de los símbolos teniendo en cuenta la frecuencia con que aparecen simultáneamente en cada estado del sistema. Más información sobre este tema se puede encontrar en el capítulo 5 de esta memoria.

Para combinar las estimaciones de los parámetros obtenidos con dos modelos diferentes, es necesario aplicar pesos en las diferentes estimaciones. Estos pesos reflejan la calidad de la estimación o la cantidad de información usada para calcular cada uno de ellos. Un método que determina automáticamente estos pesos mediante el algoritmo *Forward-Backward* es el *deleted interpolation* [Jel80].

Modelado temporal

El modelado del tiempo en un modelo de Markov está contenido en las probabilidades de transición. En concreto, la probabilidad de permanecer en un estado dado decrece exponencialmente con el tiempo, lo cual es un pobre modelado temporal en el caso de la señal de voz. Se han hecho varios intentos de mejora en este sentido.

En los *Semi-Hidden Markov Models* [Fer80], un conjunto de pdf's $P_i(d)$ en cada estado i indica la probabilidad de permanecer en un estado durante un tiempo d. Este conjunto de probabilidades es entrenado junto con las probabilidades de transición y observación usando un algoritmo modificado *Forward-Backward*. Un algoritmo más simple es entrenar independientemente la probabilidad de duración y los parámetros del modelo [Rab85a].

Para permitir un modelo más fácilmente entrenable, pueden usarse también para el modelado temporal funciones de densidad de probabilidad continuas, como la distribución de Poisson [Rus85] o la gamma [Lev86].

Otro modo de tener en cuenta indirectamente el tiempo es incluir las características dinámicas de la voz como nuevo parámetro, tanto del espectro como de

la energía. Por ejemplo, en el caso de modelos discretos se han usado múltiples diccionarios con pdf's independientes [Lee88a]

Unidades de decisión

La idea más inmediata es usar un HMM para cada palabra. Un ejemplo de modelo de Markov para una palabra es el de R. Bakis [Bak76]. En ese modelo el número de estados es igual a la duración media de la palabra, medida en número de vectores de la señal parametrizada, y las probabilidades de transición a_{ij} sólo toman valores distintos de cero para $i \leq j \leq i+2$, es decir, desde un estado sólo se puede saltar a él mismo o a los dos siguientes (con ello se intenta modelar los fenómenos de inserción y borrado observados en DTW). Más tarde, se han propuesto con éxito modelos con menos estados [Rab85b].

Sin embargo, del mismo modo que en la técnica de comparación de patrones, el uso de la palabra como unidad de decisión plantea inconvenientes importantes. Como ya se ha mencionado, cuando el vocabulario es grande el uso de modelos de palabras plantea problemas de entrenabilidad y almacenamiento de los modelos, aumenta el coste computacional y la complejidad de la búsqueda.

La solución consiste en la utilización de unidades subléxicas y representar cada palabra por una cadena o red de estas unidades. Es conveniente que estas unidades tengan pocos problemas de coarticulación y detección de bordes y además que el número de ellas no sea excesivo para asegurar su entrenabilidad. En el caso de los HMM, el proceso segmentación por reconocimiento mediante el algoritmo de Viterbi permite eludir el problema de la segmentación previa y permite, por tanto, el uso de unidades tan pequeñas como los fonemas, difonemas, sílabas, semisílabas, etc.

Los fonemas independientes del contexto son interesantes porque son muy poco numerosos [Bah83]. Desafortunadamente, los fonemas están muy afectados por el contexto y los parámetros del modelo de fonema reflejan señales acústicas muy diferentes.

Para resolver el problema anterior se han utilizado los fonemas dependientes del contexto [Bah80] [Sch84]. Si construimos un modelo para cada contexto (*triphone model*), el número de modelos es demasiado elevado y aparecen problemas de entrenabilidad. Puede usarse conocimiento fonético para reducir su número, pues algunos contextos tienen un efecto similar [Der87]. Alternativamente, se han usado

para este cometido medidas de la entropía de los modelos (*generalized triphone*) [Lee88a].

Si el vocabulario es pequeño y el número de repeticiones disponibles por palabra es grande, es posible entrenar fonemas dependientes de la palabra. Esta aproximación ha sido usada por CNET [Dut87] y BBN [Cho86]. En CMU, K.F. Lee ha utilizado modelos de fonemas de ciertas palabras, que el autor llama *function words* [Lee88a]. Se trata de palabras cuyo papel es principalmente gramatical, normalmente cortas y mal pronunciadas, y , por tanto, difíciles de reconocer. Son muy frecuentes en el habla fluida y afectan gravemente al comportamiento global del reconocedor. Sin embargo, precisamente por su frecuencia, pueden entrenarse sin dificultad.

Adaptación al locutor

La adaptación a un nuevo locutor puede realizarse usando técnicas basadas en el mapeado de diccionarios. El método usado inicialmente por BBN realizaba la adaptación cuantificando una frase de entrada desconocida con el diccionario del locutor de referencia y aplicando un algoritmo *Forward-Backward* modificado para calcular la matriz de transformación que representa la probabilidad condicionada de un espectro cuantificado del nuevo locutor, dado un espectro cuantificado del locutor de referencia [Sch87]. El método fue mejorado construyendo el diccionario del nuevo locutor, alineando mediante DTW una frase conocida pronunciada por el nuevo locutor y el de referencia y contando las coocurrencias de sendos diccionarios [Fen88].

Modelado del lenguaje

El modelo de lenguaje puede también representarse como un proceso de Markov. En un modelo bigrama, la probabilidad de una palabra, dada la palabra anterior, se calcula como la frecuencia de secuencias de dos palabras [Bah83]. En un modelo trigrama, se calcula la probabilidad de una palabra, dadas las dos palabras anteriores. Un modelo unigrama es simplemente la probabilidad de una palabra. Un modelo más simple es el modelo de par de palabras (*word-pair model*, en la literatura inglesa), donde se asigna la misma probabilidad a todas las palabras que pueden seguir a una palabra dada.

Los N-gramas deben entrenarse con grandes bases de datos, mayores cuanto mayor es N. Si la base de datos no es lo suficientemente grande y el número de palabras del vocabulario es alto, muchas sucesiones de palabras existentes de hecho no aparecerán y el modelo, especialmente en el caso de los trigramas, tendrá muchas

probabilidades nulas. Esto puede mejorarse utilizando técnicas de suavizado, como la estimación de Turing-Good [Nad85]. También puede usarse *deleted interpolation* para combinar las probabilidades de los unigramas, bigramas y trigramas en el modelo de lenguaje completo [Der86].

Por otro lado, en un modelo biclase o triclase la probabilidad de sucesión de palabras es reemplazada por la probabilidad de sucesión de clases gramaticales [And79]. La probabilidad de una palabra dada en una clase puede usarse para refinar el modelo [Der86]. Una aproximación intermedia es el modelo trigrama suavizado donde se ligan las probabilidades de las palabras largas (tres o más sílabas), ya que son fáciles de reconocer y normalmente no tienen homófonos [Dum88].

La ventaja de los modelos basados en palabras es que pierden menos información sintáctica y semántica. Además se entrenan de una manera muy simple, ya que el texto no necesita ningún etiquetado gramatical inicial. Sin embargo, la cantidad de datos necesaria para entrenar el modelo es muy grande, especialmente en el caso del trigrama. Cuando se usan categorías gramaticales, el texto deber ser etiquetado pero puede ser más corto. Además, si una nueva palabra es introducida en el diccionario puede heredar las probabilidades calculadas anteriormente para las palabras de la misma categoría gramatical.

2.1.3.3. APROXIMACION CONEXIONISTA

En la aproximación conexionista, los datos de referencia son representados como patrones de actividad distribuidos sobre una red de unidades de procesado sencillos. Por su similitud con el funcionamiento del cerebro, a estas redes se les suele llamar redes neuronales y a los elementos de procesado neuronas.

Perceptrones

Los orígenes de esta aproximación se encuentran en el perceptrón, un modelo de percepción visual propuesto por Rosenblatt [Ros59], que fue abandonado más tarde porque se comprobó que no era capaz de sintetizar operaciones sencillas como la XOR. Recientemente, se ha renovado el interés por este sistema debido a que el perceptrón multicapa no presenta esta limitación y tiene, por tanto, mayores capacidades de clasificación que el perceptrón original [Lip87a]; y a la reciente propuesta de un algoritmo para su entrenamiento llamado *Back-propagation* (retropropagación) [Rum86].

Un perceptron multicapa está compuesto de una etapa de entrada, una de salida y una o varias ocultas. Cada etapa está compuesta de varias células, llamadas neuronas. Cada neurona i en una etapa dada está conectada a cada neurona j de la etapa siguiente mediante enlaces, llamados sinapsis, que tienen un peso w_{ij} que puede ser positivo o negativo, según se trate de una sinapsis de excitación o de inhibición. El estímulo es introducido en las neuronas de la etapa de entrada (puestas a 0 o a 1 si el modelo es binario) y se propaga en la red. En cada neurona se calcula la suma de la energía ponderada transmitida por las sinapsis que llegan a ella. Si esta energía supera un umbral T_j, la célula reacciona y, por turno, transmite energía a las neuronas de la capa superior. La respuesta de una neurona a la energía de entrada viene dada por una función de activación o logística f(.).

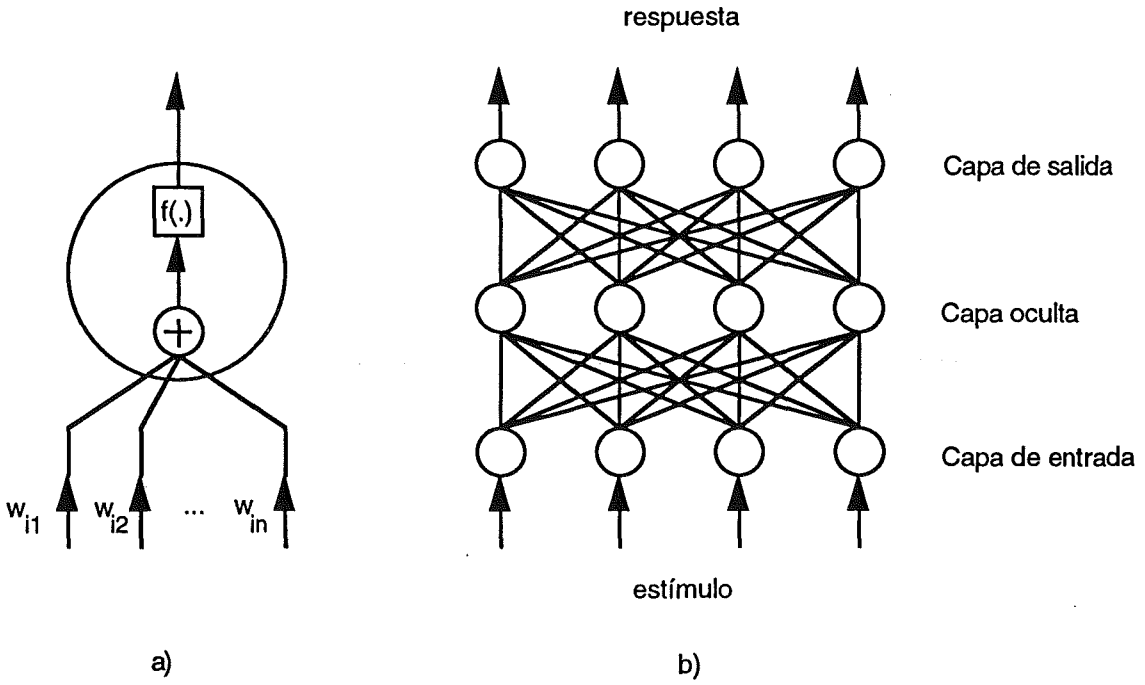

Fig 2.4. a) Neurona. b) Perceptrón multicapa

En la fase de entrenamiento, el estímulo propagado cuando llega a las neuronas de la etapa de salida es comparado con la respuesta deseada, generando una señal de error que es propagada hacia atrás a las etapas inferiores para ajustar los pesos de las

sinapsis y el umbral de excitación de cada neurona. Se itera el proceso hasta que los parámetros de la red alcanzan suficiente estabilidad. Esto se hace para todos los pares estímulo-respuesta.

En la fase de reconocimiento, se propaga el estímulo hacia la etapa de salida. En algunos sistema, la neurona de salida con mayor valor identifica el patrón reconocido. En otros, el vector de valores de las neurona de salida se compara con los vectores que representan cada patrón de referencia con una medida de distancia (p.e. la distancia de Hamming, para neuronas binarias).

Redes neuronales con procesado temporal

Aunque la capacidad discriminativa de los perceptrones es interesante para el reconocimiento del habla, presentan problemas para modelar adecuadamente la evolución temporal de la señal de voz. Para paliar este problema, se han propuesto varias métodos.

Para tener en cuenta la variabilidad de la duración de las señales de voz, la aproximación más sencilla consiste en diseñar una capa de entrada con un número de neuronas suficiente para acomodar la secuencia temporal de mayor longitud [Pee88]. Otra posibilidad es utilizar alguna técnica de compresión lineal o no lineal para acomodar la duración de las secuencias temporales al tamaño de la capa de entrada de la red.

Por otro lado, se puede modelar la información contextual haciendo que la entrada a la red incluya el contexto en que ocurre el estímulo. Esta aproximación da lugar a los llamados perceptrones multicapa contextuales. T. Sejnowski usó este método para la conversión fonema-grafema en inglés [Sej86] y posteriormente se ha aplicado a reconocimiento de voz [Bou 87]. Una aproximación similar es el perceptrón multicapa con retardo temporal [Wai87]. Otra aproximación alternativa al procesado temporal la constituyen las redes neuronales recurrentes [Brid90].

Por último, para englobar las buenas propiedades discriminativas de las redes neuronales y las buenas propiedades de alineamiento temporal de los algoritmos DTW [Sak89] y de Viterbi [Lip87b] [Mon92], se han hecho varios intentos de usarlos en el mismo entorno.

Otros sistemas

Otro tipo de redes neuronales que han tenido gran difusión son los mapas de características o fonotópicos [Koh84], que se basan en la hipótesis de que, para el reconocimiento del habla, la información que está relacionada debería situarse topológicamente próxima, tal como ocurre probablemente en el cerebro humano.

El proceso de creación de un mapa fonotópico es similar a un agrupamiento. La red puede representarse en una rejilla bidimensional, en la que cada nodo representa un prototipo espectral. Cuando se presenta un nuevo espectro de los datos de voz, es comparado con todos los prototipos existentes usando una medida de similitud. Cuando se encuentra el prototipo más cercano, este prototipo es promediado con el nuevo vector, teniendo en cuenta como ponderación el número de espectros que ya han sido promediados con ese prototipo. Los restantes nodos son también promediados con una ponderación que depende inversamente de la proximidad topológica al citado prototipo. Al final del proceso se obtiene un cuantificador, como en un agrupamiento, pero en el cual las palabras-código similares están topológicamente próximas. De este modo, un proceso de reconocimiento corresponderá con una trayectoria en esta red, que previamente se habrá etiquetado reconociendo frases etiquetadas.

Otro ejemplo es la red Hopfield que tiene una sola etapa, estando cada neurona conectada con todas las demás, y que se usa como memoria asociativa y puede restaurar entradas ruidosas. La red Hamming es similar a esta última, pero en primer lugar calcula la distancia de Hamming para comparar el vector de entrada con los patrones de referencia [Lip88].

Las redes neuronales se han utilizado incluso en el modelado del lenguaje; una aproximación consiste en extender los modelos bigrama o trigrama a modelos Ngrama [Nak88]. Sin embargo, aunque la aproximación conexionista parece muy atractiva y prometedora, varios problemas permanecen todavía sin resolver: qué arquitectura debe escogerse, cuántas etapas, cuántas neuronas, cómo tratar el procesado temporal, cuál debería ser la representación de los pares estímulo-respuesta, cómo es posible reducir el tiempo de cálculo,... Hasta el día de hoy ningún experimento definitivo ha probado la superioridad de las redes neuronales sobre los modelos de Markov o la comparación de patrones.

2.1.3.4. METODOS BASADOS EN EL CONOCIMIENTO

La aproximación basada en el conocimiento se hizo muy popular cuando se propusieron los sistemas expertos en Inteligencia Artificial. La idea consiste en separar el conocimiento que va a usarse en un proceso de razonamiento (la "Base de Conocimiento") de la estrategia o mecanismo de razonamiento sobre ese conocimiento (basado en el "Motor de Inferencias", que produce reglas). Dicha estrategia también se refleja en el modo en que la información de entrada (los "Hechos") es procesada y el orden en que las reglas son introducidas.

Esta aproximación implica que el conocimiento se ha de incorporar manualmente, a menos que se encuentre algún procedimiento de aprendizaje automático. A principios de los 80, se estimo que el esfuerzo de obtener suficiente cantidad de conocimiento para el reconocimiento de habla continua independiente del locutor y con grandes vocabularios podía durar unos 15 años.

Sistemas expertos en lectura de espectrogramas

Ante la evidencia de que algunas personas expertas en lectura de espectrogramas, registros de la energía de la señal de voz en función de la frecuencia y el tiempo, consiguen altas tasas de reconocimiento (80-90 %), se han hecho varios intentos de imitarlas mediante un sistema experto basado en el conocimiento [Col80].

El sistema experto dialoga con un "ingeniero cognitivo" (normalmente un informático), que tiene la función de extraer los hechos, el conocimiento y las estrategias con que el experto va a aplicar el conocimiento sobre los hechos. Principalmente, estas aproximaciones están dedicadas al estudio de un conjunto específico de fonemas de un locutor específico [Ste86], o un conjunto de fonemas en un contexto específico para cualquier locutor [Zue86], etc.

Un problema es el hecho de que el experto, antes de aplicar las reglas, usa pistas visuales, que son difíciles de representar por reglas aplicadas sobre símbolos. Una manera de evitar el problema es verificar manualmente todos las características medidas por el sistema [Zue86], o tomar como entrada una lista de características dadas por el usuario.

Otras aproximaciones

Aparte de los sistemas expertos en la lectura de espectrogramas, se realizó en el MIT un trabajo sobre segmentación y etiquetado del habla usando métodos basados en el conocimiento [Zue88]. El proceso de segmentación realiza una representación multinivel llamada "dendrograma". El espectrograma es segmentado en unidades de diferentes niveles, de fino a grueso, siendo el último segmento la frase entera. Este proceso está basado en el cálculo de una medida de similitud entre segmentos adyacentes, usando una distancia euclídea sobre los vectores espectrales promediados de cada región previamente delimitada, y en la unión de los segmentos similares. Usando un clasificador estadístico se obtiene después una celosía de fonemas. La representación léxica tiene diferentes pronunciaciones para cada palabra. El resultado es una celosía de palabras.

Se han realizado intentos de integrar la aproximación basada en el conocimiento con la estocástica, HMM [Hat87]. Otros trabajos tienden a usar arquitecturas de sistemas basadas en el conocimiento más complejas como la estructura *Specialist Society* [Gon88] o la *Expert System Society* [Mor86].

2.2. ESTADO DEL ARTE DEL RECONOCIMIENTO DEL HABLA EN ENTORNOS ADVERSOS

El problema del reconocimiento automático del habla en entornos adversos ha atraído la atención de muchos investigadores en los últimos años [Jua91]. La razón principal es que el comportamiento de los sistemas actuales de reconocimiento, que han sido diseñados suponiendo que las condiciones ambientales en que dichos sistemas van a operar no van a afectar sustancialmente la señal de voz, se degrada sustancialmente cuando las condiciones ambientales son adversas.

En general, tales entornos adversos consisten en la presencia de ruido ambiente (de oficina, de coche,...), en la reverberación de la propia sala y en distorsiones y ruidos introducidos por los transductores y el canal de transmisión (micrófonos, canal telefónico,...). Además, también han de tenerse en cuenta las variaciones en el modo de articular del hablante debidas a su reacción psicológica al entorno ruidoso (efecto Lombard).

Estos problemas constituyen las principales causas de degradación de los sistemas de reconocimiento automático del habla cuando se usan en la práctica. A pesar de que el oído humano es capaz de reconocer el habla en condiciones notablemente adversas, incluso si un sistema de reconocimiento automático del habla funciona razonablemente bien en las pruebas de laboratorio se producen problemas en el "mundo real".

El estudio realizado por Dautrich, Rabiner y Martin [Dau83] únicamente sobre los efectos del ruido ilustran algunos consideraciones clave en el reconocimiento del habla en entornos adversos en general. En este estudio se comprobó que un reconocedor de palabras aisladas entrenado en condiciones limpias (virtualmente libres de ruido) y capaz de alcanzar una tasa del reconocimiento del 95% experimentaba un incremento de un orden de magnitud en la tasa de error cuando las señales a reconocer estaban contaminadas con una SNR (relación señal-ruido) de 18 dB.

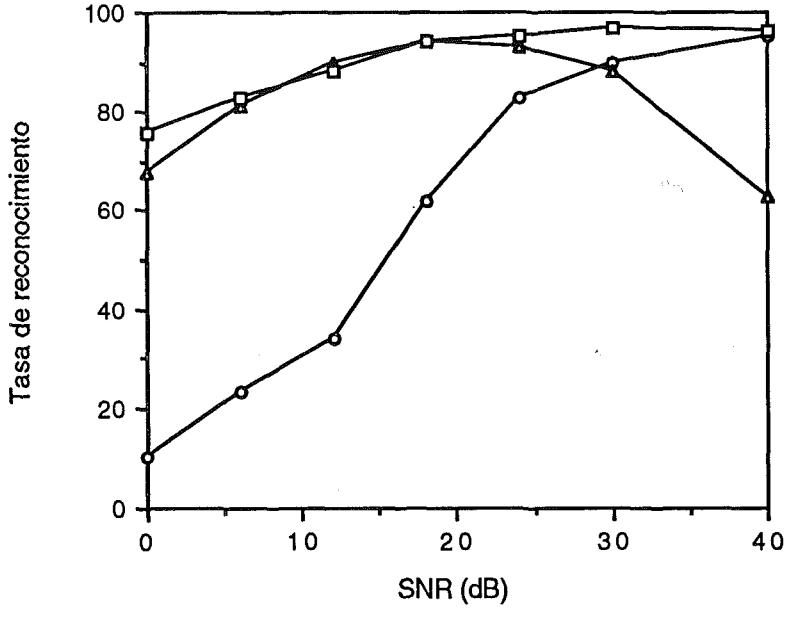

Fig. 2.5. Comportamiento del reconocimiento del habla en entornos adversos

Los resultados de este estudio pueden visualizarse en la figura 2.5. La línea que une los círculos muestra una drástica degradación del comportamiento del reconocedor

cuando se entrena el sistema con señal limpia y la SNR de las señales de test es la indicada en abscisas. Por otro lado, la línea que une los cuadrados refleja una degradación moderada del comportamiento del sistema si se conoce las condiciones ambientales de test, en este caso la SNR indicada en abscisas, y se entrena al reconocedor en esas mismas condiciones. Por último, la línea que une los triángulos muestra la tasa de reconocimiento que se obtiene cuando se entrena al sistema con la SNR indicada en abscisas y la SNR de test es siempre de 18 dB. La evolución de esta curva refleja que la degradación del comportamiento del reconocedor puede reducirse respecto al caso de entrenamiento con señales limpias si este se entrena en condiciones que se aproximen a las de test.

Estos resultados sobre el reconocimiento del habla en presencia de ruido pueden extenderse a otros tipos de entornos adversos: la robustez de un reconocedor de voz frente a los entornos adversos aumenta ostensiblemente si se usan referencias entrenadas en similares condiciones a las de reconocimiento. Incluso en el caso de variaciones del modo de articulación debido al entorno o al estado emocional del locutor, se ha observado una mejora de resultados entrenando el reconocedor con señales de diferentes modos de habla (rápida, lenta, suave, alta, con efecto Lombard,...). Es el llamado entrenamiento multiestilo (*multi-style training*, en la literatura inglesa) [Lip87c]. Se puede concluir, por tanto, que el principal problema es el desajuste entre las condiciones de entrenamiento y reconocimiento, es decir, la variación sufrida por el habla entre ambas fases.

Aunque estos resultados apuntan a una posible manera de mejorar el comportamiento del reconocedor en entornos adversos, el problema estriba en el hecho de que la disponibilidad de datos de entrenamiento que reflejen las condiciones de reconocimiento es raras veces realista (por ser desconocidas estas condiciones, por ser difíciles de obtener, por ser variables en el tiempo,...) y la utilización de referencias no limpias conduce a un comportamiento inaceptable cuando el reconocimiento no se realiza en condiciones adversas.

Por estas razones, se requieren soluciones elaboradas para resolver el problema. En los últimos años, se han propuesto algunos métodos y algoritmos en varias de las etapas del sistema de reconocimiento en la dirección de desarrollar un sistema que opere siempre robusta y fiablemente como si hubiera sido entrenado en las mismas condiciones en que se realiza el reconocimiento. Sin embargo, el reconocimiento del habla en entornos adversos no ha encontrado todavía una solución

satisfactoria incluso en el caso de reconocimiento de palabras aisladas dependiente del locutor y con vocabularios pequeños.

En los siguientes apartados se hará una revisión de estos temas. El apartado 2.2.1 está dedicado a los potenciales entornos adversos y los diferentes modos de articulación dependientes del entorno con que el sistema de reconocimiento puede enfrentarse. En los distintos subapartados del apartado 2.2.2 se revisan los principales técnicas y algoritmos que se han propuesto para tratar este problema.

2.2.1. ENTORNOS ADVERSOS

En este apartado se revisarán los principales fenómenos físicos que provocan entornos adversos para el reconocimiento automático del habla. Se seguirá la distinción clásica entre ruido y distorsión, a pesar de que se ha comprobado que la distinción entre ambas categorías no es clara en cuanto sus efectos en la variación del habla [Fur92] y, por tanto, muchas técnicas para combatirlos sean comunes. En concreto, la distorsión lineal puede considerarse como un ruido aditivo sobre el espectro en escala logarítmica.

Además, también se comentarán las variaciones del modo de articular del hablante debidas a su reacción psicológica al entorno, que como ya se ha mencionado también provoca degradación en el comportamiento del reconocedor.

2.2.1.1. RUIDO

El ruido de ambiente acústico suele considerarse aditivo y es la más importante de las posibles condiciones adversas con que el reconocedor puede enfrentarse. Las fuentes de este tipo de ruido son abundantes.

En una oficina, por ejemplo, la máquina de escribir, la impresora, ordenadores personales o puestos de trabajo que usualmente tienen componentes móviles como discos o ventiladores, el teléfono y la conversación de fondo de otras personas emiten suficiente ruido acústico (45-70 dBA) para degradar el comportamiento del reconocedor de voz.

Dentro de un automóvil, el nivel de ruido debido al motor, el ventilador de la calefacción, el aire, las ruedas y la carretera es todavía más alto.

Por último, hay situaciones, como en la cabina de un avión de combate en vuelo, en que se pueden llegar a niveles de ruido (superiores a 100 dB SPL) tales que la señal es casi ininteligible para el oído humano y menos aún para una máquina.

El espectro del ruido de ambiente acústico no es, en general, plano. En el caso del automóvil, por ejemplo, el motor, las ruedas y el ventilador generan ruido de baja frecuencia, mientras que el ruido debido a los efectos aerodinámicos presenta un espectro plano por encima de 1 KHz [Lec89].

Otros tipos de ruido tales como el ruido eléctrico o el ruido de cuantificación, que naturalmente están presentes en cualquier sistema de reconocimiento moderno, están en general a un nivel por debajo de nuestro umbral de interés. Sin embargo, el ruido debido a los equipos de transmisión y conmutación de la red telefónica sí pueden afectar al comportamiento del reconocedor.

2.2.1.2. DISTORSION

Además de la contaminación aditiva de señales ruidosas, la señal de voz puede sufrir una serie de distorsiones antes de ser registrada y procesada para su reconocimiento. También son abundantes las fuentes de distorsión, tanto lineales cono no lineales.

La habitación en que funciona el sistema de reconocimiento casi seguro que tiene un cierto grado de reverberación que puede alterar el espectro de la señal.

El micrófono empleado, dependiendo del tipo y la colocación, también puede añadir ruido y distorsionar significativamente el espectro de la señal. En este sentido, se ha comprobado que la utilización de distintos tipos de micrófonos en las etapas de entrenamiento y reconocimiento puede provocar graves problemas. Se han propuesta técnicas efectivas basadas en la normalización del vector de parámetros, como la CDCN (*Codeword-Dependent Cepstral Normalization*) [Ace92].

Asimismo, cuando el reconocedor funciona en la red telefónica, el canal telefónico a través del cual se transmite la señal puede causar aún más distorsión sobre la señal. Se han hecho estudios al respecto en que se muestran las características frecuenciales medias del canal telefónico de usuario a usuario y las grandes variaciones entre unos canales y otros. Las grandes variaciones entre canales pueden provocar desajustes de consecuencias desastrosas entre las condiciones de entrenamiento y reconocimiento.

2.2.1.3. EFECTOS ARTICULATORIOS

Muchos factores afectan al modo de hablar del locutor: reacción psicológica al ambiente acústico exterior, estado emocional,... Incluso el simple hecho de ser consciente de que se está estableciendo comunicación con una máquina puede hacer que el locutor produzca diferencias notables en los formantes de los sonidos y el ritmo.

Los cambios articulatorios debido a la influencia del entorno, conocidos como efecto Lombard, pueden tener efectos dramáticos en los resultados de reconocimiento. Se han realizado estudios para modelar estos efectos. Así, por ejemplo, se ha observado que cuando un locutor habla en presencia de ruido el primer formante de una vocal tiende a crecer mientras que el segundo decrece [Pis85], y que la caída espectral decrece en las frecuencias bajas y aumenta en las altas para la mayoría de las vocales y líquidas [Jun90].

Sin embargo, estas variaciones de las características del habla presentan una gran dificultad a la hora de cuantificarlas. El ruido acústico o la distorsión del canal, que usualmente no varían tan rápidamente como la señal misma en términos de características espectrales, pueden ser modelados o medidos hasta cierto punto. En cambio, estos efectos articulatorios constituyen un proceso inherente al proceso de producción de la voz y son dependientes del contexto. Por ello, sólo han sido objeto de caracterizaciones cualitativas no suficientemente específicas para proporcionar soluciones satisfactorias.

2.2.2. TECNICAS DE RECONOCIMIENTO DEL HABLA EN ENTORNOS ADVERSOS

Como ya se ha mencionado, si las características del entorno son conocidas con un cierto grado de aproximación, un sistema de reconocimiento automático del habla que sea entrenado en esas condiciones se comporta en general de forma más robusta que un sistema que use referencias limpias. Sin embargo, se ha argumentado que la disponibilidad de datos de entrenamiento que reflejen las condiciones adversas en que se va a realizar el reconocimiento es raras veces realista. Por tanto, soluciones directas como el entrenamiento ruidoso o multiestilo pueden no resolver el problema del reconocimiento del habla robusto en entornos adversos.

En los apartados siguientes se revisarán algunos métodos y algoritmos que se han propuesto en los últimos pocos años para combatir las a menudo desconocidas, variables y severas condiciones en que el sistema de reconocimiento automático del habla ha de trabajar en la práctica. Naturalmente, no se pretende confeccionar un lista exhaustiva de todos los trabajos que se han realizado en el tema de reconocimiento robusto del habla, pues esta lista sería enorme y no entra dentro de los objetivos de esta memoria confeccionarla. Simplemente, se revisarán las ideas principales en que se inspiran los algoritmos y técnicas que han tenido más resonancia en la literatura dedicada al tema.

En la figura 2.6 se muestran los principales métodos que se han propuesto y que se desarrollarán en apartados posteriores, situándolos a lo largo de la secuencia básica del procesado de la señal de voz que se realiza en el sistema de reconocimiento.

En la exposición, los distintos métodos serán agrupados en mejoras en los transductores, nuevas representaciones de la señal de voz, métodos de preprocesado para la mejora de la señal de voz, enmascaramiento de ruido y modelos adaptativos, medidas de distorsión robustas y compensación de estrés. No existe, sin embargo, una total unanimidad en la literatura sobre la clasificación de los diversos métodos de reconocimiento robusto.

Es importante destacar que aunque se han propuesto muchos métodos para el reconocimiento robusto del habla en entornos adversos, estos métodos no han sido apenas comparados. Es una tarea pendiente recoger bases de datos suficientemente grandes y variadas en entornos adversos (reales o simulados) y comparar con ellas la efectividad de los métodos propuestos en condiciones idénticas. También es necesario determinar claramente las áreas de aplicación de los métodos principales, estudiar la combinación de los mismos y mejorar los más prometedores.

2.2.2.1. TRANSDUCTORES ESPECIALES

Usando un micrófono de gradiente de presión cancelador de ruido en un coche de pasajeros, Dal Degan y Prati [Dal88] confirmaron que la señal recogida está prácticamente libre de ruido si el micrófono se mantiene muy cerca de la boca del locutor y paralelo al frente de ondas. Sin embargo, con sólo unos 10 cm de distancia y 30 grados de giro la potencia de la voz cae 15 dB, lo cual produce una degradación.

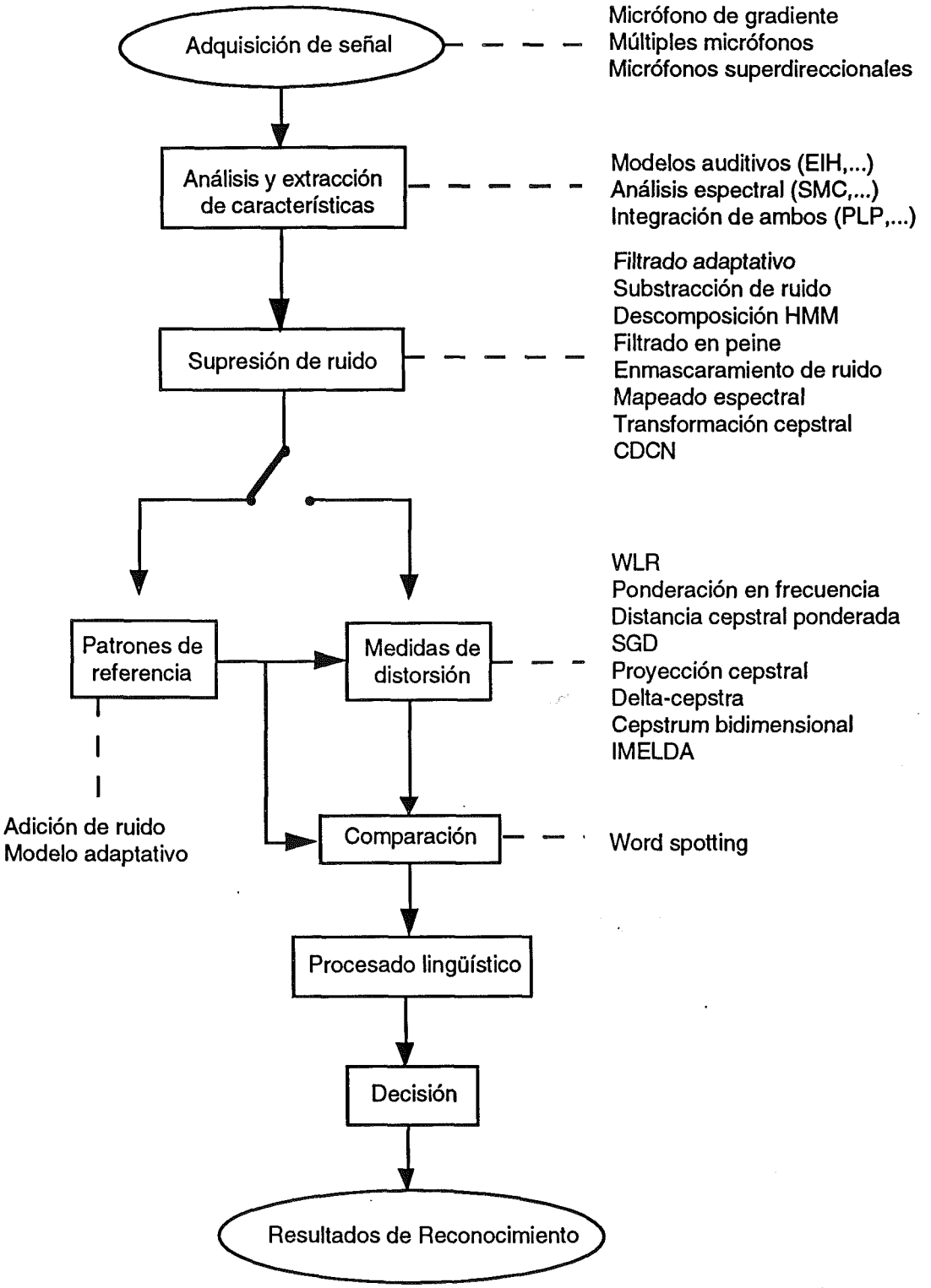

Fig. 2.6. Principales métodos de reconocimiento robusto del habla

Viswanathan y Henry [Vis86] comprobaron que otros tipos de micrófonos de gradiente son efectivos en un ambiente de ruido moderado (95 dB SPL de ruido acústico de banda ancha) si se optimiza y se fija la localización del sensor. Sin embargo, cuando el ruido es severo, como en la cabina de un avión de combate con 105 dB SPL, observaron que el uso de micrófono cancelador de ruido únicamente no proporcionaba resultados de reconocimiento satisfactorios y sugirieron el uso de dos sensores, un acelerómetro para bajas frecuencias, hasta 1.5 KHz, y un micrófono de gradiente para frecuencias superiores. El acelerómetro se coloca en la piel (cerca de la garganta) del locutor para medir las vibraciones de la piel y es insensible al ruido acústico. Otra posibilidad sugerida por los mismos autores es el uso en paralelo para el reconocimiento de las salidas de los dos sensores como vectores de coeficientes espectrales concatenados. Sin embargo, aunque se consiguieron mejoras mediante el uso de los dos sensores, las pruebas se hicieron sólo en el caso de condiciones iguales para entrenamiento y test y, por tanto, los resultados no pueden extrapolarse al caso más general de desajuste entre ambas condiciones.

El uso de micrófonos superdireccionales puede ser útil en el caso de que el ruido provenga de una dirección determinada. Una posibilidad es el uso de un *array* de micrófonos adaptativos para la reducción de ruido, que consiste en un conjunto de micrófonos seguidos cada uno de ellos por un filtro digital. Las características de estos filtros se ajustan automáticamente para asegurar que el array de micrófonos sea insensible al sonido proveniente de la dirección de la fuente de ruido [Kan86]. Cuando el ruido es difuso, sin embargo, como en el caso de coches y aviones, la efectividad de los micrófonos superdireccionales es relativamente baja [Pow87].

Uno de los métodos convencionales para la cancelación adaptativa de ruido usa el algoritmo de filtrado adaptativo para procesar dos señales de entrada [Wid75] [Pow87]. En este método (Fig. 2.7), se aproximan las características de transferencia desde la fuente de ruido al micrófono primario mediante un filtro transversal y se cancela el ruido en el primer micrófono estimado a partir de la entrada del micrófono de referencia. El ruido que debería ser cancelado tiene que ser estimado correctamente y cuanto mayor es la distancia entre los micrófonos primario y de referencia más difícil es estimar el ruido. Sin embargo, si la distancia entre micrófonos es pequeña, es difícil evitar que la señal de voz alcance el micrófono de referencia.

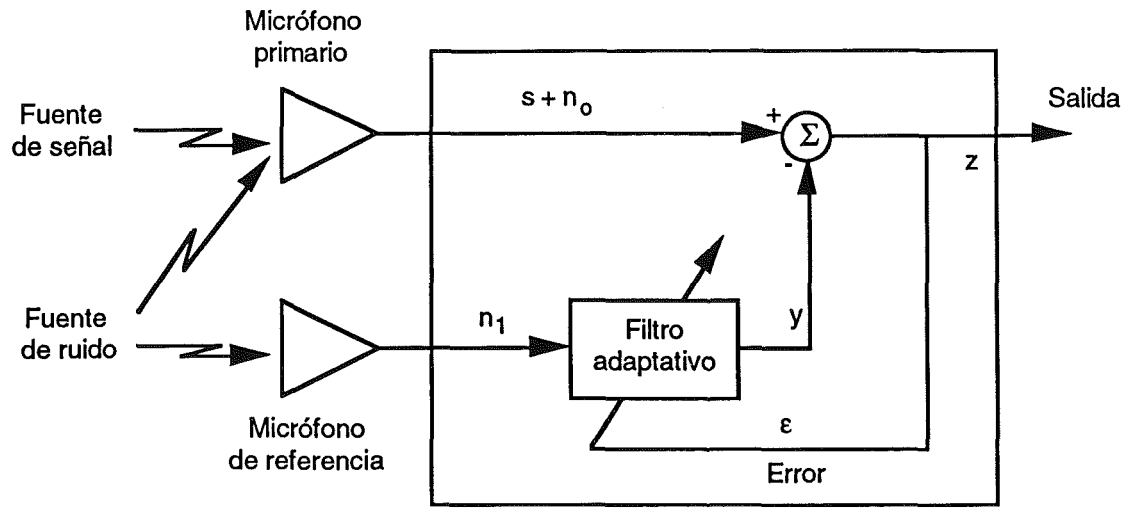

Fig. 2.7. Cancelación adaptativa de ruido con dos micrófonos

La cancelación adaptativa de ruido fue probada en un experimento de reconocimiento del habla en un coche. Para habla con una SNR de 13 a 24 dB, la tasa de reconocimiento era superior que en el caso convencional de un sólo micrófono [Nak90]. Sin embargo, aunque este método puede reducir la componente principal del ruido del motor de coche, una fuente de ruido que se captura fácilmente, es difícil reducir el ruido difuso.

Cuando los reconocedores de habla se usan como controles remoto de equipos acústicos, como aparatos de estéreo o de televisión, el sonido producido por el equipo se convierte en ruido de interferencia. En estas situaciones el ruido es conocido y, por tanto, el filtrado adaptativo es una buena solución. Se ha confirmado que un filtro adaptativo FIR controlado por un algoritmo LMS normalizado puede reducir el ruido de radio o televisión de 15 a 20 dB [Usa91].

Yamada et al. [Yam91] han propuesto un método para eliminar reverberación mediante filtrado inverso de señal de voz reverberante en bandas separadas obtenida a partir de varios micrófonos. Este método se denomina MINT subbanda, ya que la construcción del filtro inverso está basada en el teorema MINT (entrada múltiple/salida inversa) [Miy88] para minimizar la diferencia entre la señal inversa-filtrada y la señal de referencia en cada banda de frecuencias. Con un tiempo de reverberación entre 0.55 y 0.86 s, usando este método con 512 subbandas y dos micrófonos se recuperó la señal de voz original aceptablemente de 0.5 a 8.5 KHz.

Este método ha sido simplificado [Wan91] y aplicado a reconocimiento del habla. Usando un sólo micrófono la reverberación bajó la tasa de reconocimiento hasta el 10%, pero cuando se utilizaron de cinco a siete micrófonos y se aplicó el filtrado inverso la tasa de reconocimiento sólo descendió en un 2-3%.

2.2.2.2. NUEVAS REPRESENTACIONES DE LA SEÑAL DE VOZ

Esta aproximación al problema consiste en encontrar representaciones de la señal del habla que sean invariantes o resistentes a ruidos, distorsiones, etc... En otras palabras, se trata de diseñar una etapa de parametrización del sistema de reconocimiento que extraiga de la señal de voz unas características que sean robustas por sí mismas a variaciones del entorno.

Dentro de esta aproximación, pueden distinguirse claramente dos enfoques: uno deriva del análisis espectral desde el punto de vista de procesado de señal y otro trata de emular la capacidad auditiva humana.

Dentro del primer enfoque, Mansour y Juang [Man89a] han propuesto recientemente una técnica de estimación espectral robusta de la señal de voz, la Coherencia Modificada Localizada (SMC,*Short-Time Modified Coherence*). Este técnica consiste esencialmente en un modelado AR en el dominio de la autocorrelación, basándose en el hecho de que dicha secuencia se ve menos afectada por el ruido que la señal original y conserva la estructura de polos de la misma. En reconocimiento de habla ruidosa a 10 dB de SNR, la representación SMC mantiene una tasa de acierto del 98.3% para el reconocimiento monolocutor de los dígitos, mientras que la predicción lineal clásica sufre una degradación importante llegando a un 39.8% de reconocimiento.

En cuanto al segundo enfoque, es un hecho bien conocido que el sistema auditivo humano es más robusto que cualquier sistema automático no sólo frente al ruido aditivo y las distorsiones en general sino también frente a cualquier factor de variabilidad de la voz (modo de articulación del locutor debido a sus características personales, su estado emocional, la influencia del entorno, etc.). Por tanto, es de esperar que un sistema de reconocimiento del habla sea más robusto a todos estos factores si la etapa de representación de la señal de voz imita las características fisiológicas o psicoacústicas del oído humano.

Basándose en esta premisa, se han realizado varios intentos de modelar el sistema auditivo humano para aumentar la robustez del sistema de reconocimiento. Entre ellos destaca el modelo computacional propuesto por Ghitza [Ghi86], llamado Histograma de Intervalos Conjunto (EIH, *Ensemble Interval Histogram*), que pretende representar el patrón temporal de descarga de las fibras del nervio auditivo. En primer lugar, el modelo EIH separa la señal de voz en la banda de 100-3200 Hz usando 85 filtros que simulan el poder discriminativo en frecuencia de la cóclea; seguidamente, se detectan los intervalos en que la salida de cada filtro excede un cierto umbral; finalmente, se calculan los histogramas de estos intervalos. El histograma conjunto resultante es asimilable a un espectro, pero con las no-linealidades y la resolución no uniforme en frecuencia que son característicos del procesado auditivo humano. Cuando se aplicó a reconocimiento de habla ruidosa utilizando el algoritmo DTW tradicional, se consiguieron significativas mejoras en el caso de voz masculina.

Un intento de combinar ambos enfoques es el análisis de predicción lineal perceptivo (PLP, *Perceptual Linear Predictive*) [Her85], que estima el espectro "auditivo" utilizando tres evidencias psicoacústicas del oído: la resolución espectral en bandas críticas, la curva de ecualización de volumen y la ley de potencia intensidad-volumen. Seguidamente, el espectro auditivo es aproximado con un modelo autorregresivo. La combinación de esta representación con la medida de distancia cepstral ponderada, de la que se hablará en el apartado dedicado a medidas de distorsión robustas, ha resultado efectiva en reconocimiento de habla ruidosa.

2.2.2.3. PREPROCESADO DE MEJORA DE LA SEÑAL DE VOZ

Cuando las condiciones adversas consisten únicamente en ruido aditivo, pueden usarse técnicas para la mejora (*enhancement*, en la literatura inglesa) de la señal de voz antes de aplicar el algoritmo de reconocimiento.

Pueden englobarse dentro de este grupo de técnicas los algoritmos de cancelación adaptativa de ruido basados en la utilización de varios micrófonos, que se han revisado en el apartado 2.2.2.1. Sin embargo, también se han considerado técnicas para la mejora de la señal de voz que no se basan en la existencia de una referencia de ruido simultánea y separada. De ellas hablaremos en este apartado.

Estas técnicas obtienen modelos espectrales mejorados de la voz a partir de señales ruidosas, utilizando algún tipo de estimación de las características del ruido.

Suele considerarse que el ruido es aditivo, incorrelado con la señal de voz, continuo y de banda ancha. Los tonos y los ruidos impulsivos pueden cancelarse, por ejemplo, con filtros *notch*.

Se han propuesto una gran variedad de técnicas [Bol92]. Entre ellas, han tenido gran difusión la substracción espectral de ruido [Bol79]; el filtrado de Wiener, con una estimación MAP del espectro todo-polos [Lim78] [Lim79] o basada en cumulantes [Mas92]; y la estimación por mínimos cuadrados de las componentes espectrales [McA80] [Eph84] [Por84].

Muchas de estas técnicas han sido diseñadas para la mejora de la señal de voz en general, no específicamente para mejorar el comportamiento de los reconocedores en presencia de ruido. Sin embargo, la aplicación a reconocimiento robusto ha proporcionado en muchos casos buenos resultados. Así, por ejemplo, con la técnica de Porter y Boll [Por84] se ha conseguido reducir los errores de reconocimiento a 10 dB de SNR de un 40% a un 10% en el reconocimiento monolocutor de los dígitos.

Un ejemplo de técnica diseñada específicamente para el reconocimiento robusto es la de Ephraim, Wilpon y Rabiner [Eph87]. Estos autores han desarrollado un método que estima iterativamente el nivel de ruido y el modelo espectral todo-polos de la señal de voz limpia localizados minimizando la distancia de Itakura-Saito entre el espectro ruidoso y un modelo aditivo, que es suma del espectro todo-polos de la señal limpia y el nivel de ruido estimados. Por tanto, sean $Y(\omega)$, $\sigma^2/|A(\omega)|^2$ y N las densidades espectrales de potencia de la señal ruidosa, el modelo todo-polos de la señal limpia y el nivel de potencia de ruido, respectivamente, los valores de $\sigma^2/|A(\omega)|^2$ y N se estiman iterativamente minimizando la siguiente expresión:

$$d(Y(w),\sigma^2/|A(\omega)|^2+N) = \int_{-\pi}^{\pi} \left\{ \frac{Y(w)}{\sigma^2/|A(\omega)|^2+N} - \log \frac{Y(w)}{\sigma^2/|A(\omega)|^2+N} - 1 \right\} \frac{d\omega}{2\pi} \qquad (2.2)$$

El término $\sigma^2/|A(\omega)|^2$ resultante es utilizado posteriormente en el reconocedor como medida espectral del habla. Una ventaja de este método es que puede aplicarse tanto a las frases de entrenamiento como a las de test sin un conocimiento explícito del nivel de ruido. La principal limitación consiste en la hipótesis de modelo aditivo. En pruebas de reconocimiento monolocutor de palabras aisladas, la técnica mejoró significativamente la tasa de reconocimiento, desde un 42% cuando se usaron

referencias limpias y señales de test contaminadas con 10 dB de ruido blanco aditivo hasta casi un 70% cuando las referencias y las señales de test ruidosas fueron procesadas con este algoritmo. Una desventaja de este método es, sin embargo, su elevado coste computacional.

Se han propuesto también otras técnicas más sofisticadas para la mejora de la señal de voz, que hacen uso de modelos ocultos de Markov en conjunción con la estimación MAP o por mínimos cuadrados [Eph89], mapeado vectorial probabilístico [Gis90], redes neuronales [Tam88], etc. Sin embargo, muchas de estas técnicas no han sido aplicadas al reconocimiento robusto.

Otro grupo de técnicas integra la mejora de la señal de voz dentro del proceso de reconocimiento. Un ejemplo es la descomposición de la señal mediante modelos ocultos de Markov [Var90]. El habla y el ruido son cada uno de ellos modelados por un HMM, y la señal ruidosa se modela combinando estos modelos. Se utiliza el algoritmo de Viterbi para decidir qué secuencia de estados del modelo compuesto es el más probable. Este método puede ser aplicado no sólo para ruido estacionario sino también para ruido no estacionario, como la voz de otro locutor. También se ha propuesto la integración del filtrado de Wiener dentro del sistema de reconocimiento, tanto en el caso de comparación de patrones [Ber91] como en el de modelos ocultos de Markov [Bea91].

Un caso particular es el efecto *cocktail party*, fenómeno que permite a una persona extraer la voz de un locutor del ruido de fondo. Un modelo de ingeniería de este efecto fue construido usando filtrado en peine [Par76]. En un experimento, usando señal sumada de dos locutores se extrajo el tono de la voz de un locutor utilizando la continuidad temporal del tono y, posteriormente, se extrajo la voz del locutor mediante el filtro en peine controlado por el tono. El método de filtrado en peine ha sido mejorado añadiendo la capacidad de estimar y substraer el sesgo de ruido de cada armónico [Nag88]. Para habla con una SNR de 12 dB, este método incrementó la tasa de reconocimiento de palabras aisladas de un 25% a un 80% en el caso de ruido blanco, y de un 60% a un 95% en el caso de ruido de sala.

Por último, debido a que no es necesario en reconocimiento de habla la obtención de una señal de voz libre de ruido, se han desarrollado métodos de adición de ruido estimado a los patrones de referencia en lugar de reducir ruido en la señal de test [Roe87] [Tak91]. En este sentido, se ha estudiado también la actualización continua de referencias [Dvo91]. Estos métodos tienen la ventaja de estar libres del problema de la

obtención de potencias negativas, que ocurre en algunas ocasiones en los métodos de reducción de ruido cuando se sobreestima el nivel de ruido.

2.2.2.4. ENMASCARAMIENTO DEL RUIDO Y MODELOS ADAPTATIVOS

En presencia de ruido de banda ancha, las regiones del espectro de la voz que tienen un nivel bajo de potencia son las más afectadas por el ruido. Estas regiones más contaminadas de ruido dan lugar a medidas espectrales poco fiables, que provocan que el cálculo de la similitud espectral sea difícil.

Teniendo en cuenta este hecho, Klatt [Kla76] propuso el uso del enmascaramiento de ruido en conjunción con un analizador de banco de filtros. La idea clave de Klatt es que sólo aquellas bandas de frecuencia cuyos niveles de energía son superiores a los niveles de enmascaramiento se usen en el cálculo de la distancia en cada trama. Estos niveles de enmascaramiento se obtienen para cada banda a partir de los niveles de ruido estimados de antemano en un tramo en el que no hay señal de voz. Para cada trama, en cada banda de frecuencias en que el nivel de la señal de test o el de la de referencia es inferior al nivel de ruido de enmascaramiento el nivel de señal tanto en el test como en la referencia es reemplazado por este nivel de enmascaramiento. De esta manera aquellos canales que probablemente están seriamente contaminados de ruido no colaborarán en el cálculo de una distancia euclídea entre espectros.

Este esquema tiene, sin embargo, limitaciones prácticas. Este método supone que el nivel de ruido es estable y que el nivel de ruido de la señal de test no es alto. Si el nivel de ruido es demasiado alto, no puede calcularse una distancia significativa ya que no hay suficientes bandas de frecuencias cuyos niveles sean mayores que el nivel de enmascaramiento.

Debido a ello se han propuesto diversas mejoras [Bri84] [Hol86] al método inicial de Klatt: mantener una estimación actualizada del espectro de ruido, tanto durante el test como en el entrenamiento; marcar separadamente, en lugar de enmascarar, los valores espectrales de los canales de las tramas de test y referencia como voz o ruido de acuerdo con las estimaciones respectivas de ruido; diseñar reglas de cálculos de distancia individuales en las diferentes situaciones de marcaje (voz-voz, voz-ruido, ruido-voz, ruido-ruido); en caso de múltiples referencias, usar el máximo de las estimaciones de ruido para cada canal durante el entrenamiento.

Se han hecho intentos de incorporar el enmascaramiento de ruido en un entorno de modelado probabilístico [Hol86] [Nad88] [Com89]. Se revisará aquí la técnica propuesta por Nadas et al.

El esquema anterior de enmascaramiento de ruido puede formularse analíticamente en un modelado probabilístico. Sean las variables aleatorias X, Y y Z el espectro de la señal limpia, el espectro que modela el ruido y el espectro ruidoso enmascarado, respectivamente. El procedimiento de enmascaramiento consiste en modelar Z como

$$Z = \max (X,Y). \qquad (2.3)$$

La función de densidad de probabilidad de Z, h(.), cuyos parámetros han de estimarse, puede expresarse como

$$h(z) = f(z)\ G(z) + F(z)\ g(z), \qquad (2.4)$$

siendo F(.) y G(.) las funciones de distribución acumulada de X e Y, respectivamente, y f(.) y g(.) , sus correspondientes funciones de densidad de probabilidad. G(.) se supone conocido o bien se estima con datos de ruido.

La pdf h(.) puede considerarse como una versión compensada de f(.), y se usa para aproximar la verdadera pdf ruidosa. De forma similar al diseño convencional de un diccionario de un cuantificador vectorial, pueden diseñarse un conjunto de h(.)'s como las palabras-código compensadas de la voz dada la estimación de ruido G(.).

En el sistema de HMM's discretos de Nadas et al. se realizó una comparación en cuanto al uso de palabras-código limpias, compensadas o ruidosas, usando ruido de oficina. Se asoció la condición limpia a 41 dB de SNR y la ruidosa a 31 dB. Se comprobó que la compensación de ruido no alteraba significativamente el comportamiento del reconocedor para condiciones iguales de entrenamiento y reconocimiento, pero era capaz de reducir los errores en casi un 70%, desde un 32%, usando palabras-código limpias para señales de test ruidosas, hasta un 10% con palabras-código adaptadas.

La técnica de compensación de ruido ha sido también empleada por Roe [Roe87] para adaptar los palabras-código al ruido en el dominio de la autocorrelación. La hipótesis subyacente, a diferencia del modelo de enmascaramiento anterior, es que el

espectro de potencia del habla y el ruido son aditivos y, por tanto, también las autocorrelaciones. Este método, que ha obtenido mejoras en el reconocimiento de habla ruidosa similares a las de Nadas et al., también puede extenderse a la compensación de estrés, como se verá en el apartado 2.2.2.6.

Otra técnica similar es el mapeado espectral, que transforma habla ruidosa en habla limpia mediante reglas de correspondencia. La aproximación propuesta por Juang et al. [Jua87b] consiste en obtener dichas reglas mediante técnicas de cuantificación vectorial y utiliza la teoría de la estimación para obtener la señal y el ruido. Aplicando este método a una señal ruidosa con una SNR de 14 dB se mejoró la SNR en 10 dB aproximadamente. Otra posibilidad está basada en el agrupamiento espectral jerárquico [Shi90]. También se ha estudiado la utilización de redes neuronales para el mapeado no lineal de parámetros espectrales [Sor91].

2.2.2.5. MEDIDAS DE DISTORSION ROBUSTAS

En los últimos años, se han propuesto en la literatura muchos estudios en la dirección de definir una medida de distorsión, o distancia, apropiada para el reconocimiento robusto del habla. El objetivo de estas medidas de distorsión es enfatizar selectiva y automáticamente la distorsión/similitud perteneciente a ciertas regiones del espectro que son menos contaminadas por el ruido.

Debido a que varios de los métodos descritos en apartados anteriores extraen partes estables de los espectros de voz contaminados por el ruido o transforman el habla en representaciones estables, estos pueden ser considerados en sentido amplio como medidas de distorsión robustas. Por ejemplo, puede interpretarse que un esquema de compensación de ruido sobre componentes espectrales independientes, como las obtenidas mediante un banco de filtros, define implícitamente una medida de distorsión robusta, ya que desenfatiza la medida de distorsión en aquellas regiones que están contaminadas por el ruido. Para otras respresentaciones espectrales, el énfasis selectivo sobre ciertas regiones del espectro para abordar el problema del ruido se realiza de múltiples maneras.

La ponderación de las regiones de picos espectrales, que están menos afectados por el ruido, para aumentar la robustez frente al ruido ha sido considerada ampliamente. Evaluaciones experimentales de varias medidas de distancia en reconocimiento de palabras aisladas basadas en predicción lineal y banco de filtros han

confirmado que la medida de distorsión LR ponderada (WLR,*Weighted Likelihood Ratio*) es robusta frente al ruido blanco [Sug82] [Mat86].

También se ha propuesto una medida de distancia similar a la LR ponderada asimétrica con un factor de expansión del ancho de banda en el espectro de ponderación adaptativo al ruido [Soo87].

Sin embargo, el tipo de distancia más empleada por su eficiencia y buen comportamiento es el de las medidas de distancias euclídeas cepstrales ponderadas, que tienen la forma general

$$d = \sum_{n=1}^{N} w^2(n) \; (\; c_n^{(f)} - c_n^{(g)} \;)^2 \qquad (2.5)$$

donde $c_n^{(f)}$ y $c_n^{(g)}$ son los coeficientes cepstrales de los dos espectros comparados. Cuando se utilizan funciones de ponderación w(n) adecuadas, estas distancias han resultado ser ventajosas para el reconocimiento del habla tanto en condiciones limpias como ruidosas [Han86] [Jua87a]. El éxito de estas distancias en reconocimiento de habla ruidosa puede atribuirse al hecho de que el ruido aditivo normalmente afecta a los coeficientes cepstrales de orden bajo y, por tanto, la aplicación de una ponderación que desenfatice estos términos es beneficioso.

La medida de distancia de retardo de grupo suavizado (SGD, *Smoothed Group Delay*) fue propuesta para el reconocimiento en presencia de ruido y distorsiones [Ita87] y no es más que un caso particular de distancia euclídea cepstral ponderada en la que la función de ponderación se define como

$$w(n) = n^s \exp\left(\frac{n^2}{2\tau^2}\right) \qquad s \geq 0 \qquad (2.6)$$

Experimentos con habla ruidosa y distorsionada han indicado que la ponderación más efectiva se obtiene con valores de s entre 1 y 2 y τ alrededor de 5. La tasa de reconocimiento obtenida con ruido multiplicativo y una SNR de 10 dB fue de un 82%, mientras que sin ponderación era sólo de un 62%. Es interesante destacar que con estos valores de los parámetros la función de ponderación SGD es muy similar a la ponderación seno realzado propuesta en [Jua87a].

También se ha propuesto una medida de distancia espectral de retardo de grupo ponderado (WGD, *Weighted Group Delay*), que combina las ventajas de las medidas WLR y SGD y se ha mostrado efectiva en entornos ruidosos [Mat91].

Por otro lado, evidencias tanto analíticas como experimentales indican que el ruido blanco aditivo provoca una reducción de la norma del vector cepstral utilizado en reconocimiento (término del origen excluido) pero deja la orientación del vector más o menos intacta [Man89b]. La reducción de la norma del vector es perjudicial cuando se utiliza la distancia euclídea cepstral. Además, al ser la reducción de la norma del vector función del nivel de ruido, la misma norma del vector puede usarse para facilitar una ponderación no uniforme para cada trama en la distancia acumulada durante la comparación. Este resultado sugiere el uso de una operación de proyección para formular varias medidas de distorsión en el caso de que el sistema sea entrenado en condiciones libres de ruido pero las condiciones de test sean desconocidas. En el trabajo de Mansour y Juang [Man89b], se propuso la siguiente medida de proyección cepstral como una buena elección de compromiso para habla limpia y ruidosa:

$$d(C^{(f)},C^{(g)}) = |C^{(g)}| \left(1 - \frac{C^{*(f)}C^{(g)}}{|C^{(f)}||C^{(g)}|}\right) \qquad (2.7)$$

donde $C^{(f)}$ y $C^{(g)}$ son los vectores cepstrales de referencia y test, respectivamente, y * denota la operación de transposición. En pruebas de reconocimiento en ambiente ruidoso, esta distancia ha mostrado un comportamiento superior a las medidas de distorsión clásicas [Man89b] [Car91], incluso en presencia de efecto Lombard [Jun89].

También se ha mostrado efectiva en reconocimiento de habla ruidosa la distancia IMELDA (*Integrated MEL-scale representation using Linear Discriminant Analysis*), basada en la técnica de análisis discriminante en combinación con una parametrización mel-cepstrum [Hun89].

Por último, también ha resultado efectivo en reconocimiento robusto del habla el uso de la evolución temporal de los parámetros de reconocimiento. En concreto, si se usan como parámetros los coeficientes cepstrales, han resultado muy efectivos los delta-cepstra, definidos como los coeficientes de regresión de las funciones temporales de los coeficientes cepstrales sobre un intervalo de unos 50 ms [Fur86] [App90]. En la misma línea están enmarcados los trabajos de Hirsch [Hir91] y Hermansky [Her91] sobre el filtrado de los componentes espectrales. También se ha conseguido

robustez frente al ruido usando el análisis cepstral bidimensional en los dominios de la frecuencia y el tiempo [Ari88].

2.2.2.6. COMPENSACION DE ESTRES

El objetivo de las técnicas de compensación de estrés es compensar los cambios espectrales que ocurren como consecuencia de los efectos articulatorios debidos al esfuerzo que realiza el locutor como reacción a los entornos adversos. Debido a las dificultades de modelado de tales cambios, no hay estudios analíticos disponibles. Sin embargo, se han propuesto varias técnicas heurísticas que destacan por su efectividad.

El sistema de reconocimiento de Roe [Roe87] está basado en técnicas de comparación de patrones e incorpora cuantificación vectorial y programación dinámica (DTW). La técnica de compensación de estrés consiste en el mapeado del diccionario de patrones libres de estrés sobre un nuevo diccionario compensado. Para obtener las reglas de correspondencia entre ambos diccionarios, se ha de disponer de las mismas palabras pronunciadas en las dos condiciones: sin y con ruido inyectado mediante auriculares al locutor. En primer lugar, se calcula el cuantificador vectorial para el habla libre de estrés; seguidamente, se obtienen los grupos de espectros de habla con estrés correspondientes a cada palabra código de este diccionario libre de estrés alineando temporalmente, sin cuantificación vectorial, las dos pronunciaciones de cada palabra; por último, se calculan las palabras-código compensadas promediando estos grupos en el dominio de la autocorrelación. Además, se suma un espectro de ruido a las palabras-código del nuevo diccionario para combatir el ruido aditivo. En pruebas de reconocimiento de palabras aisladas, este esquema de compensación de estrés redujo la tasa de error de 29.9% a 9.6%, cuando el estrés era causado por el ruido en un coche circulando a 60 mph con el ventilador de la calefacción en funcionamiento.

Otra técnica de compensación de estrés, que opera en el dominio cepstral, es la propuesta por Chen [Che87]. La hipótesis básica de esta técnica es que la distorsión espectral introducida por esfuerzos del locutor inusuales puede compensarse mediante una simple transformación lineal del cepstrum. Aunque esta hipótesis sea indudablemente dura, se observó que las medias y las varianzas de los vectores cepstrales que definen las probabilidades de observación de un modelo oculto de Markov muestran una modificación sistemática en varios estilos de habla. Por tanto, el modelo de palabra compensado puede construirse desplazando las medias y escalando las varianzas del modelo original de acuerdo con las modificaciones observadas. En

reconocimiento monolocutor de palabras aisladas, incluyendo seis estilos diferentes de habla, los modelos compensados redujeron la tasa de error de 25.9% a 16.4%.