Información Acentual para el Reconocimiento Automático del Habla

por

Diego H. Milone

Memoria de Tesis presentada al Departamento de Electrónica y Tecnología de Computadores de la Universidad de Granada, como requisito para obtener el grado académico de

Doctor en Ciencias

Departamento de Electrónica y Tecnología de Computadores

Universidad de Granada

Granada, marzo de 2003

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Información acentual para el reconocimiento automático del habla"
Departamento de Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, 2003.

Universidad de Granada

Departamento de Electrónica y Tecnología de Computadores

Antonio J. Rubio Ayuso
Catedrático de Teoría de la Señal y Comunicaciones

CERTIFICA:

Que la presente memoria titulada "**Información acentual para el reconocimiento automático del habla**" ha sido realizada por **Diego H. Milone** bajo mi dirección en el Departamento de Electrónica y Tecnología de Computadores de la Universidad de Granada. Esta memoria constituye la Tesis que Diego H. Milone presenta para optar al grado académico de Doctor en Ciencias.

Antonio J. Rubio Ayuso
Director de la Tesis

Granada, marzo de 2003

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Información acentual para el reconocimiento automático del habla"
Departamento de Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, 2003.

Dedicado a Cecilia, Marcos y Malena

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Información acentual para el reconocimiento automático del habla"
Departamento de Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, 2003.

Dedicado a Umberto, Ana y Amorina

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Información acentual para el reconocimiento automático del habla"
Departamento de Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, 2003.

Reconocimientos

Grupo de Procesamiento de Señales y Comunicaciones, Facultad de Ciencias, Departamento de Electrónica y Tecnología de Computadores, Universidad de Granada.

Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos.

Cátedra de Computación II, Departamento de Matemática e Informática, Facultad de Ingeniería, Universidad Nacional de Entre Ríos.

Cátedra de Bioingeniería I, Departamento de Bioingeniería, Facultad de Ingeniería, Universidad Nacional de Entre Ríos.

Departamento de Informática, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral.

Diego H. Milone

Departamento de Electrónica y Tecnología de Computadores
Granada, marzo de 2003.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Información acentual para el reconocimiento automático del habla"
Departamento de Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, 2003.

Información Acentual para el Reconocimiento Automático del Habla

Diego H. Milone

Director de la Tesis: Antonio J. Rubio Ayuso
Departamento de Electrónica y Tecnología de Computadores, 2003

A lo largo del tiempo, los sistemas de reconocimiento automático del habla se han ido beneficiando de la incorporación de numerosos aspectos relacionados con la producción y la percepción natural del habla. Aún lejos de alcanzar las habilidades humanas en el reconocimiento del habla, actualmente se sigue incorporándoles más y más conocimientos acerca del habla natural. Los rasgos prosódicos, y en particular la acentuación, forman parte de un gran grupo de conocimientos acerca del habla que aún no se utilizan en forma explícita para el reconocimiento automático. En esta Tesis se realiza un estudio de la relación entre las tres manifestaciones físicas más importantes de la prosodia y la acentuación en el discurso continuo. En base a estos estudios se diseña un sistema para obtener de forma automática la acentuación a partir de la señal de voz. Luego, esta información es utilizada para mejorar el rendimiento de un sistema de reconocimiento automático del habla en discurso continua. La incorporación de esta información acentual se realiza a través de los modelos de lenguaje y los resultados finales muestran una significativa reducción del error de reconocimiento en un corpus de habla en español.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. H. Milone; "Información acentual para el reconocimiento automático del habla"
Departamento de Electrónica y Tecnología de Computadores, Facultad de Ciencias, Universidad de Granada, 2003.

Índice general

Índice de tablas

Índice de figuras

Prefacio

El reconocimiento automático del habla (RAH) ha experimentado un fuerte desarrollo en las últimas décadas. Actualmente existen algunos sistemas comerciales capaces de reconocer el habla de forma automática utilizando un simple ordenador personal. Esto ha motivado que algunos investigadores abandonen prematuramente el RAH. Si bien ya no es un terreno virgen, como lo fue a fines de los 80, está claro que queda mucho por hacer cuando se observan los resultados que se pueden obtener actualmente para habla espontánea y condiciones ambientales naturales. En este sentido, se han publicado algunos trabajos muy motivadores donde se comparan las capacidades para RAH de "humanos y máquinas" [Lippmann, 1997] y se argumenta que las investigaciones en RAH han caído en un "mínimo local", donde solamente se realizan pequeñas —aunque costosas— adaptaciones de un modelo básico y para escapar es necesario explorar nuevos paradigmas aceptando que inicialmente aumenten los errores de reconocimiento [Bourlard et al., 1996].

Si se considera que en todo sistema de RAH se utiliza la señal acústica de la voz como punto de partida, se podría pensar que en forma implícita todas las características del habla son tenidas en cuenta. Sin embargo, la experiencia ha mostrado que la incorporación explícita de la información contenida en el habla a diferentes niveles de análisis, favorece el rendimiento de todo sistema de RAH. Es así como históricamente se han ido considerando progresivamente más y más características del habla. Los sistemas actuales de RAH incorporan muy diversos niveles de análisis del habla, desde el fonético hasta el gramatical. Los rasgos prosódicos se encuentran en uno de los niveles de análisis que aún no se ha integrado completamente al RAH. En particular, la acentuación es una característica importante de nuestra lengua cuya incorporación explícita en el RAH aún no se ha investigado profundamente. En esta memoria de Tesis Doctoral se incluye la descripción de un conjunto de investigaciones dirigidas en este sentido y se presenta un sistema de RAH donde se ha incorporado con éxito la información prosódica y acentual.

Se ha realizado un gran esfuerzo para que esta memoria quede conceptualmente autocontenida y formalmente detallada. El lector podrá acceder a un amplio rango de profundidad en tratamiento de los temas. En este sentido se han hecho pocos presupuestos en cuanto a los conocimiento previos

del lector y se intenta dejar claro tanto los aspectos más elementales como los más complejos.

Se ha dividido esta memoria en 6 capítulos y 2 apéndices. En el primer capítulo se expone un revisión general de los conocimientos actuales relacionados con el RAH. En su última sección se expone la motivación principal de la Tesis Doctoral y se realiza un primer análisis del problema de la incorporación de información acentual al RAH. Sin embargo, este capítulo tiene una finalidad introductoria y no contiene detalles acerca de muchas de las técnicas mencionadas.

En el segundo capítulo se explican con mayor detalle las técnicas utilizadas en esta Tesis. En general este capítulo no posee aportes novedosos salvo, claro está, por el enfoque particular que el autor propone en las exposiciones.

El Capítulo 3 trata sobre las relaciones entre prosodia y acentuación en el discurso continuo. Es importante conocer inicialmente cómo se presenta esta información en el idioma de estudio, para luego buscar sistemas automáticos de análisis y extracción de características. Para ello se ha realizado un estudio sobre un corpus de habla en español con frases leídas y es en este capítulo donde se presentan los resultados obtenidos.

En el Capítulo 4 se describen diversos experimentos realizados con el fin de obtener un sistema automático que relacione prosodia y acentuación en habla continua. Los resultados de este capítulo se utilizan como punto de partida para la incorporación de información acentual a un sistema de RAH. A partir de la señal acústica de la voz se obtienen los rasgos prosódicos y con éstos las estructuras acentuales de cada frase.

Es en el Capítulo 5 donde se describe el sistema de referencia y un método según el cual, a partir de la información acentual, se imponen restricciones estructurales que favorecen el rendimiento de un sistema de RAH. En este capítulo también se presentan los resultados finales de la Tesis. Para terminar, en el Capítulo 6 se exponen las conclusiones y un resumen de los aportes originales.

Capítulo 1

Introducción

En esta introducción se presentan los conocimientos básicos que permiten contextualizar científica y tecnológicamente el trabajo que se desarrollará en los capítulos siguientes. Considerando la comunicación oral como el marco donde se realiza el reconocimiento del habla, se tratan en primer lugar las etapas a través de las cuales se puede modelar este proceso entre seres humanos. A partir de estas etapas es posible comenzar a analizar al habla desde tres perspectivas: desde dentro del hombre a través de los mecanismos de fonación y percepción, desde fuera a través del estudio de su organización estructural y desde dentro de los ordenadores por medio de los modelos que se han utilizan con mayor éxito en el contexto del reconocimiento automático del habla. Estas tres perspectivas guían las tres secciones centrales del capítulo. La última sección está dedicada a la presentación del problema de incorporar los rasgos prosódicos y la acentuación a un sistema de reconocimiento automático del habla. Para terminar se presentan resumidamente los objetivos de la Tesis.

En este capítulo se han dejado de lado algunos formalismos con la intención de motivar y acercar al lector más rápidamente la problemática del reconocimiento automático del habla. Las principales técnicas en que se basó la presente Tesis serán abordadas con mayor detalle en el Capítulo 2.

1.1. El lenguaje y el habla

A través del lenguaje nos hemos diferenciado definitivamente de cualquier otro ser vivo en la tierra. Los estudios acerca de la evolución en nuestra especie han mostrado que las áreas del cerebro asociadas al lenguaje se vieron notablemente expandidas a partir del *Homo ergaster* y con relación al *Australopitecus*. Los humanos poseemos así la muy especial capacidad de hablar y transmitir de esta forma cualquier tipo de información a nuestros semejantes. De este modo el habla se constituye como una de las manifestaciones más compleja y antigua de la inteligencia humana. Veamos como se realiza este proceso en el ser humano y cuánto podemos aprender de ello para luego diseñar sistemas de reconocimiento automático del habla (RAH).

1.1.1. El ser humano bajo estudio

Existe un modelo comúnmente aceptado para ilustrar el proceso de la comunicación oral [Rabiner y Juang, 1993]. En este modelo intervienen dos humanos, uno como emisor y el otro como receptor. A partir de alguna idea o abstracción mental, el locutor genera un mensaje hablado y lo transmite por medio de ondas sonoras. El oyente capta estas ondas sonoras e interpreta o decodifica el mensaje para recuperar la idea original. Entre la idea original en la mente del emisor y la idea recuperada por el receptor se ponen en juego muchos mecanismos que confieren una estructura muy particular al mensaje. Todos estos mecanismos y estructuras han sido materia de estudio para las más diversas ramas de la ciencia. Consideremos el punto de vista desde el que cada disciplina ha estudiado el fenómeno, para luego retomar estos conocimientos desde la ubicación de un diseñador.

Desde el campo de la *biología* se han estudiado tanto las estructuras anatómicas [Rouvière y Delmas, 1988a] como los procesos fisiológicos [Cingolani y Houssay, 1988a] encargados de la generación y comprensión del mensaje. En este sentido se comprende principalmente al aparato fonador y al sistema auditivo, estudiando las regiones del cerebro relacionadas con el lenguaje, las vías eferentes que controlan los músculos del aparato fonador, el aparato respiratorio y el tracto vocal, las distintas partes del oído, la transducción mecánico-nerviosa y las vías auditivas aferentes.

En *lingüística* se estudia principalmente la estructura del mensaje, despojándolo de los mecanismos que lo han generado. En este sentido, la fonética y la fonología [Quilis, 1993] estudian los sonidos elementales de una lengua tanto en lo que respecta a su acústica como a su función en el sistema de comunicación. Pero hasta aquí no se considera el significado que transmiten

estos sonidos y los símbolos asociados. La gramática [Llorach, 1999] estudia esto desde una perspectiva más amplia donde se considera también la sintaxis, la semántica e incluso la pragmática de las palabras que componen un mensaje.

1.1.2. Imitando al ser humano

Si los pájaros baten las alas al volar, ¿por qué no lo hacen los aviones?

Esta pregunta plantea interesantes discusiones a la hora de diseñar sistemas que pretenden realizar tareas que el ser humano ya bien sabe hacer. En primer lugar, como cualquier principio de "camino medio" lo indicaría: no es necesario que los aviones batan las alas al volar pero seguramente deberán contar con un par de ellas. También surgen naturalmente las cuestiones acerca de la imposibilidad de que nuestra inteligencia pueda lograr abarcarnos completamente en una investigación introspectiva. Nuevamente, sin pretender desarrollar una discusión en el terreno filosófico, debemos reconocer que cada una de las partes intervinientes en el modelo de la comunicación oral entre humanos ha dado origen a algún avance en el RAH, pero se han aplicado también muchas otras buenas ideas a partir de principios algo alejados de este esquema.

Ahora sería necesario modificar el modelo de la comunicación oral de forma de incluir una máquina como receptor. Existen dos enfoques que nos posibilitan esta incorporación. El primero, el más directo, es encontrar un modelo del receptor y reemplazarlo en una simulación de su funcionamiento. Sin embargo, existe una visión muy interesante en un segundo enfoque: estamos interesados en recuperar la idea original del emisor y por lo tanto pretendemos "invertir" el proceso llevado a cabo por éste. Si queremos invertir el proceso según el cual la idea se convirtió en mensaje hablado, entonces necesitaremos un modelo inverso del emisor y, antes que esto, un modelo del emisor en si mismo. Volviendo a aplicar algún tipo de regla de punto medio, construiremos un modelo que posee tanto partes del emisor como del receptor humano. Con el tiempo, en los sistemas de RAH se han modelado cada vez más partes no sólo del receptor, sino también del emisor. Es bajo esta concepción que se ha dado origen a la idea central de la presente Tesis.

Los modelos de receptor y emisor se han construido fundamentalmente a partir de abstracciones matemáticas. Diversas ramas de la matemática aplicada, la física y la informática han contribuido a la formalización y puesta en marcha de estos modelos como sistemas de RAH. En el área del *procesamiento de señales* se han desarrollado muy diversos métodos para extraer

la información de la señal de voz. En el área del *reconocimiento de patrones* se han propuesto técnicas para agrupar los datos y obtener prototipos que forman parte de los modelos de la comunicación oral. En *teoría de la información y las comunicaciones* se han establecido las formas de estimar los parámetros que definen el funcionamiento de estos modelos, en general, desde una perspectiva probabilística.

Como el lector puede observar, las investigaciones sobre las que se basa el RAH pueden agruparse en tres grandes categorías desde el punto de vista epistemológico: el estudio centrado en el ser humano, el estudio del mensaje en si mismo y el estudio de las herramientas para la construcción de un modelo. En las próximas secciones revisaremos brevemente cada una de estas categorías y luego haremos hincapié en las fallas del modelo actual para presentar finalmente la idea central de esta Tesis.

1.2. Percepción y fonación

El ser humano posee básicamente dos sistemas relacionados con el habla. Ambos realizan transducciones inversas: el aparato fonador convierte en ondas mecánicas la información codificada en estímulos nerviosos; el oído convierte las ondas mecánicas del sonido en estímulos nerviosos. A continuación se revisarán brevemente las estructuras anatómicas y los procesos fisiológicos que intervienen en estas transducciones.

1.2.1. Anatomía del órgano de la audición

El órgano del oído es separado anatómicamente en tres partes: el oído externo, el oído medio y el oído interno (véase Figura 1.1). En el oído externo y en el medio se realizan transducciones puramente mecánicas del sonido para presentar esta información al oído interno. En el oído interno es donde se realiza la transducción mecánico-nerviosa y la codificación fisiológica del sonido.

Laberintos del oído interno

El oído interno comprende un *laberinto óseo*, compuesto por varias cavidades comunicadas entre sí y un *laberinto membranoso* formado por cavidades de paredes membranosas, contenidas dentro del laberinto óseo.

Del laberinto membranoso nacen las vías nerviosas acústicas y vestibulares. Las cavidades del laberinto membranoso están llenas de un líquido llamado endolinfa y, como el laberinto membranoso no llena completamente

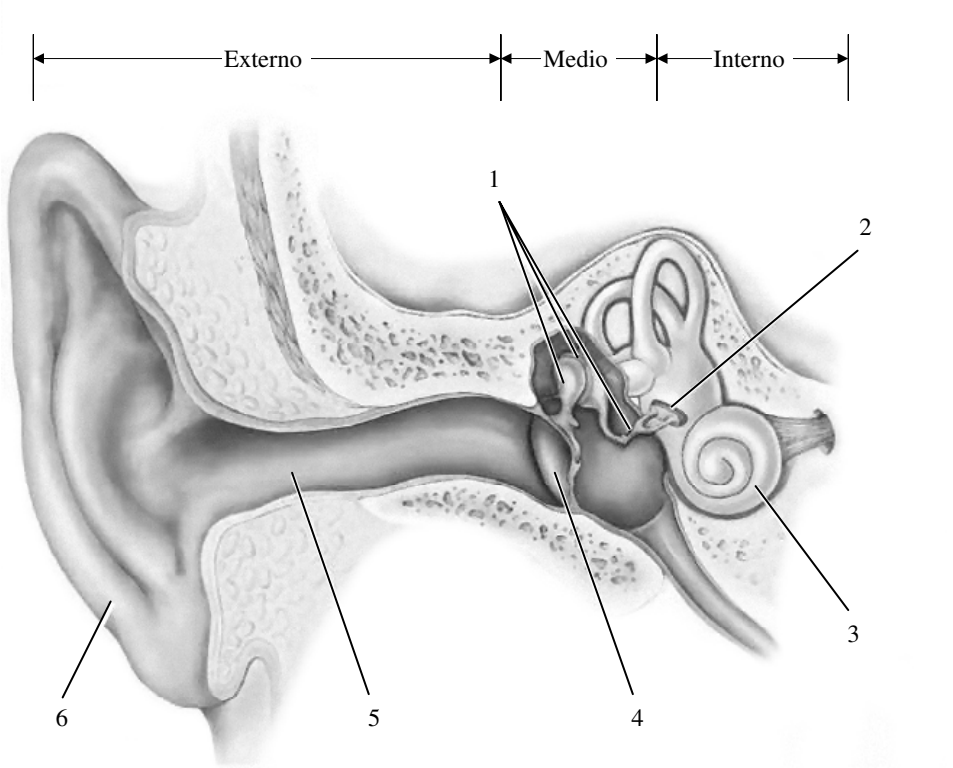

Figura 1.1. Ilustración donde se pueden observar las tres partes del oído. 1: Cadena de huesecillos del oído medio; 2: Ventana oval; 3: Cóclea; 4: Membrana timpánica; 5: Conducto auditivo externo; 6: Pabellón de la oreja.

al óseo, el espacio que deja está lleno de un líquido similar a la endolinfa llamado perilinfa. Ambos laberintos constan de tres partes: el vestíbulo, los conductos semicirculares y el caracol. Los dos primeros no son relevantes para el estudio de la audición; se centrará la descripción en el caracol o cóclea. En la parte ósea del laberinto de la cóclea se pueden discriminar tres estructuras fundamentales: el tubo óseo del caracol, la columela y la lámina espiral.

El tubo óseo del caracol es un conducto enrollado que describe un poco más de dos vueltas y media alrededor del eje cónico que conforma la columela. La lámina espiral divide a este tubo en dos rampas denominadas vestibular y timpánica. La rampa vestibular está situada por arriba de la lámina espiral y se comunica con la cavidad vestibular que luego se conecta a través de la ventana oval con la base del estribo. Éste es el último de una

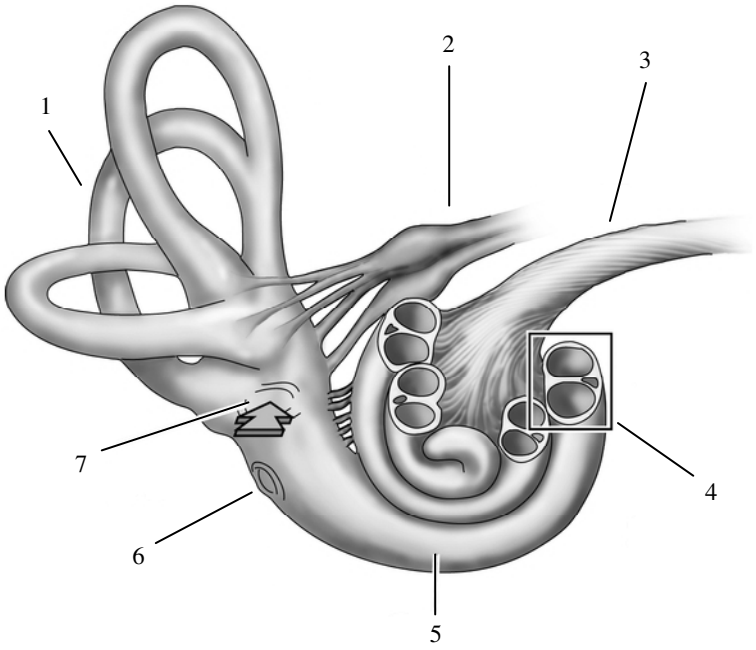

Figura 1.2. Laberinto del oído interno. 1: Sistema vestibular; 2: Nervio vestibular; 3: Nervio auditivo; 4: Sección del caracol óseo y membranoso; 5: Cóclea; 6: Ventana redonda; 7: Ventana oval.

cadena de tres huesecillos que están encargados de conducir las vibraciones mecánicas del sonido en el oído medio. La rampa timpánica está situada por debajo de la lámina espiral y se comunica con la cavidad subvestibular que luego, por medio de la ventana redonda, se conecta con la caja del tímpano.

El caracol membranoso o conducto coclear, es un tubo de sección triangular que se enrolla dentro del conducto óseo. Su base ocupa el espacio entre el borde libre de la lámina espiral y la lámina del contorno completando el tabique que separa las dos rampas del caracol. El conducto coclear es aveces llamado rampa media y su pared inferior se conoce como membrana basilar.

El órgano de Corti

La membrana basilar sirve de apoyo al órgano de Corti, donde llegan las prolongaciones protoplasmáticas del *ganglio de Corti*. Este ganglio se encuentra a lo largo de todo el conducto espiral de Rosenthal y sus prolongaciones cilindroaxilares dan origen a la rama coclear, que junto a la

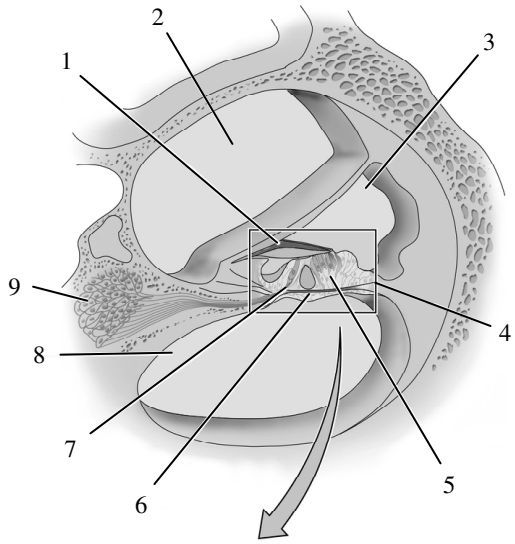

Figura 1.3. Corte transversal de una espira de la cóclea. 1: Membrana tectorial; 2: Rampa vestibular; 3: Rampa media; 4: Órgano de Corti; 5: Células ciliadas externas; 6: Membrana basilar; 7: Células ciliadas internas; 8: Rampa timpánica; 9: Ganglio de Corti.

vestibular, conforman el nervio auditivo.

Los componentes del *órgano de Corti* pueden clasificarse en: estructuras de soporte y células sensoriales. Las estructuras de soporte consisten básicamente en células de morfología diversa y elementos no celulares. Para mayores detalles véase la Figura 1.4.

El mayor interés lo merecen los dos tipos morfológicamente diferentes de células sensoriales: las *ciliadas internas* y las *ciliadas externas*. El órgano de Corti consta de unas 3000 células ciliadas internas dispuestas en una sola hilera y rodeadas completamente por células de soporte. En la parte apical presentan de 40 a 60 estereocilios que no se encuentran anclados a la membrana tectorial. Las células ciliadas internas están inervadas por la rama coclear del nervio auditivo y sus fibras aferentes representan el 95 % de la inervación total del órgano de Corti.

Las células ciliadas externas son más numerosas (unas 9000) y están dispuestas en 3 o 4 hileras libres de células de soporte, formando pequeñas V, por debajo de la membrana tectorial. De 100 a 120 estereocilios en su parte apical se unen firmemente a la membrana tectorial.

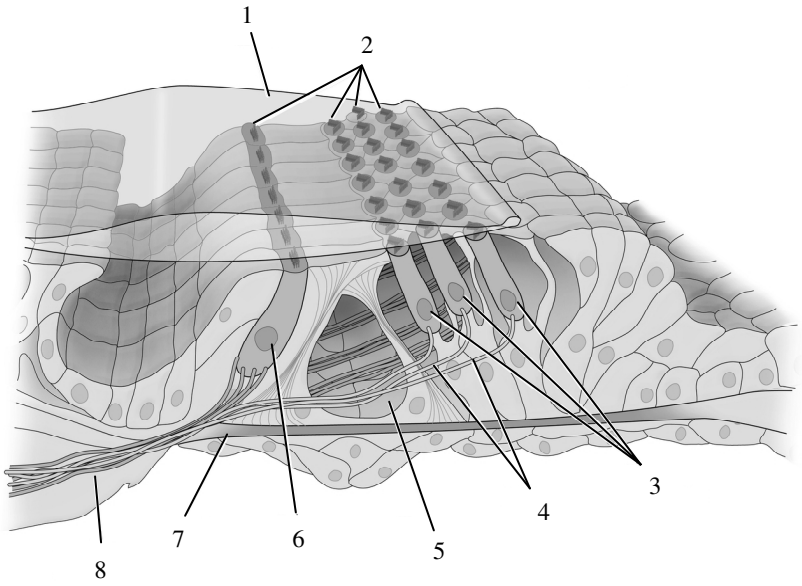

Figura 1.4. Ilustración del órgano de Corti. 1: Membrana basilar; 2: Estereocilios; 3: Células ciliadas externas: 4: Axones eferentes; 5: Túnel de Corti; 6: Células ciliadas internas; 7: Membrana basilar; 8: Axones eferentes.

1.2.2. Fisiología de la cóclea

A lo largo del tiempo, se han ido incorporado al RAH diferentes características de los mecanismos de procesamiento y codificación del sonido que tienen lugar en el ser humano. A continuación se hará una breve revisión de la forma en que se codifica fisiológicamente el sonido. En la fisiología de la cóclea se pueden encontrar los siguientes tópicos de importancia:

- la mecánica vibratoria de la membrana basilar,
- la fisiología de las células ciliadas y la respuesta al sonido en el nervio auditivo,
- las teorías de la percepción de la frecuencia fundamental,

Mecánica vibratoria

Antes de comenzar con la mecánica vibratoria de la membrana basilar será útil repasar brevemente el trayecto de las ondas de presión del sonido.

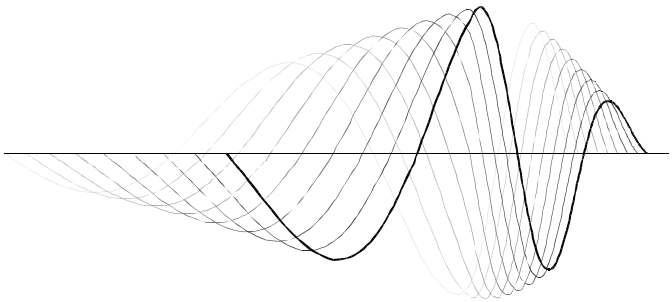

Figura 1.5. Nueve posiciones de la onda viajera desplazándose a lo largo del conducto coclear (supuestamente estirado). La línea central indica la posición de reposo de la membrana basilar. Los tonos de gris en las curvas dan una idea de los instantes de tiempo en que la onda estuvo en cada posición, cuanto más oscuro, más reciente. La base de la cóclea se encuentra a la izquierda y el ápex a la derecha. La estimulación consiste en un tono puro.

Desde el exterior el sonido se conduce a través del conducto del oído externo para hacer vibrar al tímpano. La membrana del tímpano transmite las vibraciones a la cadena de huesecillos y el último de éstos, el estribo, las transmite mediante la ventana oval a la perilinfa que se encuentra en la rampa vestibular.

Las vibraciones en la ventana oval forman ondas de presión en la perilinfa que se equilibran poniendo en movimiento al conducto coclear. Las ondas de presión pasan así a la rampa timpánica y transmiten a la ventana redonda un movimiento opuesto al producido por la ventana oval.

Las regiones en que la perilinfa de la rampa vestibular tiene mayor presión se corresponden con un mayor desplazamiento del conducto coclear hacia abajo. En cambio en las regiones en que la perilinfa de la rampa timpánica tenga mayor presión se desplazará el conducto coclear hacia arriba. De esta manera el caracol membranoso tendrá una forma que acompaña las diferencias de presión entre la perilinfa de la rampa vestibular y la de la rampa timpánica. Dado que este es un proceso dinámico, las diferencias de presión se desplazan en forma de *ondas viajeras*.

Para cada frecuencia de estimulación existe a lo largo de la membrana basilar una zona de máximo desplazamiento, debido a los cambios en el ancho y la elasticidad de la membrana basilar a lo largo de el conducto coclear. Esto se denomina *resonancia* o *sintonía mecánica* de la membrana basilar. La Figura 1.5 muestra varios instantes de una onda viajera. El pico de su envolvente de amplitud tiene una ubicación, a lo largo del conducto

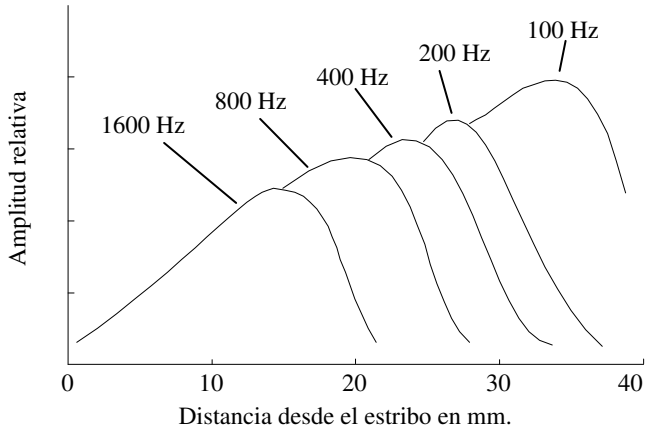

Figura 1.6. Amplitud del movimiento a lo largo de la membrana basilar para distintas frecuencias de estimulación con la misma intensidad.

coclear, dada por la frecuencia de estimulación.

Las características de la onda viajera pueden ayudar a comprender mejor la mecánica vibratoria de la membrana basilar. A continuación se destacan algunas particularidades de estos movimientos (Figura 1.6):

- Para estímulos de alta frecuencia las amplitudes máximas del movimiento se encuentran en la zona basal de la membrana basilar mientras que si la frecuencia es baja se encuentran en la zona apical. La amplitud del movimiento en distintas regiones de la membrana basilar depende de la frecuencia.

- Las amplitudes máximas alcanzadas dependen de la intensidad del sonido pero, para una misma intensidad, los sonidos de menor frecuencia producen mayores amplitudes.

- A bajas frecuencias la variación de la posición del máximo desplazamiento en función de la frecuencia es casi lineal. Sin embargo a partir de 1 KHz (aproximadamente) el comportamiento es logarítmico.

Transducción en las células ciliadas

Las células ciliadas deben su nombre a un grupo de filamentos de actina que se encuentran en su membrana apical. Estos filamentos, también denominados estereocilios, tienen la particular característica de estar unidos

por pequeños microfilamentos que restringen los movimientos entre vecinos cercanos.

Las células ciliadas responden a un modelo simple denominado modulación de resistencia. La deflexión de la membrana basilar es acompañada por la deflexión en los estereocilios. Cuando la deflexión se realiza en el sentido del estereocilio más alto, también denominado quinocilio, la resistencia se reduce y provoca la despolarización de la célula ciliada. Cuando el desplazamiento es en el sentido contrario la resistencia de la membrana aumenta y la célula se hiperpolariza.

Existen diferencias importantes entre las células ciliadas internas y las externas. Como se describió antes, las primeras no están fijas a la membrana tectorial. Mientras los estereocilios de las células ciliadas externas siguen los movimientos relativos entre la membrana basilar y la tectorial, los estereocilios de la células ciliadas internas se ven movidos por la velocidad relativa entre la endolinfa y el órgano de Corti. Dado el bajo porcentaje de fibras aferentes de las células ciliadas externas, se ha postulado que su función es la de servir como lazo de realimentación y proveer una actividad motora que contribuya al fenómeno de *sintonización mecánica* de la membrana basilar.

Ante el estímulo auditivo, en las células ciliadas se generan potenciales receptores con dos componentes principales. Uno de ellos sigue las variaciones instantáneas de los desplazamientos mecánicos del conducto coclear y es el denominado componente de corriente alterna. El otro está relacionado con la envolvente de estas variaciones y se denomina componente de corriente continua. Estos componentes confluyen para dar una de las características más importantes de los mecanismos de transducción: la *fijación de fase*. Se observa que existe cierta preferencia de las células ciliadas internas para iniciar su potencial de acción durante la primera mitad del ciclo de un estímulo senoidal. Es decir, las células ciliadas internas también estarían enviando información acerca de la frecuencia del estímulo ya que sincronizan el inicio de sus salvas de disparos con el primer medio ciclo de la onda de estimulación. El mecanismo de detección de fase consistiría justamente en estimular con el máximo positivo de la derivada del movimiento, que está relacionado con los cruces por cero de la onda con que se estimula.

La información que resta codificar es la amplitud del estímulo y esto se realiza de la siguiente forma: cuando la intensidad del estímulo aumenta o disminuye, la frecuencia de disparo la sigue según una función sigmoidea. Esta función, dependiente de la sintonía mecánica de la membrana basilar, alcanza los valores más altos para la frecuencia característica de cada zona.

Finalmente se puede concluir que los tres parámetros que caracterizan

una onda sinusoidal recibida por una célula ciliadas son codificados de la siguiente forma:

- *Amplitud*: frecuencia de disparo de cada salva según la función sigmoidea.
- *Frecuencia*: frecuencia de las salvas de disparos, correlacionadas con las primeras mitades da cada ciclo de la onda de estímulo.
- *Fase*: ocurrencia de los comienzos de las salvas en relación a sus vecinas.

Debido a la limitación en la velocidad de respuesta del mecanismo, el fenómeno de fijación de fase es posible solamente por debajo de los 5 KHz.

Percepción de la entonación

En base a la codificación fisiológica de la composición frecuencial del sonido se describen dos teorías acerca de la forma en que percibimos el tono fundamental.

La primera está basada en la descomposición frecuencial realizada por la mecánica vibratoria de la membrana basilar. Esta descomposición consistía en asignar la energía de una banda reducida del espectro de la señal a una amplitud de oscilación con una localización espacial específica en la membrana basilar. La información estaría contenida en la frecuencia de los impulsos de las salvas enviadas por las fibras del nervio auditivo y su ubicación relativa a lo largo de la rampa coclear. De esta manera se ve a la cóclea como un analizador de espectro. Este principio es conocido como *codificación por lugar* o principio de tonotopía, que puede resumirse en su similitud con un conjunto de filtros pasa bajos de poca selectividad y con frecuencias de corte según la ley lineal-logarítmica descrita anteriormente. Esta ley se obtuvo experimentalmente a partir de la frecuencia percibida para diferentes tonos puros. Así se dio origen a la denominada *escala de mel*, con la que se constituye una nueva unidad de medida para la frecuencia perceptual: el mel. En la Figura 1.7 se puede observar una representación gráfica de la asignación de frecuencias en las diferentes regiones de la rampa coclear.

La segunda teoría se basa en el fenómeno de fijación de fase. En este caso la información acerca de las componentes frecuenciales de la señal estaría contenida en la frecuencia de las salvas de impulsos y en su fase relativa. Así, la cóclea actuaría como un analizador de la señal en el dominio temporal.

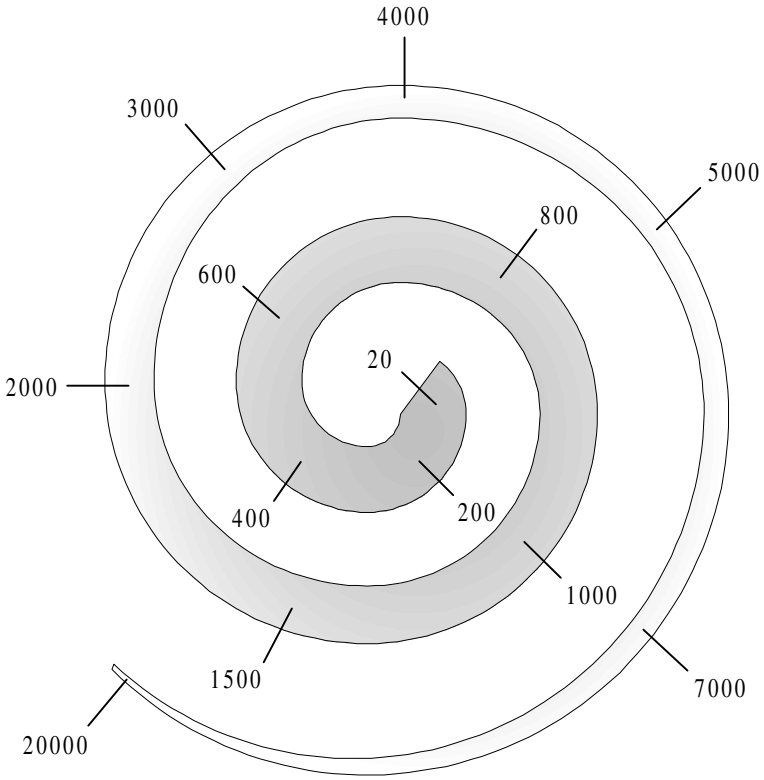

Figura 1.7. Percepción de la entonación por el principio tonotopía. Se pueden observar las dos vueltas y media de la rampa coclear y la asignación de las frecuencias percibidas (en Hz) a cada región de la misma.

Este principio se denomina *codificación temporal*, principio de periodicidad y también principio de temporización.

Para componentes frecuenciales que se encuentran por debajo de los 5 KHz ambos principios son concurrentes para realizar una codificación compuesta. Sin embargo, en este rango de frecuencias tiene mayor peso en la percepción el principio de periodicidad. Como por arriba de los 5 KHz se anula la fijación de fase, el principio de codificación por lugar comienza a predominar. Sin embargo, se sabe que la mayor parte de la información contenida en el habla se encuentra por debajo de los 5 KHz.

1.2.3. Anatomía del aparato fonador

El aparato fonador puede considerarse constituido por cuatro partes: los pulmones, la laringe, las cuerdas vocales y el tracto vocal. Los pulmones

son los encargados de proporcionar la energía necesaria para la producción de los sonidos. La laringe y las cuerdas vocales constituyen principalmente el sistema vibrante y el tracto vocal se puede ver como una caja de resonancia con morfología variable, que termina de dar forma a los sonidos de la voz. Se revisarán primeramente algunas particularidades anatómicas del aparato fonador.

Tórax y pulmones

El aparato fonador comparte la mayoría de sus estructuras anatómicas con el sistema respiratorio. Algunas de estas estructuras adquieren mayor relevancia en la producción de la voz mientras que otras cumplen roles más secundarios. En relación a esta observación se puede restringir la amplitud de la siguiente descripción.

El sistema respiratorio puede dividirse en: vías respiratorias superiores y vías respiratorias inferiores. Las vías respiratorias superiores comprenden a las fosas nasales, boca y faringe. Las vías respiratorias inferiores comprenden a los pulmones, bronquios, tráquea y laringe. Desde el punto de vista de la fonación, las estructuras de mayor interés en las vías respiratorias inferiores son los pulmones y la laringe.

El *tórax* es la estructura óseo-muscular que contiene a los principales componentes de las vías respiratorias inferiores. El esqueleto del tórax está constituido por las vértebras dorsales, las costillas, los cartílagos costales y el esternón. En cuanto a los músculos tórax, se pueden distinguir tres grupos principales: los de la pared posterior del tronco, los de la región anterolateral del tórax y el diafragma [Rouvière y Delmas, 1988b].

Los *pulmones* se encuentran en la cavidad del tórax, separados por las pleuras. Pocos órganos presentan tanta variabilidad en su volumen como los pulmones. Existen diferencias importantes de acuerdo a la capacidad del tórax y los procesos fisiológicos de la inspiración y espiración. Después de una inspiración normal, la capacidad de los pulmones llega a 3.5 litros y mediante una inspiración forzada puede llegar a los 5 litros.

Laringe y faringe

Aún dentro de las vías respiratorias inferiores se incluye a la *laringe*: el órgano esencial de la fonación. Este órgano está constituido por varias piezas cartilaginosas, sus ligamentos, músculos y repliegues membranosos. Los cartílagos de la laringe son once y pueden separase en medios o impares y laterales o pares. Los músculos pueden ser agrupados en: extrínsecos, que

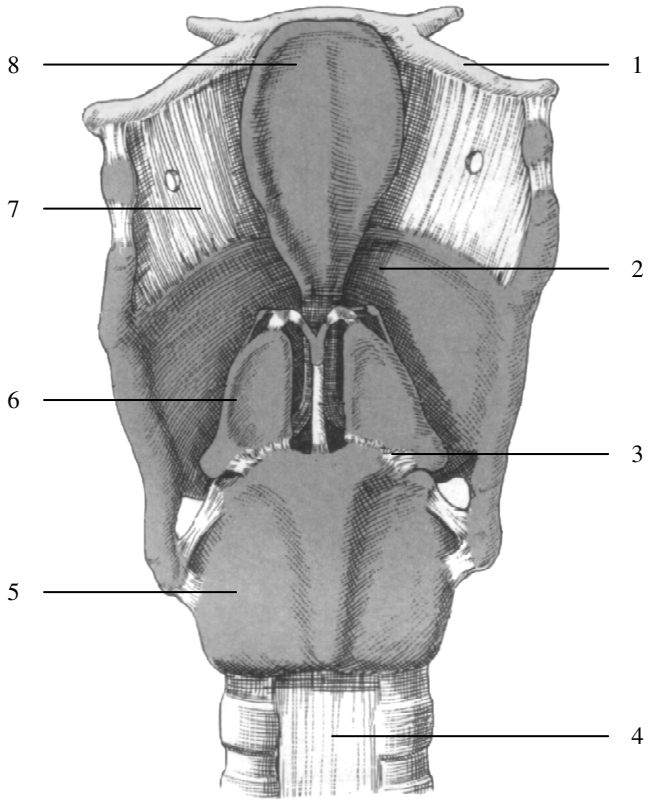

Figura 1.8. Vista posterior de los cartílagos y ligamentos de la laringe. 1: Hueso ioides; 2: Cartílago tiroides; 3: Ligamento cricoaritenoideo; 4: Pared membranosa de la tráquea; 5: Cartílago cricoides; 6: Cartílago aritenoides; 7: Membrana tirohioidea; 8: Epiglotis. (Modificado de [Latarjet y Liard, 1989])

unen la laringe con otros órganos vecinos, e intrínsecos, que le pertenecen a la laringe en su totalidad. Los músculos intrínsecos pueden distribuirse en tres grupos de acuerdo a su acción sobre las cuerdas vocales y sobre la glotis: los tensores de las cuerdas vocales, los dilatadores de la glotis y los constrictores de la glotis. Véase un detalle de la laringe en la Figura 1.8.

Para terminar con la laringe se deben revisar algunos detalles de su configuración interna. En su parte media, la laringe presenta dos repliegues superpuestos que forman las bandas ventriculares y las cuerdas vocales. Las bandas ventriculares se encuentran más cerca de la epiglotis, arriba de las cuerdas vocales, que van desde el cartílago tiroides al aritenoides. Las *cuerdas vocales* tienen forma prismática y sus bordes internos sobrepasan, hacia

adentro, a los de las bandas ventriculares. En su interior se encuentran el ligamento tiroaritenoideo inferior y el músculo tiroaritenoideo inferior. Tomando como referencia a las cuerdas vocales se suele dividir la laringe en tres pisos, uno superior o vestíbulo de la laringe, uno medio y uno inferior o subglótico [Rouvière y Delmas, 1988a].

Si se continúa subiendo en las vías respiratorias se encuentra la primera región de la *faringe*, denominada laringofaringe o hipofaringe. La faringe es un embudo musculomembranoso irregular que asciende verticalmente unos 15 cm. En su parte superior se comunica por detrás con las cavidades oral y nasal. En base a estas relaciones se describen, además de la laringofaringe, dos regiones consecutivas más: la orofaringe y la rinofaringe. Por abajo de la rinofaringe se encuentra el velo del paladar, separándola de la orofaringe y restringiendo selectivamente el paso de aire hacia la cavidad nasal [Stevens, 1998]. Los músculos de la faringe pueden agruparse en: constrictores y elevadores. Los músculos constrictores estrechan los diámetros anteroposterior y transversal de la faringe. Por la acción de los músculos elevadores la faringe puede reducir su longitud hasta 3 cm [Rouvière y Delmas, 1988a].

La faringe, junto con el vestíbulo de la laringe y las cavidades oral y nasal, constituye el tracto vocal (Figura 1.9). En el piso de la *cavidad oral* se encuentran la lengua y la mandíbula. La lengua —el articulador por excelencia— es una formación muscular compleja que no se une a ninguna estructura ósea en su dorso, su ápice, los costados y en la parte anterior de la superficie inferior [Manrique, 1980]. Limitando hacia adelante de la cavidad oral se encuentran los labios, que también son formaciones musculares complejas compuestas por varios músculos faciales que se unen en una banda que rodea a la boca.

Para terminar con las partes del tracto vocal hay que describir brevemente a la *cavidad nasal*, que se extiende desde la rinofaringe hasta los orificios de la nariz. Esta cavidad está dividida en dos espacios aproximadamente iguales y paralelos. En la estructura ósea de sus paredes se encuentran los cornetes, que son unos huesos curvados que dividen a cada fosa en varios canales. Los cornetes se conectan a través del tabique y hacen que la cavidad nasal pueda verse como una única estructura resonante. Todos estos espacios están recubiertos por una mucosa gruesa y húmeda.

1.2.4. Producción del sonido articulado

La voz puede estudiarse desde la perspectiva del mecanismo de su producción. En este proceso estarán presentes una fuente de energía, los generadores del sonido y los modificadores del sonido. La fuente de energía la

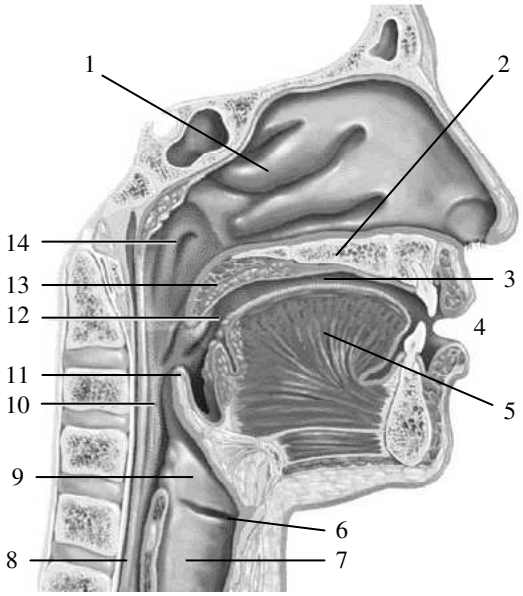

Figura 1.9. Ilustración de las diferentes partes del tracto vocal. 1: Cavidad nasal; 2: Paladar duro; 3: Cavidad oral; 4: Labios; 5: Lengua; 6: Ventrículo de Morgagni (por arriba está la banda ventricular y por abajo la cuerda vocal); 7: Extremo superior de la tráquea; 8: Esófago; 9: Vestíbulo de la laringe; 10: Laringofaringe; 11: Epiglotis; 12: Orofaringe; 13: Paladar blando (velo del paladar); 14: Rinofaringe.

constituyen los músculos torácicos, que impulsan el aire a través de las vías respiratorias. La generación del sonido puede realizarse en varios puntos a lo largo del tracto vocal, pero son principalmente las cuerdas vocales las que cumplen con esta función. La morfología variable del tracto vocal y todas sus cavidades determinan qué frecuencias del sonido generado van a ser realzadas y cuales serán atenuadas. El tracto vocal en su conjunto trabaja así como un resonador que termina de dar forma a los sonidos de la voz.

Fuente de energía

La energía que proporcionan los músculos del sistema respiratorio se manifiesta en la forma de un flujo de aire que interactúa con las diferentes partes del aparato fonador. Durante la respiración entran en juego tanto las vías aéreas y pulmones como todo el sistema mecánico de la caja torácica y los centros nerviosos bulbares y medulares. Para la generación de las diferencias de presión que dan lugar al habla se requiere un ciclo de inspiración en el que la caja torácica aumenta todos sus diámetros. En general se consi-

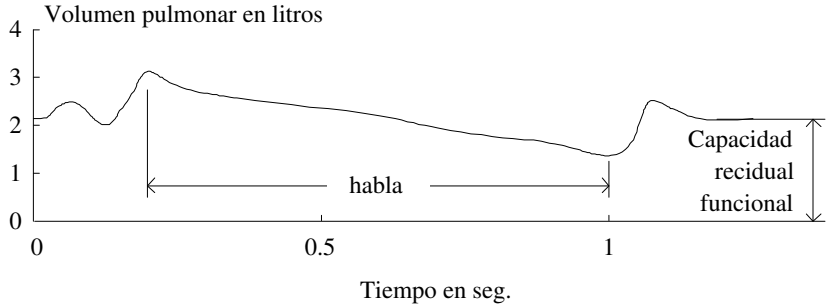

Figura 1.10. Variaciones del volumen pulmonar durante la respiración normal y durante la fonación. (Modificado de [Manrique, 1980] y [Cingolani y Houssay, 1988b]).

dera que los músculos intercostales externos son inspiradores mientras que los intercostales externos actúan más bien en la espiración.

La fonación se produce normalmente durante la espiración y afecta a todo el ciclo respiratorio. Si bien durante la respiración normal los tiempos de cada ciclo son casi iguales, durante la fonación la fase espiratoria puede llegar a ser 8 veces más larga que la inspiratoria (obsérvese la Figura 1.10). La inspiración se hace más profunda y la espiración se produce con cambios continuos de acuerdo a la intensidad de la voz y otros fenómenos importantes como el acento o la separación entre palabras. Cuando el aire es forzado a salir de los pulmones atraviesa los bronquios, la tráquea, la glotis, la laringe y sigue hacia las cavidades de la faringe, la boca y la nariz para terminar atravesando los labios y las fosas nasales. En este trayecto la energía que transporta esta corriente de aire puede tomar formas sonoras muy diferentes gracias a la participación de generadores y modificadores del sonido. En la Figura 1.11 se muestra la variación de la energía de la señal de voz a lo largo de una frase.

Generadores del sonido

Como se mencionó antes, el principal generador de sonido se encuentra en la laringe y está constituido por las cuerdas vocales. Desde el punto de vista funcional se puede dividir la laringe en tres partes que actúan sobre las cuerdas vocales: el aparato fibroso de soporte y su esqueleto, el aparato tensor y el aparato motor.

En el aparato fibroso y esqueleto de la laringe está la articulación cricotiroidea, que permite movimientos de balanceo alrededor de un eje transversal que pasa por las articulaciones a la derecha e izquierda. Los cartílagos

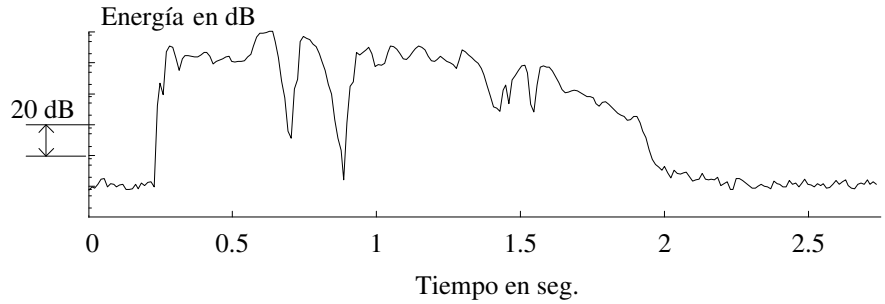

Figura 1.11. Energía a lo largo de la frase: *Comunidad autónoma más grande.* Esta frase se extrajo del corpus de habla Albayzin, que se detalla en el Apéndice A.

aritenoides se balancean hacia atrás y ayudan así a poner en tensión a la cuerda vocal. El aparato tensor se suma a la generación del movimiento de balanceo de los cartílagos fonatorios y agrega la tensión del músculo tiroaritenoideo.

En el caso de que las cuerdas vocales se encuentren relajadas y separadas el aire pasará por ellas sin provocar ningún sonido, pero si están tensas modularán el aire en pulsos cuya frecuencia dependerá fundamentalmente de la tensión y del tamaño del órgano. Se puede decir que estas emisiones son soplos de aire cuasi-periódicos de banda muy ancha. En la Figura 1.12 se muestra la forma de onda y un análisis en frecuencia de los pulsos glóticos.

En el hombre, la frecuencia de vibración de las cuerdas vocales está entre 100 y 170 Hz, en las mujeres suele ir desde 180 a 280 Hz y en los niños puede superar los 300 Hz. Los valores de esta vibración glótica se modifican en forma voluntaria durante el canto y son los responsables de la frecuencia fundamental (F_0) producida al hablar. En la Figura 1.13 se muestra la variación de F_0 a lo largo de una frase leída por una mujer.

Existen otros dos generadores de sonido que merece la pena mencionar. Se produce un flujo turbulento que provoca ruidos cuasi-aleatorios cuando el aire pasa a través de constricciones estrechas en el tracto vocal, como en algunas posiciones del velo del paladar, los dientes, la lengua, los labios y otros. Por ejemplo, al pronunciar la /s/ se provocan ruidos entre la lengua y el paladar mientras que las cuerdas vocales están relajadas. Sin embargo, al pronunciar la /p/ se cierra completamente el tracto vocal en los labios y al abrirlos se libera la presión en un punto provocando un breve impulso.

De acuerdo con la zona en donde se genera el sonido se ha realizado una primera división de los sonidos de la voz. Se denominan *sonoros* cuando en la generación intervienen las cuerdas vocales y se habla de *sordos* cuando

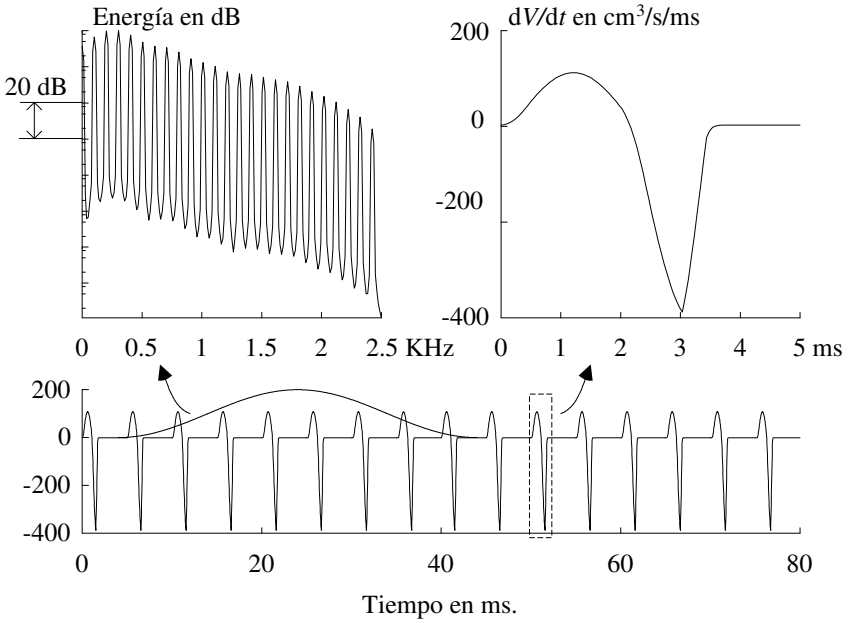

Figura 1.12. Pulsos glóticos en el tiempo y en la frecuencia. En la parte de abajo se observa un tren de pulsos glóticos con una frecuencia fundamental de 100 Hz. La forma de onda del pulso glótico (presión sonora proporcional a la derivada de la velocidad del volumen de aire a través de la glotis) está ampliada en el detalle de arriba, a la derecha. En la parte de la izquierda se observa un análisis frecuencial de los primeros 8 pulsos del tren.

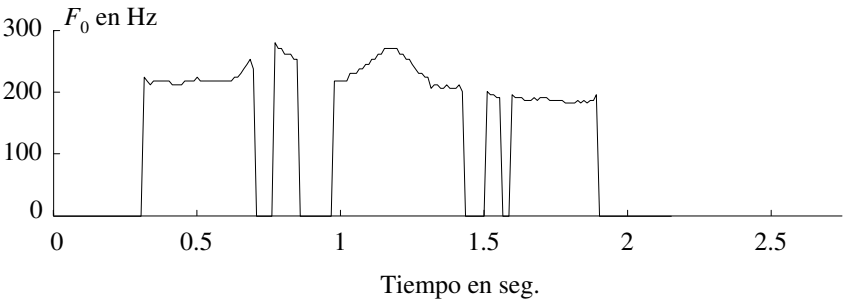

Figura 1.13. Frecuencia fundamental para la misma frase de la Figura 1.11: *Comunidad autónoma más grande.*

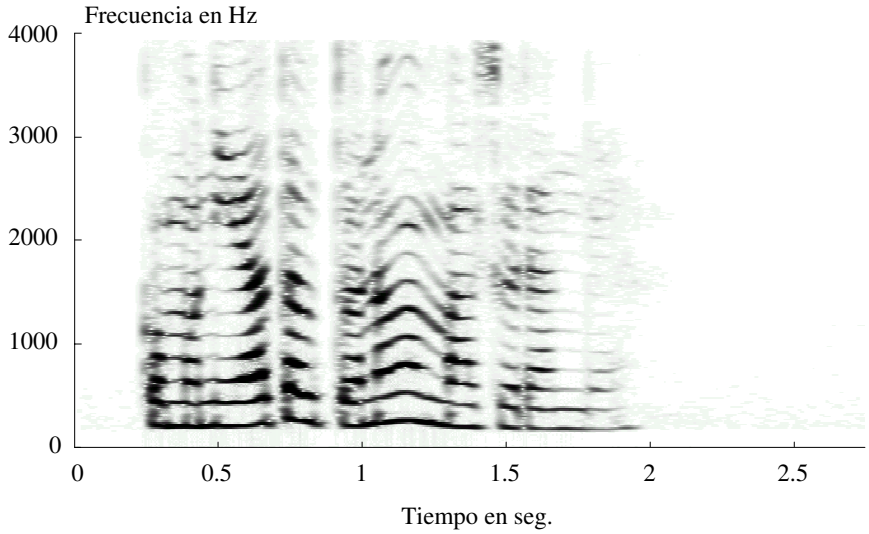

Figura 1.14. Espectrograma de la misma frase de las Figuras 1.11 y 1.13: *Comunidad autónoma más grande.*

el generador están en otra parte del tracto vocal.

Modificadores del sonido

El sonido de todos los generadores es muy rico en componentes frecuenciales cubriendo en conjunto toda la banda del espectro sonoro del habla. Estos sonidos atraviesan todo el tracto vocal en donde reciben muchas modificaciones debido a sus irregularidades. Algunas frecuencias contenidas en la señal original son fuertemente atenuadas mientras que otras pueden reforzarse por resonancias acústicas, dependiendo de la disposición de las irregularidades que varían constantemente cuando se articula una palabra. Si para un sonido en particular se grafica su espectro de frecuencias, se podrán ver algunos picos de resonancia y otros valles donde hubo predominantemente atenuaciones. Es posible ver al tracto vocal como un conjunto de resonadores que se encarga de reforzar o atenuar ciertas frecuencias según sea el sonido que se desea pronunciar. A lo largo de una frase los cambios en la morfología del tracto vocal y las alternancias entre las diferentes fuentes de sonido dan como resultado un cambio permanente del espectro de la señal resultante. En la Figura 1.14 se puede apreciar cómo varían las componentes frecuenciales a lo largo de una frase completa.

Cuando se excita al tracto vocal con los pulsos glóticos, sólo aquellas

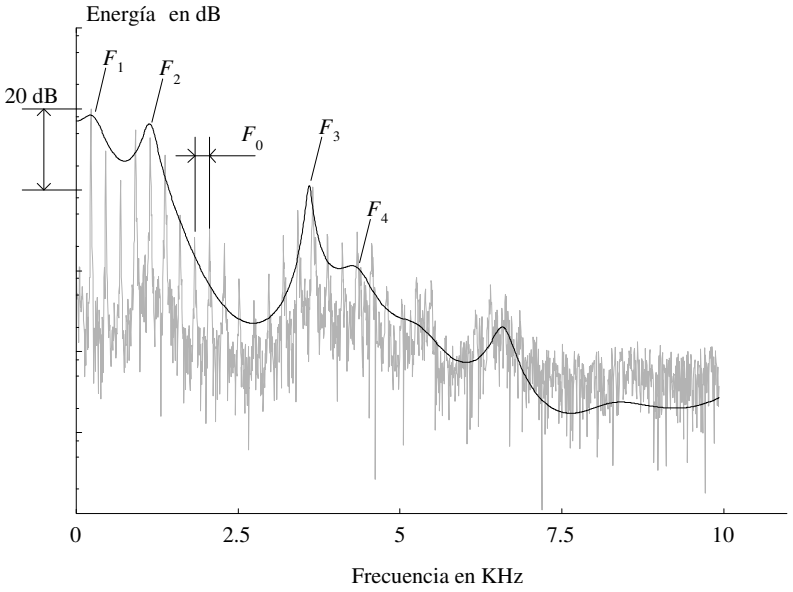

Figura 1.15. Espectro de energías para la vocal /a/ con $F_0 \approx 250$ Hz.

pocas bandas de frecuencia que coincidan con la frecuencia de resonancia de alguno de sus resonadores no serán atenuadas. Como resultado, en la salida predominarán algunas ondas sinusoidales amortiguadas que se verán como picos en el espectro de frecuencias. Este es el concepto de *formante*, que puede definirse más precisamente como: energía que se concentra en una banda de frecuencia por efecto de un resonador del tracto vocal. Algunas veces también se define a la formante en el dominio del tiempo como: una de las ondas sinusoidales que se observan en la señal de salida del resonador estimulado por pulsos glóticos. Las formantes se notan con una F seguida por un número que indica su orden de aparición desde las frecuencias más bajas. Esta enumeración sigue de F_0, notación que se utilizó para la frecuencia de la emisión glótica. Generalmente se pueden ver en forma clara varias formantes en los sonidos vocálicos[1] y ciertos sonidos consonánticos conservan las formantes de su contexto vocálico. En la Figura 1.15 se puede observar un análisis en frecuencia de la vocal /a/ con una estimación suavizada del espectro en la que se pueden apreciar claramente las cuatro primeras formantes.

[1]Luego se definirán más precisamente los sonidos vocálicos y consonánticos.

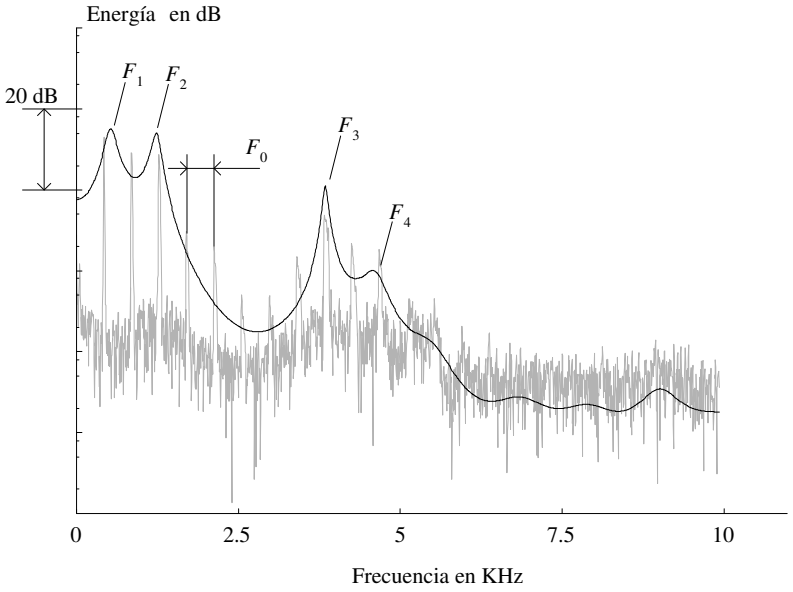

Figura 1.16. Espectro de energías para la vocal /a/ con $F_0 \approx 415$ Hz. Compárese con la Figura 1.15 y obsérvese como la posición relativa de las formantes se mantiene al cambiar la F_0.

La importancia de las formantes radica en que su posición identifica a los sonidos vocálicos. En la Figura 1.16 se muestra otro análisis frecuencia para una /a/ que ha sido pronunciada con una F_0 más alta[2]. A pesar de este cambio en la F_0, se puede observar como las cuatro primeras formantes quedaron prácticamente en el mismo lugar que estaban en la Figura 1.15. Para contrastar, obsérvese la posición de las formantes para una /i/ en la Figura 1.17. En la próxima sección se estudiará con mayor detalle la caracterización de los sonidos vocálicos en base a la posición de sus formantes.

Resta por mencionar el fenómeno de radiación a partir de las cavidades oral y nasal. Un modelo sencillo adopta la característica de radiación como proporcional a la frecuencia, a razón de unos 6 dB por octava [Stevens, 1998]. Las pérdidas por radiación se manifiestan principalmente en las bajas frecuencias y en su conjunto el fenómeno compensa, en parte, la menor energía en las componentes de alta frecuencia del pulso glótico.

[2]La F_0 en este caso es de aproximadamente 415 Hz. Esta frecuencia no es normal en el habla para un adulto pero se ha utilizado la voz de un niño para que las diferencias sean más notorias en este ejemplo.

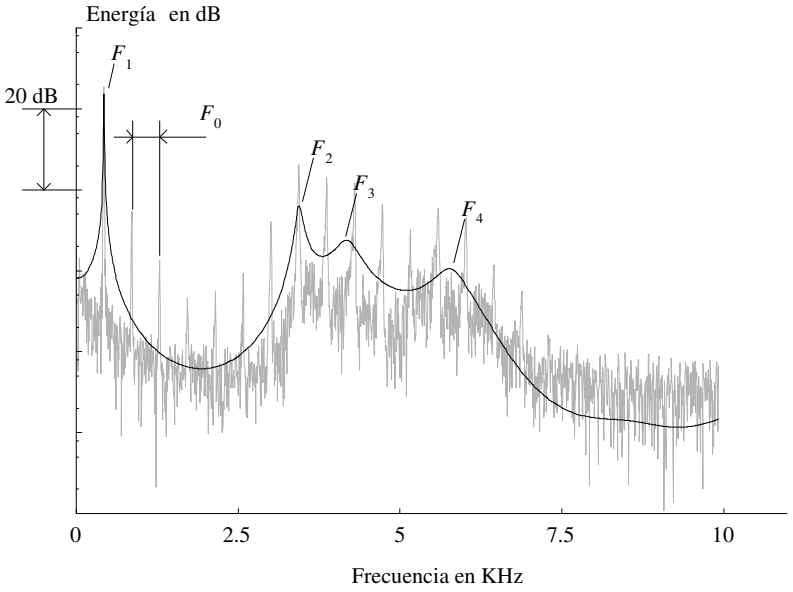

Figura 1.17. Espectro de energías para la vocal /i/ con $F_0 \approx 415$ Hz. Compárese la posición de las formantes con la vocal /a/ de las Figuras 1.15 y 1.16.

1.3. Organización estructural

El habla puede organizarse según distintas estructuras jerárquicas de acuerdo con el aspecto que se considere como central. De todas las formas de estructurar el conocimiento acerca del habla, la lingüística nos provee de una jerarquía en base a la que se pueden desarrollar muchos otros estudios. En la Figura 1.18 se muestra una estructura donde quedan sintetizados muchos de los aspectos que se estudian a diferentes niveles de análisis del habla. En esta figura se incrementa de arriba hacia abajo, no solamente el nivel de abstracción sino también la cantidad de elementos que son objeto de análisis. Las divisiones que se han realizado no son estáticas y tampoco se encuentran definidas de forma absoluta, en general los bordes son borrosos como en todo intento de clasificación de objetos de la realidad. En los últimos niveles entran en juego varias ramas de la lingüística, como la dialectología o la sociolingüística e incluso se puede observar una frase en un idioma diferente.

A continuación se describirán brevemente los distintos niveles y las ramas de la ciencia que los tratan, intentando introducir las complejidades que entraña su estudio y haciendo especial énfasis en aquellos que están más

relacionados con la presente Tesis.

1.3.1. La señal de voz y el análisis por tramos

En este nivel comienza todo con la señal continua de voz. Esta señal es la que se encuentra en forma de ondas de presión y se dice que es continua ya que para cada intervalo de tiempo que se considere, por pequeño que éste sea, siempre podrá medirse un valor de presión sonora. La física acústica ha estudiado la forma en que estas ondas de presión se generan y propagan en el medio. Dado que las herramientas actuales de análisis trabajan en general con señales digitales, el primer paso consiste en convertir la señal de voz a una representación discreta (no representado en la Figura 1.18). Para esto se mide la presión sonora entre 8000 y 44000 veces por segundo y se almacenan todos estos valores. Este proceso también suele denominarse *muestreo* de la señal de voz.

No tiene mucho sentido analizar la señal muestra por muestra, ni tampoco resulta muy útil analizar toda una frase de varios segundos como una sola cosa. Es por esto que se analiza la voz por *tramos*, en donde se puede considerar que la morfología del tracto vocal ha permanecido invariable. Estos tramos miden generalmente entre 10 y 30 ms. En la parte superior de la Figura 1.18 se destaca esta primera separación.

En general no se analizan directamente los tramos de voz en su evolución temporal sino que se aplican técnicas de procesamiento de señales para obtener representaciones que ponen de manifiesto las características más relevantes de la voz. Un ejemplo de estas representaciones es el análisis frecuencial, como se muestra en la Figura 1.18. También a partir de técnicas de procesamiento de señales suelen extraerse otras características de interés, como valores de F_0 o energía para cada tramo, pudiendo construir así, por ejemplo, curvas melódicas. A partir de cada uno de los tramos de voz, de sus características y en base a conocimientos de la física acústica, se puede hacer una primera distinción entre sonidos del habla, silencios y otros sonidos que no son de interés para el análisis y se descartan como ruidos.

Se ha revisado la acústica relacionada con la producción de la voz en la Sección 1.2.4 y se dará un trato formal al procesamiento de señales en la primera sección Capítulo 2.

1.3.2. Fonos y fonemas

En el siguiente nivel ya pueden distinguirse las primeras unidades del habla. A partir del análisis del proceso de generación y el resultado acústico

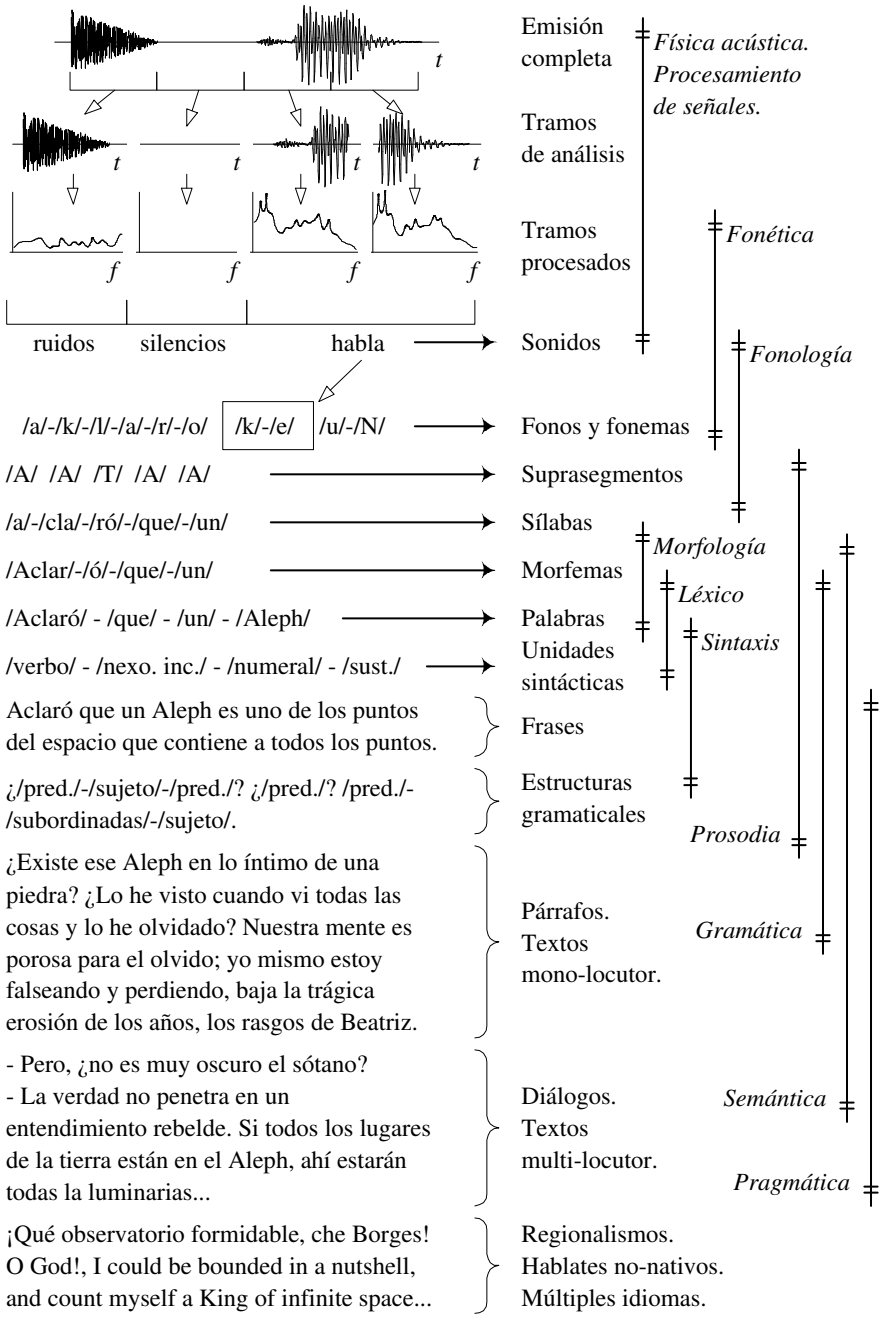

Figura 1.18. Organización estructural del habla. (Los textos fueron extraídos de *El Aleph*. La última frase es una cita de *Hamlet* que Jorge Luis Borges hace en su cuento.)

se han establecido modelos para los sonidos elementales del habla. Estos modelos son los denominados *fonemas*. Existen algunas reglas simples para identificar a los fonemas [Quilis, 1993]:

1. Dos sonidos que aparecen en el mismo contexto y pueden ser sustituidos uno por el otro sin que cambie el significado de la palabra.

2. Dos sonidos que son similares articulatoria o acústicamente, que nunca se encuentran uno al lado del otro y que nunca se presentan en el mismo contexto.

La función que cumplen estos modelos en un sistema de comunicación lingüística es estudiada por la *fonología*[3]. Una de las funciones principales de la ortografía es la de relacionar los símbolos que se utilizan en la escritura —grafías o letras— con los fonemas.

Por otro lado, es necesario considerar que estos modelos son pronunciados de diferentes formas dependiendo del contexto o del hablante. Por ejemplo, en la palabra *laba* el fonema /b/ no se pronuncia de igual forma que en la palabra *bala*. En este caso se observan dos realizaciones diferentes de un mismo fonema[4] y así se llega al concepto de *alófonos*, como diferentes realizaciones o variedades de un mismo fonema. También se conoce a los alófonos como fonos y como variantes. Es la *fonética* quién se encarga de estudiar los diferentes elementos fónicos de una lengua desde el punto de vista de su producción, caracterización acústica y percepción.

En base a los patrones de pronunciación, los modos articulatorios y los sonidos producidos, se ha clasificado a los sonidos del habla en dos grandes familias: los sonidos vacálicos y los sonidos consonánticos[5]. Las vocoides son las realizaciones acústicas de las vocales y se definen como aquellos sonidos que se producen sin estrechar o cortar el pasaje del aire que circula desde los pulmones hasta el espacio exterior. Las contoides son las realizaciones acústicas de las consonantes y corresponden a los sonidos producidos con algún estrechamiento u oclusión de la vía aérea en el tracto vocal. Una conclusión que surge de estas definiciones es que los sonidos vocálicos son producidos fundamentalmente utilizando a las cuerdas vocales como fuente de sonido. En contraste, los sonidos consonánticos poseen más componentes generadas por turbulencias y oclusiones en el tracto vocal.

[3] O también fonética funcional.

[4] El fonema /b/ posee dos realizaciones en el español, en el primer ejemplo es oclusiva y en el segundo fricativa. Más adelante se darán los detalles del caso.

[5] O también vocoides y contoides.

Clasificación de los sonidos vocálicos

Las formantes F_1, F_2 y F_3 son las más importantes para la caracterización de los sonidos vocálicos. Más aún, es posible realizar una buena clasificación con solamente las formantes F_1 y F_2. Las formantes superiores, con frecuencias generalmente mayores a los 3200 Hz, son bastante diferentes para distintos hablantes y caracterizan factores personales. En la Tabla 1.1 se pueden ver los rangos normales de estas dos formantes para las 5 vocoides del español.

Vocoide	F_1 en Hz	F_2 en Hz
/i/	200 a 400	1800 a 3500
/e/	400 a 700	1600 a 2700
/a/	600 a 1000	1000 a 2000
/o/	500 a 700	600 a 1000
/u/	250 a 400	600 a 1100

Tabla 1.1. Valores típicos para la primera y segunda formante de los sonidos vocálicos del español.

En la Figura 1.19 se muestra una representación más completa de las regiones que ocupan los sonidos vocálicos en el plano formántico, describiendo el *triángulo acústico*. Además, en esta figura también se muestra el denominado *triángulo articulatorio* —en coincidencia con el anterior—, según el cual se representa el modo y lugar en que se articula cada uno de los sonidos vocálicos. Los sonidos más graves coinciden con los que son articulados con la lengua en la región posterior. Así, la /o/ y la /u/ son clasificadas como graves, posteriores o velares. Cuando la lengua se articula en la región anterior se pronuncian la /i/ y la /e/, que se clasifican como agudas, anteriores o palatares. En cuanto a la forma de la articulación se distingue entre abiertas y cerradas, según la proximidad entre la lengua y el paladar. La /i/ y la /u/ son clasificadas como cerradas o altas, mientras que la /a/ es abierta o baja. Los restantes se clasifican como casos medios.

Clasificación de los sonidos consonánticos

La variedad y las características que identifican a los sonidos consonánticos son mucho más amplias que en el caso de los vocálicos. Una clasificación general según el alfabeto fonético internacional [Quilis, 1993, Manrique, 1980] muestra:

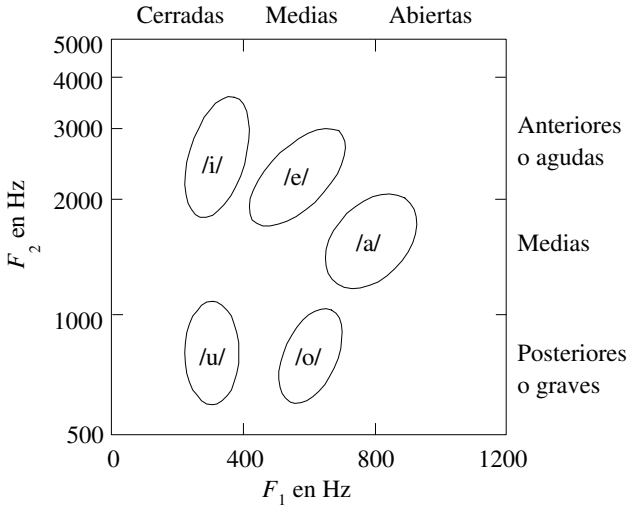

Figura 1.19. Características de las vocales del español. A la izquierda se representan las frecuencias de F_2 en escala logarítmica y abajo F_1 en escala lineal. A la derecha se han representado los tres lugares de articulación y arriba los modos de la articulación.

- Oclusivas suaves: [b], [d], [g]. Las oclusivas suaves son en parte sonoras ya que en la constricción no se anula completamente la frecuencia glótica. Duran aproximadamente 75 ms y sus armónicas son de frecuencia baja. Ejemplos de sus realizaciones son: *bata* [bata], *data* [data] y *gata* [gata].

- Oclusivas fuertes: [p], [t], [k]. Las oclusivas fuertes no son tan sonoras y se inician con el silencio de la oclusión total, que dura entre 30 y 100 ms. Pueden durar en total hasta 120 ms, por ejemplo: *palo* /palo/, *talo* /talo/ y *calo* /kalo/.

  Las bilabiales [p] y [b], cuando preceden a una vocoide, concentran su energía entre 2000 y 3600 Hz, según la vocoide asociada. En las velares [k] y [g] predomina su parte explosiva, cuya energía en el espectro se encuentra entre 700 y 2500 Hz según la vocoide asociada. Las alveolares [t] y [d] poseen su mayor energía entre 3000 y 4000 Hz.

- Nasales: [m], [n], [ñ]. El canal nasal provoca una importante atenuación de la banda entre 1000 a 2300 Hz y refuerza las formantes del alófono que oscilan en 240, 1020 y 2460 Hz. Como ejemplos se pueden citar: *mama* /mama/, *mana* /mana/ y *maña* /maña/.

- Líquidas laterales: [l], [λ]. Las formantes de las líquidas laterales siguen, con algunas modificaciones, a las formantes de las vocoides de su contexto. La F_1 de la [l] está siempre debajo de la F_1 de la vocal que la acompaña, con una media en 333 Hz. La F_2 oscila en 1550 Hz y la F_3 en torno a los 2550 Hz. Su duración total es de 100 a 200 ms. Ejemplos: *lata* /lata/, *plata*, /plata/ y *tal* /tal/. Con la [λ] se produce frecuentemente el fenómeno del *yeismo*, ampliamente difundido en todo el dominio del habla española[6]. En estos casos se reemplaza: pollo-poyo, rallar-rayar, callado-cayado, etc. [Quilis, 1993].

- Líquidas vibrantes: [r], [$\bar{\text{r}}$]. Las líquidas vibrantes consisten en una o varias oclusiones seguidas ([r] o [$\bar{\text{r}}$] respectivamente) a razón de unos 30 golpes por segundo. Como ejemplos de vibrante simple y múltiple se pueden citar: *coro* /koro/ y *corro* /ko$\bar{\text{r}}$o/. En estos casos también suelen conservarse las formantes del contexto, con las interrupciones propias en el caso de la vibrante múltiple.

- Fricativas sordas: [f], [s], [θ], [x]. La [f] es ruido de banda ancha que comienza por arriba de los 2100 Hz, por ejemplo: *fácil* /faθil/ y *difícil* /difiθil/. La [s] se pronuncia como un ruido creciente a partir de 2500 Hz con máximos alrededor de 3500 y 4500 Hz, por ejemplo: *solo* /solo/, *si* /si/. El fenómeno del seseo en las zonas meridionales de la Península española y los territorios atlánticos ha hecho perder en parte la [θ], una variante fricativa sorda de la [s] que se pronuncia con la lengua más adelante, por ejemplo en *cerca* /θerca/ o *zona* /θona/[7][Llorach, 1999]. La [x] posee varias realizaciones dependiendo de la región geográfica. En términos generales es un ruido de banda estrecha en baja frecuencia con algunas componentes de alta frecuencia. Por ejemplo: *jirafa* /xirafa/ y *general* /xeneral/.

- Fricativas sonoras: [y], [β], [∂], [γ]. La [y] posee una componente de la frecuencia glótica más un ruido con un descenso de la F_1 de las vocoides que la rodean, por ejemplo: *ese yeso* /ese yeso/. La [β] es una versión suave de la [f], con un ruido similar y frecuencia glótica, por ejemplo: *ese beso* /ese βeso/. La [∂] es un ruido cuyo espectro se aproxima al de las vocales del contexto, agregando la energía de la frecuencia glótica.

[6] Adicionalmente, en la región central de Argentina se han reemplazado [y] y [λ] por [ʒ], una fricativa sorda similar a la [ŝ] del inglés *she* /ŝi/.

[7] La /θ/ posee una posición linguointerdental mientras que la [s] es preferentemente linguoalveolar. En España se pronuncia con la lengua más adelante mientras que en Hispanoamérica la constricción es más cercana al paladar.

Por ejemplo: *ese dado* /ese ∂a∂o/. La [γ] es una versión suave de la [g], con ruido similar y frecuencia glótica, por ejemplo: *esa gata* /esa γata/.

- Africadas [ŷ], [ĉ]. Poseen una oclusión inicial (suave en [ŷ] o fuerte en [ĉ]) y una fricación que la sigue. Son también denominadas, desde el punto de vista articulatorio, semioclusivas. La banda con mayor energía es la misma para ambas y está alrededor de los 2200 Hz. Por ejemplo: *el yeso* /el ŷeso/ y *hacha* /aĉa/.

Como se ha podido observar, en varios de los sonidos consonánticos se conservan las formantes del contexto vocálico. Este fenómeno de *transiciones formánticas* sigue ciertas reglas y aporta información adicional acerca de la identidad de las contoides.

Existen algunos casos en que de forma sistemática la combinación de dos fonemas da como resultado una neutralización y se crea un nuevo modelo denominado *archifonema*, que también posee sus propias variantes alofónicas. Un ejemplo simple es el de la neutralización de fonemas nasales en posición silábica prenuclear: *un trombo*, que fonéticamente se transcribe como /uN troNbo/. También pueden neutralizarse [p] y [b] en [B], [t] y [d] en [D] y [k] y [g] en [G] o las líquidas vibrantes en el archifonema[R].

En las secuencias vocálicas sucede un fenómeno similar con la formación de diptongos, triptongos e hiatos. En el siguiente apartado se tratará este tema en relación a la división silábica en español.

1.3.3. Suprasegmentos y sílabas

En fonología se consideran suprasegmentos o prosodemas a elementos de un nivel superior al de los fonemas, relacionados con la expresión y representados principalmente por el acento, la cantidad y la entonación [Quilis, 1993]. Al igual que en el nivel anterior, estos elementos poseen diversas manifestaciones físicas y sus correspondientes modelos y símbolos lingüísticos. Cuando se distingue como un nuevo *nivel* se hace referencia principalmente a tres hechos: un *nuevo* conjunto de manifestaciones físicas que se *superponen* y poseen *duraciones* superiores a las del nivel anterior.

Estos hechos no se verifican estrictamente —en un sentido matemático o algorítmico— ya que, como bien se sabe, en el lenguaje natural es común encontrar excepciones a la regla. Sin embargo, existen reglas generales que rigen su uso y se agrupan bajo la denominación general de *prosodia*. Desde un punto de vista físico se define la prosodia como el efecto resultante de las diferentes combinaciones de energía, frecuencia fundamental y duración de

suprasegmentos, aplicadas al lenguaje hablado. Estas tres manifestaciones físicas constituyen los denominados *rasgos prosódicos*, dentro de los cuales también suele hacerse alusión a las pausas relacionadas a la puntuación y asociadas tanto a los finales de palabra, frases y párrafos.

La superposición de éstas implica que algunos de los rasgos de la emisión de un fonema son modificados sin que este fonema pierda su identidad[8]. Es decir, sin que se modifiquen aquellos rasgos que le son característicos, con lo que los suprasegmentos son unidades distintas e independientes de los fonemas. En el caso de la acentuación, se puede ver fácilmente que cuando se pronuncia una palabra aislada, la misma vocal puede emitirse con mayor o menor energía dependiendo de que se encuentre acentuada o no, respectivamente. Compárese la palabra *tomo* con la palabra *tomó*. Los fonemas (segmentos) /o/ son los mismos en ambos casos, sin embargo no son los mismos suprasegmentos los que se manifiestan en la acentuación. También se puede apreciar este fenómeno de superposición cuando se cambia el tono en una palabra para convertir una frase declarativa en interrogativa. Por ejemplo compárese la afirmación *Tomo.* con la pregunta *¿Tomo?*. Con este ejemplo también puede observarse una ampliación de la idea de superposición ya que se superponen dos fenómenos del mismo nivel. En el Capítulo 3 se analizan estas relaciones con mayor detalle.

Por otro lado se puede observar que el ascender en cada nivel de abstracción se acompaña necesariamente con un ascenso en la duración de las estructuras de estudio. Los rasgos prosódicos se pueden analizar en base a tiempos de simplemente un tramo de análisis hasta una frase completa. Sin embargo, se asocia al acento con las sílabas de una forma más natural que con los fonemas o las palabras. Es cierto que existen sílabas e incluso palabras que se conforman por un único fonema. También se puede hablar de palabras acentuadas o inacentuadas. Se puede considerar la entonación de toda una frase o de su medida en simplemente un tramo de análisis en la señal de voz. En cualquier caso, se acepta que el suprasegmento es una estructura de duración mayor a la de fonemas y menor a la de morfemas o palabras, que son afectadas por rasgos prosódicos comunes. En este rango difuso, que está más allá de los fonemas pero no alcanza a las palabras, se encuentra la sílaba como una estructura que no es estrictamente un suprasegmento pero se le aproxima en su duración.

[8]Principalmente a este fenómeno de superposición alude el prefijo *supra* del término suprasegmental.

Sílabas

A pesar de que se posee una noción intuitiva bastante precisa de su definición fonética, diversos autores coinciden en que no es una tarea simple delimitar y enmarcar teóricamente a la sílaba [Llorach, 1999, Quilis, 1993].

Desde un punto de vista estructural, la sílaba puede verse constituida por un núcleo sonoro y su contexto. Particularmente para el español, el núcleo sonoro está representado por un sonido vocálico y por esto se denomina *núcleo vocálico*. Este núcleo es generalmente es el que posee la mayor apertura articulatoria y debe permitir la extensión de su duración.

La división silábica en el español no está relacionada con la agrupación de fonemas según significantes o morfemas, como suele ocurrir en el inglés. En la mayoría de los casos, la separación silábica del español tampoco se relaciona con características particulares de la pronunciación. Es por estas razones que se puede enumerar un conjunto reducido de reglas que permite obtener una separación silábica de las palabras a partir únicamente de su representación ortográfica [Quilis, 1993]:

1. Cuando una consonante se encuentra entre dos vocales: ésta se agrupa con la vocal siguiente.

2. Cuando dos consonantes se encuentran entre dos vocales: se separa entre las dos consonantes salvo los siguientes casos en que ambas consonantes quedan con la segunda sílaba: /pr, br, pl, bl, fr, fl, gr, gl, kr, kl, dr, tr, tl/.

3. Cuando tres o más consonantes se encuentran entre dos vocales: permanecen inseparables los grupos /consonante/+/r,l/ y /ns/.

4. La conjunción de dos vocales abiertas o medias /o/,/a/ y /e/ constituye un hiato y se separa formando dos sílabas.

5. Las conjunción de las vocales /i,u/+/e,a,o/ y viceversa, forma un diptongo y no se separan salvo que la vocal cerrada (/i/ o /u/) esté acentuada. La vocal más abierta forma el núcleo vocálico.

6. Las conjunciones /i/+/u/ y viceversa, forman diptongo y no se separan. El núcleo vocálico estará formado por aquella vocal en la que recae la acentuación.

7. La conjunción de tres vocales forma un triptongo y no se separa. Al igual que en el diptongo, la vocal más abierta constituye el núcleo vocálico.

Acentuación

La tipología acentual del español, al igual que en el inglés, alemán o el italiano, es libre. Esto es, el acento puede encontrarse en cualquier parte de la palabra. No es este el caso del finés, en el que el acento se encuentra siempre en la primera sílaba o el caso del francés en que se encuentra siempre en la última sílaba [Llorach, 1999]. El hecho de que en palabras aisladas exista una relación estrecha entre el acento y los rasgos prosódicos a nivel suprasegmental, hace que sea atractivo profundizar en el conocimiento de estas relaciones en el caso del discurso continuo. Existen antecedentes indicando que en el discurso continuo no se dan estas coincidencias tan marcadas en cuando se trata palabras aisladas. Por ejemplo, para la entonación del español puede encontrarse un estudio en [Quilis, 1993] y en el caso del inglés en [Ying, 1998, Yaeger-Dror, 1996].

Para distinguir los núcleos silábicos que se encuentran acentuados se utilizan diversas notaciones. Algunos autores indican la acentuación con una tilde, superpuesta al símbolo fonético correspondiente: en el caso la palabra *casa* se notaría /cása/. Con alguna pérdida de información acerca de la identidad del fonema también suele utilizarse una mayúscula que indica la vocal acentuada: el ejemplo anterior se notaría /cAsa/. Finalmente, perdiendo toda la identidad y número de los fonemas que forman una sílaba, también se utiliza la notación /TA/. Esta es la *estructura acentual* de la palabra e indica que posee dos sílabas, la primera es *tónica* y la segunda es *átona*. Es posible escribir una frase completa como: *¡La casa de mis padres, mi casa, la vieja casa inveterada de la calle Garay!* en una *secuencia* de estructuras acentuales: /A TA A A TA A TA A TA TA AAATA A A TA AT/.

En el español, como regla general, sólo puede existir una sílaba tónica por palabra. La excepción la constituyen los adverbios terminados en *-mente*, que poseen dos sílabas tónicas, por ejemplo: *prácticamente* es /TAATA/. Según [Quilis, 1993] es útil distinguir también entre palabras acentuadas e inacentuadas. Si bien todas las palabras en forma aislada poseen una sílaba tónica, cuando éstas se encuentran en un contexto determinado del discurso continuo es posible que ninguna de sus sílabas posea la carga acentual. Las palabras inacentuadas se distinguen según su función gramatical, por ejemplo: el artículo determinado (*el* perro, *un* perro), la preposición (es *para* mejorar), algunas conjunciones (más *y* mejor, *que* si o *que* no, *pero* es mejor *aunque* más caro, *puesto que* estará limpio, *luego* de un día, *aún cuando* llueva), los términos de tratamiento (*Don* Enrique), el primero de los compuestos (*Ana* María, *cincuenta* y tres), los adjetivos posesivos y las formas como *donde* y *cuando*, en el caso de que no estén en una frase interrogativa.

Volviendo a las palabras acentuadas, es necesario distinguir aquellas que, a pesar de poseer una única sílaba, en el discurso continuo sobresalen por su acento. Este es el caso del pronombre que funciona como sujeto *él* o las formas interrogativas de *qué*, *cuál* o *quién*. Las palabras multisilábicas acentuadas pueden clasificarse según la posición de la sílaba tónica en relación a la última de la palabra. Esta clasificación distingue palabras oxítonas, paroxítonas, proparoxítonas y superproparoxítonas[9]. Las palabras oxítonas poseen la forma acentual /-T/, indicando con el signo menos a cualquier secuencia de sílabas átonas. Las paroxítonas tienen la forma /-TA/, las proparoxítonas /-TAA/ y las superproparoxítonas /-TAAA/. A partir de estas definiciones de tipología acentual, en español es posible relacionar directamente el acento con su representación ortográfica a través de la tilde:

1. Todas las palabras superproparoxítonas y paroxítonas llevan una tilde en el núcleo vocálico de su sílaba tónica.

2. Las palabras paroxítonas llevan tilde siempre que no terminen en /n/, /s/ o vocal.

3. Las palabras oxítonas llevan tilde siempre que terminen en /n/, /s/ o vocal.

Desde la perspectiva inversa, es posible conocer la acentuación de una palabra a partir de su representación ortográfica[10]:

1. La tilde indica de forma inequívoca al núcleo vocálico de la sílaba tónica.

2. Cuando no existe una tilde en la palabra es oxítona si termina en /n/, /s/ o vocal y paroxítona en otro caso.

Dada la relación directa entre este acento y su representación ortográfica, suele denominarse *acento ortográfico*, cumpliendo una función *distintiva*[11] a través de todas las posibles estructuras acentuales o esquemas léxicos acentuales[12]. En un sentido más amplio el acento también cumple otras funciones en el español. Por ejemplo, el acento posee una función contrastiva,

[9] O también: agudas, graves, esdrújulas y sobresdrújulas, respectivamente.

[10] Aunque triviales, se mencionarán las reglas aquí para no dejar lugar a dudas.

[11] Aunque no posee tanta capacidad distintiva como la de los fonemas.

[12] En el inglés se denomina a este acento *lexical stress*, aunque no existen reglas simples ni representación ortográfica para su uso.

cuando se habla de palabras acentuadas e inacentuadas o culminativa, cuando se agrupan sílabas átonas en torno a una tónica. También suele utilizarse el acento en forma exagerada para hacer un énfasis o insistencia, haciendo también tónica a una sílaba que normalmente sería átona en la palabra.

Entonación

El término entonación se utiliza, en un sentido amplio, para hacer referencia a un conjunto de fenómenos lingüísticos relacionados directamente con la frecuencia fundamental (F_0) de las emisiones de voz. Desde un punto de vista lingüístico se han realizado estudios acerca de cómo se manifiesta la entonación en el lenguaje hablado. Por ejemplo, se pueden mencionar sus funciones integradora (de palabras a frases), distintiva (enunciados interrogativos o declarativos), demarcativa (como es el caso de las enumeraciones) y aquellas de nivel expresivo o incluso sociolingüístico (por ejemplo en las entonaciones regionales).

La diversidad de niveles a los que se estudia la entonación es significativamente más amplia que para el caso de la acentuación. Para describir estos niveles se puede atender al siguiente orden: F_0, tonema, grupo de entonación y curva melódica. La F_0 se mide a nivel de cada tramo de análisis y constituye el nivel más elemental de estudio, que también se corresponde con la menor duración en el análisis. El método para la medición de la F_0 se describirá detalladamente en la Sección 2.1.4. Esta es la variable física a partir de la cual se analiza la entonación en todos los otros niveles.

Los *tonemas* están más relacionados con la sílaba o algún suprasegmento entre los fonemas y las sílabas. Una primera distinción entre los tonemas de una palabra puede realizarse mediante los símbolos 'H' y 'L' que indican un tonema de frecuencia más alta (del inglés *High*) y uno de frecuencia más baja en relación a su contexto[13] (del inglés *Low*), respectivamente. Pero estos símbolos no se asocian directamente a sílabas o palabras, si bien en cuanto a su duración están más cerca de las primeras que de las segundas. La distinción entre tonemas altos y bajos no es suficiente para una descripción completa de la entonación en una lengua. A continuación se describirá brevemente uno de los estudios más citados para el inglés [Pierrehumberg, 1980] y su aplicación al español [Sosa, 1999].

Cuando estos símbolos se utilizan individual o conjuntamente en relación a una sílaba se habla de *acentos tonales*. Así, se realiza una primera distinción para aquellos tonemas que no se corresponden directamente con

[13] En general se considera el contexto a nivel silábico y, cuando es posible, dentro de una misma palabra.

una sílaba sino más bien con un *tono de juntura*, que se indican como 'H %' y 'L %'. La segunda distinción se realiza mediante un asterisco que indica la correspondencia con la sílaba tónica de la palabra. De esta forma se pueden encontrar 'H*' y 'L*' para las sílabas tónicas[14]. El último elemento notacional en esta descripción es el signo menos, que se utiliza para idicar un *acento del grupo de entonación* y representa a los cambios tonales que se realizan generalmente antes de una pausa en la elocución.

A partir de este conjunto de símbolos y un conjunto de reglas se ha descrito un sistema de transcripción conocido como ToBI (del inglés *Tone and Break Indices*). Este sistema fue desarrollado para el inglés y se ha adaptado al español permitiendo describir la entonación a partir de un *diccionario de estructuras tonemáticas*. Este diccionario puede dividirse en tres grandes categorías: cadencias de entonación (en inglés *falling pitch*), anticadencias de entonación (en inglés *rising pitch*) y mesetas de entonación (en inglés *level pitch*) [Almiñana, 1991, Portele y Heuft, 1997]. Con cada una de estas categorías se puede asociar un conjunto de estructuras tonemáticas que ocurren en el español [Sosa, 1999][15]:

- Cadencias de entonación:
  /H*L %/, /L*L %/, /H+L*L %/, /L+H*L %/ y /H+H*L %/
- Anticadencias de entonación:
  /H*H %/, /L*H %/, /H+L*H %/, /L+H*H %/ y /L*+HH %/
- Mesetas de entonación:
  /H*HL %/

El *grupo de entonación* se corresponde con la porción de la elocución que queda delimitada por dos pausas y constituye una unidad sintáctica más o menos larga. En general se entiende que un grupo de entonación comprende a dos o más palabras con un promedio de cinco sílabas. Sin embargo hay que destacar que las pausas que delimitan a los grupos de entonación no se dan de forma regular o en concordancia con estructuras de nivel superior. En general, si una frase no es muy larga (menos de 10 sílabas), es muy probable que forme un único grupo de entonación. La separación suele introducirse más bien en relación con factores semánticos y pragmáticos, como el foco

[14] Como se discutió en la sección anterior, en el caso general la sílaba tónica no tiene por qué tener mayor F_0.

[15] Se ha utilizado el signo '+' para separar las sílabas ya que cuando se encuentran dos símbolos juntos éstos corresponden a un mismo tonema.

y la distinción entre la nueva información que se introduce en la frase y la que ya era conocida de frases anteriores. Existen diferentes notaciones para los grupos de entonación pero generalmente todas consisten en separar o agrupar las palabras de la frase mediante algún caracter especial, por ejemplo:

- en [Quilis, 1993][16]:

  # La verdad # no penetra en un entendimiento rebelde #

- en [Sosa, 1999][17]:

  $[La\ verdad]_{gm}[no\ penetra\ en\ un\ entendimiento\ rebelde]_{gm}$

Cuando se extiende y restringe el análisis a frases completas se habla de *curva melódica* o entonación de la frase[18] cuya estudio se relaciona más con la modalidad del enunciado o el nivel expresivo que con el carácter demarcativo [Llorach, 1999]. Por ejemplo, la entonación distingue entre los enunciados los declarativos de los interrogativos. En el nivel expresivo, diferentes curvas melódicas caracterizan a los matices de cortesía o el caracter enfático de ciertos enunciados. Finalmente, cabe destacar que la curva melódica también contiene información de la región geográfica a la que pertenece el hablante.

1.3.4. Palabras, frases y significado

Al pasar de los tramos de análisis a los fonemas y a las sílabas no se hizo patente ningún aumento en el grado de abstracción del segmento considerado. Más aún, no apareció ningún indicio del significado en la emisión de voz. Sin embargo, al seguir formando estructuras (en promedio) más largas comenzarán a aparecer naturalmente asociaciones con nuestro entorno tangible.

Como se analizó anteriormente, puede dividirse una palabra en fonemas y, dada su pronunciación, también en fonos o alófonos. Luego se puede dividir una palabra como *relojero* en sílabas /re/-/lo/-/je/-/ro/ y aún no se alude a su relación con los relojes. Pero una siguiente división podría ser /reloj/-/ero/, en donde sí se hace una referencia explícita a la raíz: *reloj*. Esta división contempla los denominados *morfemas* y tiene una clara relación

[16] Esta es sólo una de las separaciones posibles.

[17] Este autor utiliza la denominación *grupo melódico* y de allí las letras utilizadas en los subíndices.

[18] En ciertas ocasiones se utiliza el término *entonación* en un sentido restringido haciendo referencia justamente a la curva melódica o entonación de la frase.

con los significados en la palabra. Si por ejemplo se divide la palabra *relojito* es posible encontrar el morfema lexical {reloj} que contiene el significado "dispositivo que sirve para medir el tiempo"; el morfema afijo {it} que da el significado de "diminutivo" y el morfema gramatical {o} que indica el "género masculino"[19].

Como ya se ha destacado antes, es común observar algún grado de superposición entre diferentes niveles. En este caso se superponen el fonema de la vocal /o/ y el morfema gramatical {o}. Para cada lengua existe un conjunto de reglas —y sus excepciones— que determinan la manera de concatenar morfemas para formar *palabras*[20]. Estas reglas son estudiadas en lingüística por la *morfología* y dan como resultado un *léxico* en el que se realiza una asociación entre diferentes combinaciones de evidencias acústicas y palabras [Ducrot y Todorov, 1984].

En un nivel siguiente se puede clasificar a las palabras de acuerdo a su función en las frases. El estudio de la gramática permite agrupar a las palabras según cuatro clases principales: el sustantivo, el adjetivo, el verbo y el adverbio. Se utiliza aquí a la *gramática* en un sentido restringido ya que también se habla de gramática incluyendo tanto al conocimiento lexicográfico como al sintáctico. Cuando se agrupan dos o más palabras para formar una unidad gramatical con sentido propio se constituye un *sirrema*. De forma similar a los grupos de entonación, las palabras que forman un sirrema permanecen unidas y no admiten pausas en su pronunciación. El conocimiento *sintáctico* comprende un conjunto de reglas para determinar las combinaciones de palabras que forman cadenas gramaticalmente correctas, como las frases.

Continuando el ascenso por los niveles de organización estructural, se sigue aumentando la duración promedio de los enunciados y se utiliza la denominación más general de *texto* o *discurso*. Aunque estas denominaciones suelen relacionarse más con el lenguaje escrito o hablado, en lingüística poseen un significado técnico más amplio, sin hacer este tipo de distinciones.

En un nivel de abstracción más elevado se encuentran la semántica y pragmática. La *semántica* estudia el significado que se codifica por medio del léxico y las estructuras gramaticales. En el estudio de la semántica es interesante observar cómo se accede otros niveles de organización estructural para precisar el significado de una palabra que puede poseer muchos. Por ejemplo, la palabra *puro* posee un diferente significado en la frase *sólo tomaré el vino*

[19]La lista puede continuar, por ejemplo se podría agregar que la ausencia de la {s} indica el número singular.

[20]Suele definirse a las palabras como "aquello que en la escritura aparece entre dos espacios en blanco".

puro y la frase *pero lo haré después de fumar este puro.* Además se puede requerir información de otras frases cercanas para poder lograr dar un significado preciso a una palabra. Por otro lado, suele utilizarse el "dominio" o "tema" del texto para esta determinación. Por ejemplo, en la frase *el puño estaba ensangrentado* la palabra *puño* puede adquirir significados diferentes si se trata de un relato deportivo de boxeo o si se trata de una conversación en una tintorería. Es este caso se hace uso del conocimiento *pragmático* y así se resuelve la ambigüedad. En todos estos ejemplos se han resuelto ambigüedades a nivel de las palabras pero también un adecuado conocimiento pragmático podría resolver ambigüedades en relación a la gramática. Como se discutirá más adelante, los diferentes niveles se utilizan en forma conjunta para resolver y dar sentido a los enunciados.

Se comienzan a traslucir a partir de la semántica y la pragmática unos conocimientos de nivel aún superior como el de los sentimientos, las intenciones o las ideas. También estos niveles cambian desde la acústica hasta la gramática con que se construyen las frases. En la parte inferior de la Figura 1.18 se insinúan las complejidades asociadas a los diferentes lenguajes y cómo cada nuevo nivel puede exigir una reestructuración de todos los anteriores. En esta figura no se han considerado otros niveles que también poseen su relevancia y fueron estudiados ampliamente. La estructuración de los textos en diferentes tipos de *documentos* también se relaciona con su dominio. Un texto en prosa puede dividirse en secciones, capítulos, párrafos y oraciones. Un texto en verso puede dividirse en cantos, estrofas y versos. Una vez impresas, las secuencias de prosa y verso pueden dividirse en páginas, volúmenes y colecciones.

1.4. Modelos para el reconocimiento del habla

Los modelos ocultos de Markov (MOM) constituyen una de las técnicas que se ha utilizado con más éxito en el RAH [Rabiner y Juang, 1986]. Principalmente, esta técnica ha permitido modelar adecuadamente la gran variabilidad en el tiempo de la señal de voz. En la terminología del RAH, con MOM suele hacerse referencia no sólo a la técnica de los modelos ocultos de Markov propiamente dicha, sino también a una larga lista de adaptaciones y técnicas asociadas que se fueron incorporando para solucionar el problema de RAH. En esta sección se tratarán los conceptos básicos de los MOM y su aplicación al RAH. En el próximo capítulo se tratarán con mayor detalle los aspectos relacionados directamente con la presente Tesis.

Cuando hablamos de RAH pensamos en un sistema automático que intenta transcribir en lenguaje escrito lo que un locutor ha expresado oralmente. Deben distinguirse en primer lugar los sistemas de reconocimiento del habla de los sistemas de *comprensión* del habla. Suele considerarse que la comprensión del habla es un concepto más amplio, que si bien incluye entre otras partes a un sistema de RAH, su objetivo es capturar la semántica del mensaje y no solamente transcribirlo en texto sino entenderlo correctamente. Comenzamos a ver así las potenciales aplicaciones de un sistema de RAH. En toda interfaz entre el hombre y las máquinas resulta de especial interés aprovechar aquel medio de comunicación que entre los hombres más uso ha tenido. Actualmente la mayoría de la gente sigue tecleando unas 60 palabras por minuto (en el mejor de los casos) cuando podría llegar a pronunciar unas 200 en el mismo tiempo. Las aplicaciones del RAH ya son un lugar común —tanto en ciencia como en ficción— por lo que invitamos al lector que se interese por una larga lista de éstas, a consultar el clásico libro [Rabiner y Juang, 1993].

Volvamos a considerar el proceso de la comunicación oral que tratamos al comienzo de este capítulo. Podríamos pensar que para cada texto el locutor activa un sistema y da como salida una determinada emisión sonora. Para comenzar a entender como se aplican los MOM al RAH imaginemos que para cada una de las posibles emisiones podemos encontrar un modelo capaz de imitar al sistema activado por el locutor. Es decir, un modelo que sea capaz de generar la misma emisión que generó el locutor a partir del texto que había en su mente. De esta forma vamos a suponer que contamos con tantos modelos como posibles emisiones pueda hacer el locutor y, para cada modelo un texto asociado. En caso de que conozcamos perfectamente estos modelos, podríamos utilizar el camino inverso para resolver el problema de RAH. Teniendo una determinada emisión del locutor nos pre-

guntaremos: ¿Cuál de todos mis modelos generará el sonido más parecido al que generó el locutor? Al encontrar el modelo que genera el sonido más parecido a la emisión del locutor entonces también habremos encontrado el texto, ya que habíamos dicho que todos los modelos estaban asociados a un determinado texto.

Existen dos observaciones de interés en este planteamiento. En primer lugar se debe entender que la solución propuesta es una solución que no parte de la utilización más corriente de los modelos. Generalmente utilizamos un modelo para obtener determinadas salidas a partir de ciertas entradas. Sin embargo, aquí estamos utilizando muchos modelos y una entrada fija asociada a cada uno (el texto). Luego, dada una señal de voz en particular, vemos cuál de todos genera una salida más parecida y damos como resultado la entrada de ese modelo. En segundo lugar, se puede ver claramente que este planteamiento para la solución del problema de RAH no es totalmente aplicable a casos reales, pues sería necesaria una cantidad infinita de modelos. Este problema se resuelve teniendo en cuenta que: 1) no es totalmente necesario abarcar toda la diversidad del habla (ni nosotros mismos podemos hacerlo) y 2) cada modelo no tiene por qué ser totalmente distinto e independiente de los demás.

El segundo punto puede adquirir mayor relevancia si tenemos en cuenta la organización estructural del habla en la que, como vimos en la sección anterior, existe una estructura jerárquica en la que pequeños componentes se combinan para formar otros de mayor complejidad. Esto quiere decir que sería posible construir una gran cantidad de modelos combinando un número razonable de pequeñas partes. A continuación veremos como modelar estas pequeñas partes por medio de los MOM y cómo generar grandes modelos a partir de ellas. También veremos cómo buscar el modelo cuya salida más se aproxima a la emisión del locutor y cómo encontrar los parámetros que mejor modelan un conjunto de emisiones para diversos locutores.

1.4.1. Modelos de autómatas finitos

Los autómatas son ampliamente utilizados para modelar secuencias temporales de variables discretas. Estos modelos poseen un conjunto de estados que representan las diferentes configuraciones internas en que se pueden encontrar. Si el conjunto de estados es finito entonces se habla de de autómatas finitos. Entre los estados debe distinguirse un estado inicial y un estado final. También es necesaria una función de transición de estados que determine la forma en que se realizan los cambios de un estado a otro. Para terminar de ver a los autómatas como un modelo, será necesario especificar

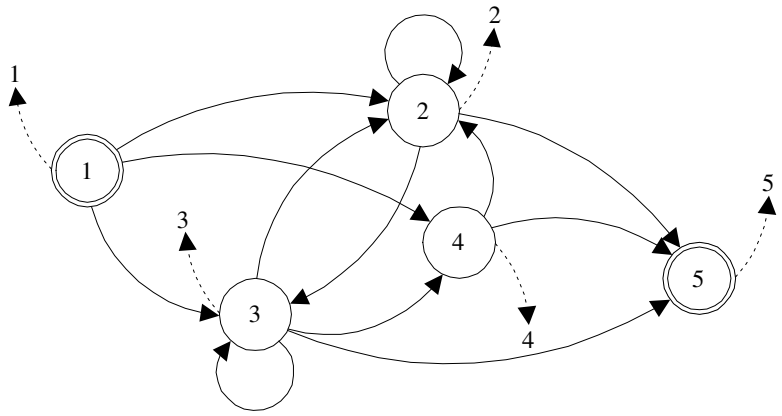

Figura 1.20. Diagrama de estados para un autómata finito. En este diagrama se puede observar un estado inicial (1), un estado final (5), los estados internos (2..4) y las flechas que indican las posibles transiciones entre los estados. También se han representado las salidas de cada estado en líneas de puntos que, para simplificar, coinciden con el número de estado.

entradas y salidas. En estos modelos cada estado puede asociar una salida para la entrada dada. La forma en que se realiza esta asociación da lugar a una gran variedad de autómatas. Por ejemplo, un caso sencillo puede consistir en que cada estado posea una función de salida que selecciona entre los elementos de un conjunto finito de símbolos de salida.

Para representar la estructura interna de un modelo de autómatas suele utilizarse un diagrama de estados como el de la Figura 1.20. En este diagrama se pueden todos observar los estados, sus salidas y las flechas que indican las posibles transiciones entre ellos.

¿Cómo se puede utilizar este modelo de autómata finito? Para entender un ejemplo sencillo se puede simplificar la función de salida de forma que de como resultado el número del estado y utilizar una función de transición que simplemente elija al estado siguiente como aquel que posee el número más cercano a la entrada actual. Así, dada una secuencia de entrada: 2, 2, 2, 4, 4, 4 se obtendrá como secuencia de estados: 1, 2, 2, 3, 4, 5 y idénticamente como secuencia de salida: 1, 2, 2, 3, 4, 5.

Otro tipo interesante de autómata es aquél que puede albergar una descripción probabilística del fenómeno que modela. Para estos autómatas es necesario realizar algunas definiciones particulares a partir de los elementos

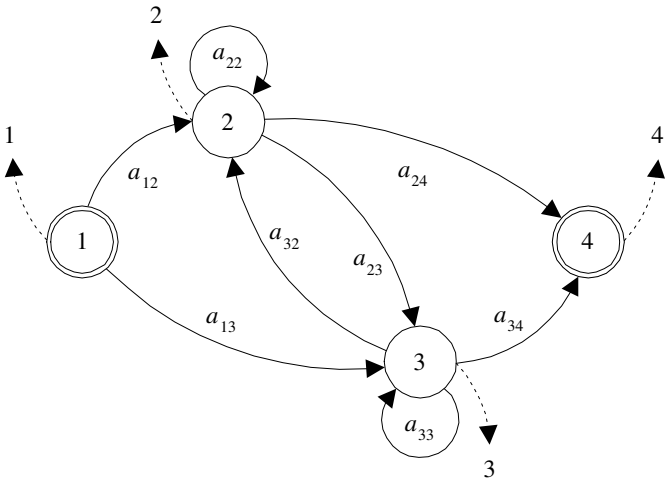

Figura 1.21. Diagrama de estados para un autómata probabilístico. Las probabilidades de transición desde el estado i al estado j se indican como a_{ij}. A cada estado se asocia un símbolo del conjunto finito de salidas. En este ejemplo la salida del estado corresponde simplemente con su número

básicos de un autómata finito. En lugar de función de transición de estados se habla de probabilidades de transición entre estados. Es común utilizar para estas probabilidades la notación a_{ij}: probabilidad de pasar al estado j dado que se está actualmente en el estado i. En cuanto a las salidas de este modelo estadístico, cada estado se asocia a uno de los posibles símbolos de un *conjunto de salidas*. Un ejemplo sencillo se puede observar en la Figura 1.21.

En este caso también cabe preguntarse: ¿Cómo se pueden utilizar estos modelos de autómatas probabilísticos? Aquí el planteamiento se invierte y se utilizan estos modelos para encontrar la probabilidad de que una determinada secuencia de salida haya sido generada por él[21]. Es decir, a partir de una secuencia de salidas observadas en el mundo real, se plantea conocer qué probabilidad existe de que el modelo en cuestión la haya generado. Para dar un ejemplo sencillo se puede suponer que en el modelo de la Figura 1.21:

[21]Estar inversión está orientada hacia la particular forma de utilizar los modelos en RAH, como se discutió en la introducción de esta sección. De esta forma se va introduciendo progresivamente la perspectiva de MOM para RAH.

$$A = \{a_{ij}\} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 0 & 1/4 & 1/4 & 1/2 \\ 0 & 1/2 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Ahora la pregunta es: ¿Qué probabilidad existe de que este modelo genere la secuencia 1, 2, 2, 3, 2, 4? Para resolver este problema deben considerarse las transiciones de estado $1 \rightarrow 2$, $2 \rightarrow 2$, $2 \rightarrow 3$, $3 \rightarrow 2$ y $2 \rightarrow 4$. Así, se obtiene la probabilidad total para la secuencia mediante la multiplicación:

$$p_{122324} = a_{12}a_{22}a_{23}a_{32}a_{24} = \frac{1}{2}\frac{1}{4}\frac{1}{4}\frac{1}{2}\frac{1}{2} = \frac{1}{128}$$

Este modelo probabilístico es también denominado *modelo de Markov* (MM). Si el tiempo transcurre entre cada transición a intervalos discretos, se dice entonces que se trata de un MM de tiempo discreto. Si además se sigue en la presunción de que las probabilidades de transición sólo dependen de los estados origen y destino, se está en presencia de un proceso de primer orden que suele denominarse cadena de Markov. Como las probabilidades de transición no se modifican con el tiempo también se trata de un sistema invariante en el tiempo o, en la terminología de la teoría de probabilidades, una cadena de Markov homogénea. Finalmente, observando el hecho de que en un MM no se especificaba una entrada, se llega a la denominación de fuente de Markov, muy utilizada en teoría de comunicaciones.

Modelos ocultos de Markov

En cada estado de un MM se emite un determinado símbolo del conjunto de salidas posibles. Es decir que la función de salida simplemente asigna uno de los símbolos dependiendo del estado en que se encuentre el modelo. Es por esto que un MM es también conocido bajo la denominación de modelo *observable* de Markov: a partir de la salida se puede "observar" en que estado se encuentra el modelo. El hecho de que en cada estado se pueda observar un único símbolo es un limitación importante que reduce las posibilidades de aplicación de los MM. Para aumentar su capacidad de modelado, se ha propuesto una extensión en donde la función que asocia a cada estado una salida sea una distribución de probabilidades sobre todas las posibles salidas. Ahora existirá un nuevo parámetro $b_j(k)$ que describe

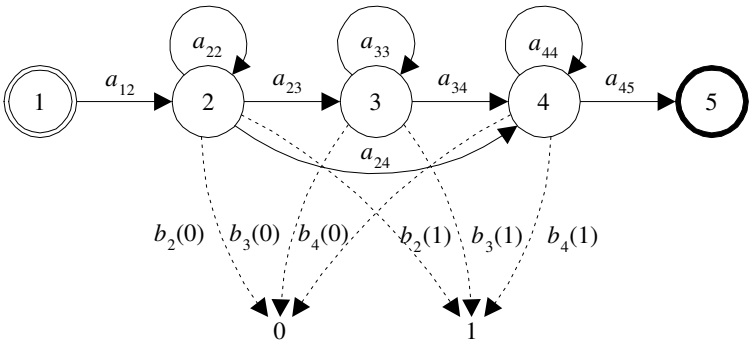

Figura 1.22. Diagrama de estados para un modelo oculto de Markov comúnmente utilizado en RAH. El estado 1 es el estado inicial y el 5 el final y se denominan no emisores. Las flechas en líneas continuas indican las posibles transiciones entre estados. Las flechas en líneas de puntos indican las probabilidades de observación para cada estado. En esta configuración se puede observar la particularidad de que las transiciones se dan solamente de izquierda a derecha.

la probabilidad de que el estado j observe el símbolo k del conjunto de salidas[22]. En estas condiciones nunca se podrá saber con certeza en que estado esta el modelo observando solamente su salida. El funcionamiento interno del modelo queda "oculto" y es por eso que se lo denomina modelo *oculto* de Markov. Los MOM más utilizados en RAH poseen una estructura muy simple denominada de izquierda a derecha. Un ejemplo de estas estructuras se muestra en la Figura 1.22.

Si para el modelo de la Figura 1.22 se dan los parámetros:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 1/3 & 1/5 & 2/3 & 0 \\ 0 & 2/3 & 4/5 & 1/3 & 0 \end{bmatrix}$$

una de las preguntas más importantes está relacionada nuevamente con la probabilidad de generar una secuencia observada: ¿Qué probabilidad existe de que este modelo genere la secuencia 0, 0, 1, 0? La respuesta no es tan obvia como en los casos anteriores. En este caso no se puede inferir directamente la

[22] En algunos casos suele hablarse de probabilidades de *emisión* en lugar de probabilidades de *observación*.

secuencia de estados que debería haber seguido el modelo para generar esa salida ya que el modelo está "oculto". Si se analiza un poco más el problema se puede deducir que la secuencia de estados que genera esa secuencia de salida no es única: ahora cada estado puede emitir cualquiera de los símbolos del conjunto de salidas (aunque con distinta probabilidad). Para resolver este problema es necesario analizar todas las posibles secuencias que pasen por 4 estados emisores y sus probabilidades asociadas (véase la Tabla 1.2). Una forma alternativa para representar estas transiciones de estados es la que se muestra en el diagrama de la Figura 1.22.

Secuencias de de estados	Probabilidades de transición	Probabilidades de observación	Probabilidades de la secuencia
1, 2, 2, 2, 4, 5	$1\ \frac{1}{4}\frac{1}{4}\frac{1}{2}\frac{1}{2} = \frac{1}{64}$	$\frac{1}{3}\frac{1}{3}\frac{2}{3}\frac{2}{3} = \frac{4}{81}$	$\frac{1}{1296}$
1, 2, 2, 3, 4, 5	$1\ \frac{1}{4}\frac{1}{4}\frac{1}{2}\frac{1}{2} = \frac{1}{64}$	$\frac{1}{3}\frac{1}{3}\frac{4}{5}\frac{2}{3} = \frac{8}{135}$	$\frac{1}{1080}$
1, 2, 2, 4, 4, 5	$1\ \frac{1}{4}\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{32}$	$\frac{1}{3}\frac{1}{3}\frac{1}{3}\frac{2}{3} = \frac{2}{81}$	$\frac{1}{1296}$
1, 2, 3, 3, 4, 5	$1\ \frac{1}{4}\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{32}$	$\frac{1}{3}\frac{1}{5}\frac{4}{5}\frac{2}{3} = \frac{8}{225}$	$\frac{1}{900}$
1, 2, 3, 4, 4, 5	$1\ \frac{1}{4}\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{32}$	$\frac{1}{3}\frac{1}{5}\frac{1}{3}\frac{2}{3} = \frac{2}{135}$	$\frac{1}{2160}$
1, 2, 4, 4, 4, 5	$1\ \frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{16}$	$\frac{1}{3}\frac{2}{3}\frac{1}{3}\frac{2}{3} = \frac{4}{81}$	$\frac{1}{324}$
Probabilidad Total			$\sum = \frac{77}{10800} \approx 0{,}007$

Tabla 1.2. Probabilidad para todos los caminos permitidos para una secuencia de 4 emisiones en el ejemplo de la Figura 1.22. Cuando se habla de caminos permitidos se hace referencia a aquellos caminos que no involucren una probabilidad nula.

1.4.2. La secuencia más probable

En la mayoría de los casos es suficiente con encontrar sólo la mejor secuencia y su probabilidad asociada. Con este fin, existen algoritmos que permiten ahorrar muchos cálculos y entre ellos, uno de los más utilizados es el algoritmo de Viterbi. En este algoritmo la idea central es recorrer el diagrama de transiciones de estados a través del tiempo, almacenando para cada estado solamente la máxima probabilidad acumulada y el estado anterior desde

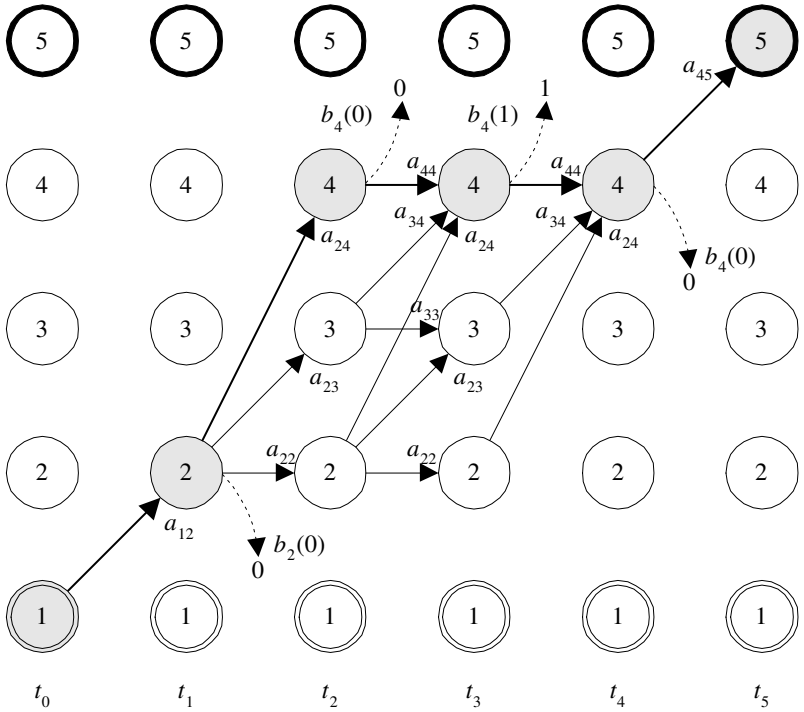

Figura 1.23. Diagrama de transiciones de estado para el modelo la Figura 1.22 y una secuencia de 4 observaciones. En este diagrama se indican todos los caminos posibles y se destaca el camino más probable encontrado mediante el algoritmo de Viterbi.

el que se llega con esta probabilidad. La máxima probabilidad acumulada se obtiene multiplicando la probabilidad de observación del estado por la máxima probabilidad acumulada entre todos los caminos que llegan hasta él. Se entenderá mejor como funciona este algoritmo de definición recursiva mediante un ejemplo.

Para este ejemplo se seguirá el diagrama de la Figura 1.23, sin olvidar que la secuencia de salida deseada es 0, 0, 1, 0. Se comienza en el estado 1, asignando una probabilidad acumulada $p_1 = 1$ y al pasar al estado 2 la probabilidad acumulada es:

$$p_{12} = b_2(0)\,[p_1 a_{12}] = \frac{1}{3}[1 \times 1] = \frac{1}{3}$$

Desde el estado 2 se puede pasar al 2, al 3 o al 4 obteniendo:

$$p_{122} = b_2(0)\,[p_{12}a_{22}] = \frac{1}{3}\left[\frac{1}{3}\frac{1}{4}\right] = \frac{1}{36}$$

$$p_{123} = b_3(0)\,[p_{12}a_{23}] = \frac{1}{5}\left[\frac{1}{3}\frac{1}{4}\right] = \frac{1}{60}$$

$$p_{124} = b_4(0)\,[p_{12}a_{24}] = \frac{2}{3}\left[\frac{1}{3}\frac{1}{2}\right] = \frac{1}{9}$$

Desde el estado 2 en el tiempo t_2 se puede pasar a los estados 2, 3, y 4:

$$p_{1222} = b_2(1)\,[p_{122}a_{22}] = \frac{2}{3}\left[\frac{1}{36}\frac{1}{4}\right] = \frac{1}{216}$$

$$p_{1223} = b_3(1)\,[p_{122}a_{23}] = \frac{4}{5}\left[\frac{1}{36}\frac{1}{4}\right] = \frac{1}{180}$$

$$p_{1224} = b_4(1)\,[p_{122}a_{24}] = \frac{1}{3}\left[\frac{1}{36}\frac{1}{2}\right] = \frac{1}{216}$$

Desde el estado 3 en tiempo t_2 se puede pasar a los estados 3 y 4:

$$p_{1233} = b_3(1)\,[p_{123}a_{33}] = \frac{4}{5}\left[\frac{1}{60}\frac{1}{2}\right] = \frac{1}{600}$$

$$p_{1234} = b_4(1)\,[p_{123}a_{34}] = \frac{1}{3}\left[\frac{1}{60}\frac{1}{2}\right] = \frac{1}{360},$$

y desde el estado 4 en el tiempo t_2 sólo se puede pasar al estado 4:

$$p_{1244} = b_4(1)\,[p_{124}a_{44}] = \frac{1}{3}\left[\frac{1}{9}\frac{1}{2}\right] = \frac{1}{54}$$

Habiendo llegado al tiempo t_3, a partir de cualquiera de los estados solamente es posible pasar al estado 4:

$$p_{12224} = b_4(0)\,[p_{1222}a_{24}] = \frac{1}{3}\left[\frac{1}{216}\frac{1}{2}\right] = \frac{1}{1296}$$

$$\begin{aligned} p_{12?24} &= b_4(0)\,\text{máx}\,\{[p_{1223}a_{34}]\,[p_{1233}a_{34}]\} \\ &= b_4(0)\,\text{máx}\,\{p_{1223}, p_{1233}\}\,a_{34} \\ &= \frac{1}{5}\,\text{máx}\left\{\frac{1}{216}, \frac{1}{600}\right\}\frac{1}{2} \\ &= \frac{1}{5}\frac{1}{216}\frac{1}{2} = \frac{1}{2160} \\ &= p_{12234} \end{aligned}$$

$$\begin{aligned} p_{12?44} &= b_4(0)\,\text{máx}\,\{[p_{1224}a_{44}]\,[p_{1234}a_{44}]\,[p_{1244}a_{44}]\} \\ &= b_4(0)\,\text{máx}\,\{p_{1224}, p_{1234}, p_{1244}\}\,a_{44} \\ &= \frac{2}{3}\,\text{máx}\left\{\frac{1}{216}, \frac{1}{360}, \frac{1}{54}\right\}\frac{1}{2} \\ &= \frac{1}{5}\frac{1}{54}\frac{1}{2} = \frac{1}{162} \\ &= p_{12444} \end{aligned}$$

Finalmente, ya en el tiempo t_4 la única opción es pasar al estado 5 que, al igual que el estado 1, es no emisor (ver Figura 1.22) y no es necesario considerar la probabilidad de observación:

$$\begin{aligned} p_{12??45} &= \text{máx}\,\{[p_{12224}a_{45}]\,[p_{12234}a_{45}]\,[p_{12444}a_{45}]\} \\ &= \text{máx}\,\{p_{12224}, p_{12234}, p_{12444}\}\,a_{45} \\ &= \text{máx}\left\{\frac{1}{1296}, \frac{1}{2160}, \frac{1}{162}\right\}\frac{1}{2} \\ &= \frac{1}{162}\frac{1}{2} = \frac{1}{324} \\ &= p_{124445} \end{aligned}$$

Así, se arriba a la misma conclusión que en el análisis exhaustivo de la Tabla 1.2: de todos los caminos posibles la mejor secuencia de estados es la 1, 2, 4, 4, 4, 5 y posee una probabilidad de 1/324.

Como se puede observar, se ha ahorrado un gran número de cálculos con este método. En la búsqueda exhaustiva de la Tabla 1.2 se realizaron 48 multiplicaciones mientras que en el ejemplo de Viterbi sólo fueron 27. Además hay que notar que esta diferencia se incrementa notablemente cuando aumenta el número de estados emisores o la cantidad de observaciones. Esto es debido a que, gracias a que sólo se sigue adelante por los caminos que tienen máxima probabilidad, muchos caminos no se analizan. Se puede ver en este ejemplo que a partir del estado 4 y el tiempo t_3, los caminos 1, 2, 2, 4, ?, ? y 1, 2, 3, 4, ?, ? ya no se analizan. Si se conoce una buena forma de llegar a ese estado, solamente se utilizará esta forma. Esto no implica que se deje de lado la evaluación de alguno de los caminos que deriva del estado en cuestión y así el método ahorra muchos cálculos sin perder generalidad.

1.4.3. Estimación de los parámetros del modelo

Hay que notar que ha quedado de lado una cuestión importante: ¿Cómo se estiman las probabilidades de transición y observación que mejor modelan un conjunto dado de secuencias observadas? Una forma muy intuitiva de entender el entrenamiento es pensar que, si el algoritmo de Viterbi provee la secuencia de estados más probable para una secuencia de símbolos de salida observada, entonces es posible estimar las probabilidades de transición y observación a partir de los símbolos que han quedado asignados a cada estado. Si se posee un conjunto de secuencias observadas para el entrenamiento, se puede encontrar todas las secuencias de estados más probables y contabilizar las veces que se ha pasado al estado j a partir del estado i. A partir de estas cuentas es posible obtener una buena estimación de la probabilidad de pasar al estado a_{ij}.

De forma similar, a partir de las secuencias más probables encontradas con el algoritmo de Viterbi, se puede contar la cantidad de veces que el k-ésimo símbolo observable a sido asignado al j-ésimo estado del modelo. Esta cuenta puede ser utilizada para obtener una buena estimación de la probabilidad de que el j-ésimo estado del modelo emita el k-ésimo símbolo observable, es decir, $b_j(k)$.

Mediante una aplicación repetitiva de la búsqueda de la mejor secuencia y posterior reestimación de las probabilidades es posible entrenar el modelo, dado un conjunto de secuencias observadas. Inicialmente se pueden considerar iguales probabilidades para todas las transiciones posibles hacia un estado. De forma similar se pueden considerar inicialmente iguales probabilidades de observación para todos los estados, obtenidas a partir de la cantidad de veces que aparece cada símbolo en el conjunto de secuencias de

entrenamiento.

Este método de búsqueda y reestimación se conoce como algoritmo de entrenamiento de Viterbi y es muy rápido en la práctica. Sin embargo, cuando se aplica el algoritmo de Viterbi se trabaja sobre una aproximación de la probabilidad del modelo para cada símbolo de cada secuencia (se ha reemplazado la sumatoria por el máximo). Es así como se obtiene la pertenencia de un símbolo observado a un estado como una función que sólo puede valer 1 o 0 (el símbolo corresponde al estado en cuestión o no corresponde). Si se utiliza una mejor estimación de esta probabilidad, es posible obtener un función de pertenencia con salida no binaria y utilizarla para pesar las evidencias de las secuencias de entrenamiento en la reestimación de las probabilidades del modelo. Éste es el algoritmo de reestimación de Baum-Welch y se tratará en detalle en el siguiente capítulo.

1.4.4. Modelado acústico de la voz

Para seguir aproximando las ideas de MOM al RAH se estudiará cómo utilizarlos para modelar una emisión acústica. Un modelo como el de la Figura 1.22 podría utilizarse para modelar un fonema y en RAH se denomina *modelo acústico* (MA). Sin embargo, hay que tener en cuenta que los MOM tal como se presentaron hasta el momento, sólo pueden modelar secuencias discretas de símbolos. Este implica dos niveles de discretización. Por un lado se requiere que los sucesos en el tiempo ocurran a intervalos discretos. Por otro lado se requiere que las manifestaciones de dichos sucesos estén dentro de un conjunto finito de símbolos.

La restricción relativa a la discretización del tiempo puede verse fácilmente superada si se considera el análisis por tramos como se describió en la primera parte de la organización estructural del habla (Sección 1.3.1, página 25). De esta forma, las observaciones del fenómeno se dan a intervalos regulares de tiempo. En cuanto a la necesidad de que las observaciones pertenezcan a un conjunto finito de símbolos, existen dos posibles alternativas: 1) representar todos los tramos de voz similares mediante un único símbolo y 2) modificar el modelo para que permita modelar valores continuos en las observaciones.

Si se opta por la primera alternativa, luego de dividir la emisión de voz se en tramos se busca un símbolo que represente a cada uno. Este proceso suele incluirse en el denominado *pre-procesamiento* de la señal de voz. Básicamente, una primera etapa del pre-procesamiento se encarga de obtener una representación adecuada del tramo mediante, por ejemplo, un análisis

en frecuencia[23] [Rabiner y Gold, 1975]. Luego, una segunda etapa clasifica el tramo de análisis y le asocia uno de los símbolos con que trabaja el MOM. Esta clasificación también puede entenderse como una cuantización, donde un grupo de valores reales se convierte en un número entero dentro de un rango acotado [Gray, 1984]. En la Figura 1.24 se observan las etapas principales y las señales involucradas. En primer lugar está la señal de voz y luego se esquematiza el análisis por tramos en el tiempo. A continuación cada segmento se analiza en el dominio de la frecuencia y finalmente se realiza una clasificación o cuantización vectorial que da por resultado una secuencia de elementos discretos.

Si se posee un MOM para cada una de las unidades acústicas a modelar (en general fonemas, sílabas o palabras), entonces se podrá aplicar el algoritmo de Viterbi y obtener el mejor camino de cada MOM. Finalmente, el MOM cuyo mejor camino presente la mayor probabilidad será el que determine de qué unidad acústica se trataba.

El esquema que hasta aquí se presenta es el que se conoce como MOM *discreto*, debido a que lo que se modela realmente es una secuencia de símbolos discretos a través de probabilidades de observación discretas. Volviendo a la segunda alternativa para solucionar estas restricciones, se elimina la etapa de cuantización vectorial y se definen los modelos ocultos de Markov *continuos* (MOMC), que utilizan directamente los vectores procedentes del análisis en frecuencia de los tramos de voz. Para esto es necesario replantear las probabilidades de observación de cada estado como, por ejemplo, vectores que contienen las medias y desviaciones para cada elemento del segmento de voz que modelan[24]. De esta manera cada estado de cada modelo tendría sus propias distribuciones de probabilidad que modelan las características acústicas de la voz [Liporace, 1982]. Finalmente, existe una alternativa intermedia denominada modelos ocultos de Markov *semicontinuos* (MOMSC), en donde todos los modelos comparten un conjunto fijo de distribuciones de probabilidad.

1.4.5. El modelo de lenguaje y el modelo compuesto

Cuando se habla del modelo de lenguaje (ML), se sitúa el estudio en niveles superiores al de las características acústicas, por encima de los fone-

[23] Como se describió antes, muchas de las características que permiten una clasificación de los sonidos del habla se hacen evidentes en el dominio de la frecuencia (Secciones 1.2.4 y 1.3.2).

[24] Este es un ejemplo muy simplificado, en el próximo capítulo se tratarán con detalle los MOMC.

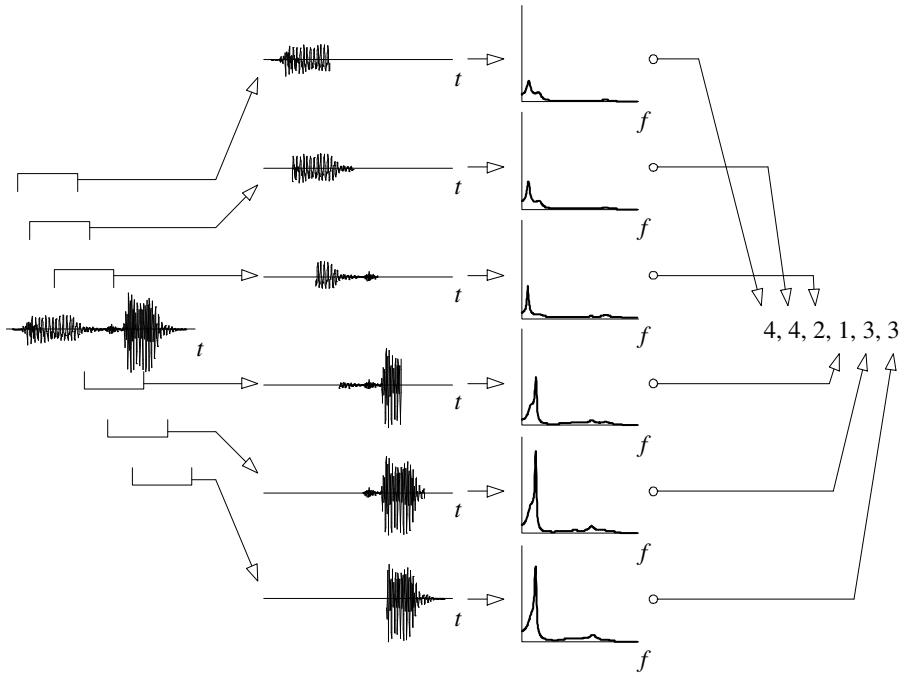

Figura 1.24. Procesamiento necesario para utilizar modelos ocultos de Markov discretos en reconocimiento automático del habla. Se pueden observar las etapas principales y las señales involucradas. En primer lugar está la señal de voz y luego se esquematiza el análisis por tramos. A continuación cada segmento se analiza en el dominio de la frecuencia y finalmente se realiza una clasificación que da por resultado una secuencia de elementos discretos. Los modelos ocultos de Markov continuos no requieren esta última etapa y trabajan directamente con los vectores en el dominio transformado.

mas y los suprasegmentos. Ahora interesan las palabras y la forma en que se combinan para formar frases. Siguiendo con la idea de los autómatas probabilísticos (finitos), es posible imaginar un autómata en el que cada estado represente (o emita) una palabra. En la Figura 1.25 se puede observar una estructura que respeta la idea general de un autómata probabilístico como el de la Figura 1.21 (página 44), utilizado para modelar secuencias temporales de palabras. Estas estructuras son conocidas como *gramáticas* en la teoría de lenguajes formales y conservan ese nombre en la jerga del RAH.

Sin embargo, se puede observar que la secuencia de estados de una de estas gramáticas es también una cadena de Markov y así se pueden extender los formalismos de MOM para incluir estas representaciones en un nivel su-

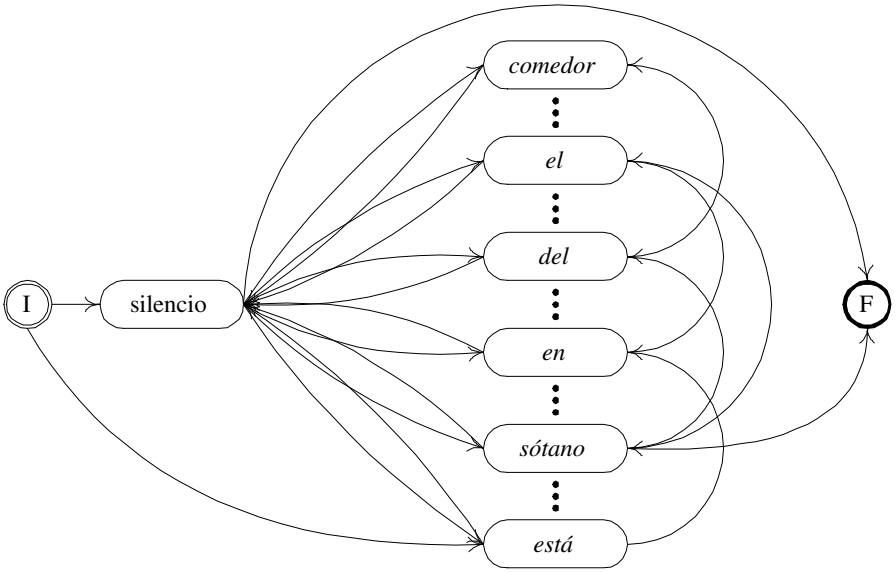

Figura 1.25. Modelo de lenguaje. Los estados de inicio y finalización se indican con las letras I y F, respectivamente.

perior al acústico (para ver los detalles formales acerca de esta generalización puede consultarse [Deller et al., 1993]). A partir de una descripción fonética de cada palabra, conocida como *diccionario fonético*, se podrían formar las palabras de este ML concatenando los MA de los diferentes fonemas. Finalmente se construiría un *modelo compuesto* (MC) capaz de modelar cualquier frase, desde los aspectos fonéticos más elementales hasta las complejidades del lenguaje hablado. En la Figura 1.26 se pueden observar los tres niveles de la composición: el ML, el diccionario fonético y el MA.

Mediante este MC es posible formar modelos para diferentes frases y evaluar, con una extensión del algoritmo de Viterbi, las probabilidades de cada frase para una emisión de voz dada. El proceso de reconocimiento culmina eligiendo el modelo de la frase que mayor probabilidad posea y dando como resultado el texto con que se formó la frase. Cabe aclarar que, nuevamente, la búsqueda sobre todas las frases posibles no se realiza de forma exhaustiva. Para esto existe una gran variedad de algoritmos que organizan y recorren de diferentes formas la expansión del MC.

Resta por comentar brevemente la extensión de los algoritmos de entrenamiento para el MC. Existen dos conjuntos de parámetros a estimar durante el entrenamiento: las probabilidades de transición y observación de

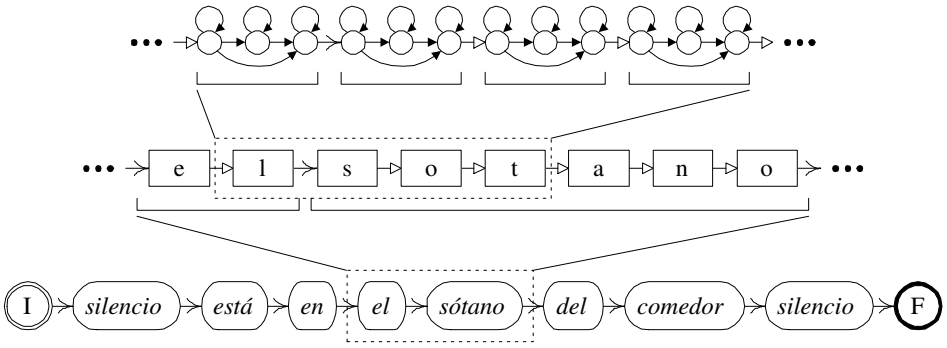

Figura 1.26. Modelo compuesto para la frase: *Está en el sótano del comedor*. Se pueden observar los tres niveles de la composición: los estados del modelo acústico, el diccionario fonético y el modelo de lenguaje. En los modelos acústicos se han eliminado los estados no emisores para simplificar el esquema.

los MA y las probabilidades de transición del ML. Estas estimaciones se realizan separadamente, es decir, se estiman las primeras dejando fijo el ML y viceversa. Para la estimación de las probabilidades de los MA, a partir de una de las frases de entrenamiento y dada su transcripción en texto es posible formar un MC para esta frase y luego aplicar el algoritmo de entrenamiento sobre este gran modelo, tal como se aplicó en el caso de un pequeño MOM. Los mismos modelos de fonemas o palabras pueden concatenarse para formar otro MC de frase y nuevamente realizar un ajuste mediante el algoritmo de entrenamiento. Las probabilidades que corresponden al ML, que habían quedado fijas durante este proceso, son estimadas directamente del texto de las frases de entrenamiento, contando la cantidad de veces que aparece una determinada secuencia y asignado una probabilidad a las transiciones que es proporcional a esta cuenta. Se verán en detalle los métodos de entrenamiento y estimación de las probabilidades del ML en el siguiente capítulo.

1.5. Acentuación y reconocimiento del habla

En esta última sección se resumen tres aspectos fundamentales para la Tesis: el contexto en el que se desarrolla la investigación, las limitaciones de las técnicas actuales y la forma en que se evitarán ciertos errores mediante aportes novedosos. Las primeras partes son una discusión en torno a diferentes temas tratados en este capítulo y su relación con el RAH. Luego se intentará hacer un especial énfasis en las falencias y limitaciones de los

sistemas de RAH basados en MOM para dar lugar a la innovación, describir el problema y plantear la estructura general de la investigación. En la última sección se resumen los objetivos de la Tesis.

1.5.1. Complejidad en el reconocimiento del habla

En las discusiones acerca de la organización estructural del habla se dejó ver lateralmente la complejidad que entraña el lenguaje hablado. También cuando se describieron los mecanismos naturales de generación y percepción del habla pudo vislumbrarse la altísima complejidad que entraña el procesamiento del habla en el ser humano. Como se discutía en el Prefacio, queda mucho por hacer en RAH.

Cuatro direcciones de avance en RAH

Hasta el momento, las investigaciones en este área se han orientado a incrementar cuatro variables fundamentales: el *vocabulario* reconocido; la *naturalidad* con que se le habla al sistema; la cantidad y variedad de *locutores* para los que el reconocimiento es aceptable y la *robustez* del sistema ante ruidos ambientales y otras condiciones adversas [Marini, 1989].

Los primeros sistemas de RAH fueron desarrollados en los laboratorios Bell y RCA en la década del 50 y reconocían 10 palabras, para un único hablante, con un porcentaje de error que podía llegar al 50 %. En la década del 60 ya se reconocía un vocabulario de unas pocas decenas de palabras. Alrededor de los años 70 se conocieron algunos sistemas comerciales capaces de reconocer unas 100 palabras, lo que actualmente se conoce como un vocabulario pequeño [Ferguson, 1980]. Años más tarde, hacia fines de los 80, los investigadores de IBM presentaron un sistema que podía reconocer unas 20000 palabras pronunciadas aisladamente. Llegando a la actualidad, los sistemas pueden reconocer vocabularios pequeños (menos de 100 palabras) y medianos (entre 100 y 20000 palabras) con tasas de error muy bajas. Por ejemplo, en condiciones de laboratorio se puede reconocer un vocabulario pequeño con errores menores al 1 %, para una gran variedad de locutores. En estas condiciones, los errores para un vocabulario mediano en iguales condiciones pueden estar entre el 1 y el 5 %. Pero actualmente los desafíos consisten en reconocer con vocabularios grandes y muy grandes. En el primer caso se intentan reconocer entre 20000 y 100000 palabras y ciertos sistemas comerciales reconocen más de 60000 palabras en condiciones especiales[25]. En

[25]Como se explicará luego, sin ruido ambiental y con un proceso previo de adaptación al hablante.

el caso de un vocabulario muy grande deben considerarse hasta un millón de palabras[26].

En cuanto a la naturalidad con que puede hablársele al sistema de RAH, los avances son más recientes. En todos los sistemas de las primeras épocas del RAH se requería que las palabras fueran pronunciadas en forma aislada, con una clara separación entre ellas. Sin embargo, en el lenguaje natural no existen separadores entre las unidades, ya que en muchos casos no existen silencios ni siquiera entre las palabras. Los trabajos pioneros en este sentido fueron realizados por [Reddy, 1966], quién introdujo el concepto de reconocimiento de habla continua. Sin embargo, en los años 60 y 70 siguió predominando la palabra como unidad de reconocimiento y se construyeron sistemas que permitían reconocer palabras conectadas con una pequeña pausa entre ellas. A fines de los 80 IBM presentó un sistema experimental capaz de reconocer un vocabulario de 5000 palabras pronunciadas en forma más natural. Como se vio en secciones anteriores, los reconocedores de voz continua utilizan a los fonemas como unidad mínima y forman las palabras a partir de un diccionario de pronunciaciones. Los sistemas actuales pueden reconocer un vocabulario mediano en habla continua con una tasa de error cercana al 5 %. Pero a pesar de tratarse de habla continua, en estos sistemas sigue siendo necesario que la pronunciación se realice cuidadosamente (por ejemplo, frases leídas). Si se asume que un reconocedor tiene un error del 5 % para frases leídas, este error puede llegar al 30 % cuando se utiliza con habla espontánea. Cuando se habla de naturalidad en la emisión se espera mucho más que habla continua. Por un lado, el habla natural debe admitir una gran complejidad en la estructura gramatical. Por otro lado, el sistema debe estar preparado para manejar sucesos comunes en el habla espontánea, como las pausas repentinas y la repetición de términos, y otros más ajenos al habla como un estornudo, la tos, el hipo o un bostezo. La complejidad de la estructura gramatical suele restringirse diseñando sistemas orientados a una tarea muy específica, como responder consultas sobre horarios de vuelos o pedidos de comida. Por ejemplo, si un sistema que puede modelar adecuadamente la complejidad gramatical de la tarea posee un error del 5 %, ese error puede llegar al 20 % en las mismas condiciones (vocabulario, ruido, locutores, etc.) pero con una complejidad gramatical que escapa a su poder de expresión como modelo. En la actualidad se intenta reconocer habla espontánea como la de una conversación telefónica sin un contexto temático particular. En estos casos los errores de reconocimiento

[26] Por ejemplo, se pueden ver las características del programa "OpenSpeech Recognizer" en `http://www.speechworks.com/products/speechrec/openspeechrecognizer.cfm`.

pueden superar el 50 %, haciendo a estos sistemas prácticamente inútiles en aplicaciones reales.

El habla presenta un amplio margen de variabilidad dependiendo de diversos aspectos relacionados con el hablante. Dentro de un conjunto de emisiones de un mismo locutor se pueden encontrar modificaciones sobre fonemas y palabras incluso en idénticas condiciones. En el fenómeno de coarticulación se puede apreciar un caso en el que una unidad fonética se pronuncia de forma diferente dependiendo del contexto en el que se encuentre. Pero más complejas y diversas son las variaciones que ocurren con múltiples locutores. En primer lugar, la voz de cada locutor posee características propias dadas las medidas antropométricas de su aparato fonador. Un primer nivel de distinción se hace entre hombres, mujeres y niños. En los primeros sistemas de RAH todo el entrenamiento y prueba se realizaba con un único locutor. Luego se conocieron los sistemas multilocutor, en los que el entrenamiento y prueba se realizaba con el mismo conjunto de locutores. En relación a las variaciones introducidas por diferentes hablantes existen dos tendencias principales en la investigación: la normalización del hablante y la adaptación al hablante. En el primer caso se intenta modificar las representaciones de los tramos de voz de forma tal de eliminar las particularidades del hablante y entregar al reconocedor una voz "normalizada". En el segundo caso lo que se intenta es adaptar rápidamente los modelos en base a la menor cantidad posible de frases del nuevo hablante. En la actualidad se avanza hacia los reconocedores independientes del hablante pero hay mucho camino por recorrer en este sentido, bien se sabe de las importantes variaciones introducidas por los hablantes de diferentes regiones y más aún si se quisiera incorporar a los hablantes no nativos. Siguiendo con el ejemplo de un reconocedor que comete errores en el 5 % de las palabras, si este se utilizara con un hablante no nativo el error puede ascender hasta un 80 %.

Es muy importante destacar que la capacidad de reconocimiento del ser humano en las mismas condiciones apenas llega a deteriorarse entre un 10 y un 20 %. Este gran deterioro en el rendimiento de un sistema de RAH ocurre también cuando ha sido entrenado con habla limpia y se prueba con habla a la que se ha sumado cantidades controladas de diversos tipos de ruido. En estos casos disminución de sus capacidades puede llevarlos a cometer hasta un 80 % de errores. De forma similar, cuando el sistema se entrena con voz registrada mediante un micrófono y un sistema de audio de alta calidad y se prueba con un micrófono comercial estándar, los errores pueden ascender al 50 %. Este es terreno de las investigaciones en reconocimiento "robusto" del habla [Junqua y Haton, 1996]. El objetivo principal es obte-

ner reconocedores que puedan ser utilizados en ambientes reales, con ruido en el ambiente, reverberación, pérdidas en el canal de transmisión, equipos de calidad comercial, etc. En los últimos años, cuando se ganó confianza en el RAH y se formaron grandes empresas con el fin de comercializar estas tecnologías, se ha puesto en evidencia la forma en que se deteriora el rendimiento de estos sistemas cuando no se encuentran en las mismas condiciones en que fueron entrenados. En un sentido más amplio se habla del "desapareamiento" entre las condiciones de entrenamiento y prueba del sistema. De forma similar a lo discutido en torno a la variabilidad entre locutores, existen dos enfoques principales que guían las investigaciones: las técnicas de transformación en el espacio de las características de la voz y las técnicas de adaptación del modelo a las condiciones de ruido. En muchos casos el ruido posee características muy especiales que permiten modelarlo fácilmente y mejorar significativamente el rendimiento de los sistemas de RAH. Sin embargo, es frecuente que el ruido consista en voces de diferentes personas que intervienen en la conversación y en esos casos vuelve a abrirse una nueva dimensión de complejidad en las investigaciones de RAH.

Debido a las diferentes restricciones impuestas sobre los sistemas de RAH actuales, no es posible obtener un único grupo de características que definan los límites claramente. Dada la aplicación práctica en que se utilizaría y las limitaciones conocidas, el sistema se debe diseñar a partir de una combinación que contemple las restricciones en el poder de cómputo de los ordenadores actuales. Si por ejemplo se conociera de antemano la identidad de cada locutor, entonces se podría utilizar modelos exclusivos para cada uno de ellos e incrementar sustancialmente otras variables, como el tamaño del vocabulario o la naturalidad con que se podría interactuar. Es así como surgen sistemas que se especializan en diferentes tareas como el dictado abierto, el reconocimiento de números o el comando de un teléfono celular.

Tres dicotomías en los sistemas de RAH

Todas las variantes y limitaciones expuestas en la sección anterior hacen que en la actualidad sea inviable abordar el problema del RAH de forma global, completa. Si bien en la década del 90 la capacidad de cómputo y almacenamiento de los ordenadores se han incrementado exponencialmente, se hace necesario establecer diversas hipótesis simplificadoras que permitan dedicar los recursos computacionales en uno u otro sentido. Una de las grandes dicotomías que se plantean en el diseño de estos sistemas es la de *estructura vs. generalidad*. Es interesante observar que a lo largo del tiempo los sistemas de RAH ha ido incorporando muchos de los niveles de la orga-

nización estructural del habla que se describió en la Sección 1.3 (página 26). Inicialmente se intentaba reconocer palabras o fonemas aislados pero luego estas unidades se incorporaron en una única estructura de reconocedor. De forma similar se incorporaron los ML, intentado capturar las relaciones entre palabras. Pero se podría pensar que un reconocedor que no cometa errores en el reconocimiento de fonemas no tiene por qué incorporar estructuras de otros niveles superiores. Este utópico reconocedor ganaría de esta forma una gran generalidad ya que no estaría restringido a un conjunto finito de palabras al punto tal de que reconocería incluso palabras inventadas por el locutor o pertenecientes a diferentes idiomas. Pero, ¿la información para tal reconocimiento existe en un único nivel? ¿acaso no dejamos de pronunciar las eses finales al hablar más rápido? Esto podría probar que tal reconocedor utópico nunca existirá y es también una buena argumentación que refuerza la evolución histórica según la cual se incorporaron progresivamente nuevos niveles estructurales en los sistemas de RAH. Queda claro que no se puede resolver el problema de RAH desde un único nivel de análisis, pero entonces ¿cuál es el costo de la incorporación de nuevos niveles en su estructura? ¿la pérdida de generalidad es obligatoria? No, en la medida en que se posean las capacidades de cómputo y almacenamiento necesarias será posible incrementar en forma conjunta la estructura y la generalidad. Esta argumentación también se sigue con las tendencias de los recientes avances en RAH.

Aún una capacidad de cómputo infinita no sería suficiente si no se posee el conocimiento suficiente acerca de los niveles estructurales del habla. En la incorporación estos nuevos niveles a los sistemas de RAH surge otra de las grandes dicotomías: los principios de estructuración *top-down* vs. los *bottom-up*. ¿Se debe esperar que el locutor emita una de las palabras de un conjunto definido por un par de fonemas iniciales o definido por la categoría gramatical de la palabra anterior? es decir ¿se debe abordar un nivel desde el adyacente más elemental o desde el más abstracto? Nuevamente la respuesta es *ambos*, no es necesario tomar partido, las dicotomías no deben entenderse como opciones excluyentes sino como soluciones confluyentes. Un claro ejemplo lo provee el fracaso que en la última década han tenido las redes neuronales artificiales como paradigma para el RAH (trabajos iniciados con mucho en los años 80 por [Kohonen et al., 1984, Waibel et al., 1989]). La principal causa de este fracaso se debe a la imposibilidad de modelar, bajo una concepción unificada, los diferentes niveles de estructuración en el habla. También surge a partir de este ejemplo la tercera dicotomía en el RAH: poder *discriminativo* vs. poder para la *expresión de la dinámica*. El problema de la discriminación consiste en encontrar un sistema capaz de dis-

tinguir entre unos y otros segmentos característicos del habla. En cuanto a la dinámica, se trata de modelar o capturar las variaciones de las características del habla en el tiempo. Las redes neuronales artificiales nunca pudieron modelar adecuadamente las dinámicas de diversos niveles simultáneamente, mientras que en los MOM la integración se dio de una forma más natural, como se discutió antes, a través de los autómatas probabilísticos[27]. Nuevamente, no se pueden separar por completo los dos aspectos porque cuanto más diferenciadas sean esas características, más fácil será seguir su evolución en el tiempo[28]. Además, volviendo a la perspectiva estructural, es necesario que tanto la dinámica como la discriminación sean consideradas en diferentes *escalas de observación* ya que esta interacción entre deferentes niveles resulta esencial para la resolución de ambigüedades.

Resolución de ambigüedades, falta de información y ruido

¿Cómo es posible distinguir la palabra *hola* de la palabra *ola*? No es posible en este caso recurrir a la información presente en niveles más bajos a la palabra, será necesario ver las palabras que la acompañan, la estructura sintáctica, o incluso saber si se está dando un saludo o hablando del mar. En estos casos se realiza una resolución de ambigüedades, sin haber aquí indicios de información confusa o ruidosa a nivel acústico. Un ejemplo de información confusa a nivel acústico lo ofrecen las palabras: *pala* y *bala*, que cuando se pronuncian en forma aislada o en el comienzo de una frase suenan muy similares. Lo mismo sucede cuando nos dictan por teléfono la letra de un departamento: *Vivo en el primero D*, es muy difícil saber si se pronunció [be], [de], [e], [θe], etc. Pero además, en este caso no sólo existe una cercanía acústica sino que el contexto y un análisis de los niveles superiores no ayuda en nada. Aunque generalmente con algo más de información acústica distintiva, esto también sucede en el dictado de números, por ejemplo, un número telefónico. Es probable que para los primeros dos o tres números se pueda recurrir a cierta información pragmática pero los siguientes sólo pueden distinguirse a nivel acústico.

Un hablante nativo utiliza subconscientemente la mayor cantidad posible de los niveles de la Figura 1.18 (página 26). Aún más, cuando los niveles superiores no existen suelen llegar a imponerlos artificiosamente: *Vivo en el primero D, D de dedo*. Cuando falta información y la acústica es confusa,

[27] Así y todo, algunos defensores de las redes neuronales artificiales argumentan que el mejor reconocedor del habla sigue siendo una red neuronal (no artificial) y de esta forma se vuelve a la discusión acerca de pájaros y aviones.

[28] De forma similar, cuando mejor segmentadas estén más fácil es clasificarlas.

hay ruido o incluso errores en la conformación del mensaje es cuando más se hace uso de la interacción entre los diferentes niveles. Estos fenómenos pueden estar auspiciados por problemas en el canal de transmisión, faltas de dicción al hablar con apuro o tensión emocional, acentos y modas regionales, o incluso el hecho de que en todos los idiomas existen palabras que adquieren su significado sólo en base al contexto. Por ejemplo, la palabra *nicho* tendrá un significado diferente si se pronuncia en una clase de biología o una visita al cementerio. Llegando a estos niveles se comienza a abandonar lo que tradicionalmente se conoció como reconocimiento del habla para llegar a la comprensión o entendimiento del habla y los sistemas de diálogo.

1.5.2. Incorporación del nivel suprasegmental

Aún falta recorrer mucho camino para que los sistemas de RAH alcancen las capacidades de reconocimiento del ser humano [Lippmann, 1997]. Es interesante contrastar las Figuras 1.18 (página 26) y 1.26 (página 56) y observar que muchos de los niveles estructurales del habla —ampliamente estudiados en el dominio de la lingüística— aún no se han incorporado al RAH. Sin llegar a los niveles que han sido incumbencia de los sistemas de diálogo y comprensión del habla, la argumentación de los apartados anteriores sugiere la tentadora opción de incorporar un nuevo nivel en los sistemas de RAH: el nivel suprasegmental.

Los sistemas de *texto a voz* constituyen un buen ejemplo de una de las tecnologías de la voz que se ha beneficiado enormemente con la incorporación de rasgos prosódicos [Rossi, 1997]. Los estudios y modelos propuestos en este área nos proveen de una amplia fuente de conocimientos acerca de como la prosodia se manifiesta en el lenguaje natural. En el caso de los sistemas de texto a voz se utilizan los rasgos prosódicos fundamentalmente para dar una mayor naturalidad a la voz sintética [Van Santen, 1997].

En este ámbito se encuentran muy diversos modelos para distintos idiomas, que tratan básicamente de generar los rasgos prosódicos a partir del texto escrito (por ejemplo en [Cahn, 1998] para el inglés, para el chino mandarín [Chen et al., 1998], para el alemán [Olaszy y Németh, 1997] y para el francés [Véronis et al., 1998]). Básicamente, en estos sistemas se intenta generar los rasgos prosódicos a partir del texto escrito. Pero, cuando se quiere utilizar la prosodia como una ayuda al RAH, el problema se plantea a la inversa. Ahora se trata de descubrir estructuras prosódicas en una emisión natural de voz, caracterizarlas e incorporarlas al proceso de reconocimiento. En el RAH no se posee el texto ya que el objetivo es justamente encontrarlo a partir de la emisión de voz. En este caso, se pretende extraer los rasgos

prosódicos de la emisión de voz de forma de obtener alguna información que ayude a determinar el texto. Aquí se llegan a distinguir las primeras grandes facetas del análisis del problema: la obtención de los rasgos prosódicos y su incorporación a un sistema de RAH. En un punto central, se encuentra a la acentuación como nexo estructural entre el texto y la manifestación física de la prosodia. De esta forma se puede dividir el problema en tres partes:

1. Estudiar la forma en que se manifiesta la acentuación en los rasgos prosódicos (Capítulo 3).

2. Encontrar un método que obtenga de forma automática la acentuación a partir de los rasgos prosódicos (Capítulo 4).

3. Incorporar la información prosódica y acentual a un sistema de RAH (Capítulo 5).

Estas tres etapas también podrían seguirse para incorporar la prosodia en cualquier otro idioma también. En cada caso habrá que descubrir primero la forma natural en que se manifiesta la acentuación y su relación con los rasgos prosódicos en el lenguaje natural. Es decir, volviendo a la idea de dicotomía estructura vs. generalidad, es necesario encontrar la *estructura natural* de lenguaje. En general esta etapa la cubren lingüistas, fonólogos y otros estudiosos dedicados al lenguaje hablado. En nuestro caso, fue necesario profundizar algunos estudios previos para descubrir fenómenos relacionados más directamente con el fin de la Tesis.

Dado que la estructura natural del lenguaje puede ser altamente compleja, puede resultar difícil encontrar un conjunto simple de reglas que relacionen las variables de interés. Aquí comienza a adquirir fuerza la tercera de las dicotomías descritas en la sección anterior. El problema de relacionar rasgos prosódicos con acentuación posee una dimensión relacionada con la capacidad discriminativa y otra con la dinámica temporal. Capturar ambas mediante una técnica que en el compromiso que aproveche tanto la información de una como de la otra es el objetivo de esta parte.

Finalmente, con alguna estimación de la acentuación que caracteriza a una emisión de voz, se requiere incorporar esta "estructura" al resto de las estructuras de un modelo estándar en RAH. Pero esta incorporación debe realizarse de forma que el compromiso estructura-generalidad de un balance positivo para los corpus de habla utilizados en las pruebas. Aquí juega un rol fundamental el principio de estructuración que se seleccione. Si se elige la dirección *bottom-up* entonces se puede pensar en incorporar a los fonemas la información acentual, por ejemplo, distinguiendo entre sonidos vocálicos

tónicos o átonos. Por el contrario, si se elige la dirección *top-down* es posible pensar en palabras inacentuadas y acentuadas, y en éstas últimas, considerar estructuras acentuales de acuerdo a la tonicidad de cada sílaba.

En esta línea de pensamiento el objetivo es introducir información a partir de niveles que antes se simplificaban. Cuando en el Capítulo 2 se describa el procesamiento de la voz para RAH se podrá apreciar que la información de F_0 —una variable tan característica de las emisiones de voz— generalmente desaparece[29]. No sucede lo mismo con la energía, que se incorpora explícitamente, pero no siempre la duración de los segmentos es modelada de la forma más apropiada. La utilización de estos rasgos prosódicos y su relación con la acentuación como vínculo estructural con el modelo de RAH, resulta en un mejor aprovechamiento de la información contenida en la señal de voz y el conocimiento a priori de la estructura del lenguaje hablado, para atacar los problema de ambigüedad, falta de información y ruido.

Antecedentes

Ya se han referido algunos antecedentes en el estudio de la prosodia y la acentuación en el español (principalmente [Quilis, 1993, Almiñana, 1991, Sosa, 1999, Llorach, 1999]). Existen experimentos muy interesantes que relacionan las habilidades de oyentes humanos para el reconocimiento en diversas condiciones de procesamiento prosódico de las frases [Hoskins, 1997] (véase para el caso de niños [Bosch y Gallés, 1997], en habla espontánea [Laan, 1997] y [Lublinskaja y Sappok, 1996] en la distinción entre diálogo y monólogo). Un caso típico al que se puede tener acceso a diario es el de la dificultad en el reconocimiento del lenguaje afectado por las diferentes modificaciones del acento regional [Arslan y Hansen, 1996]. Esto último ha sido considerado en el contexto del RAH en [Humphries y Woodland, 1998]. Otro caso en donde se pone de manifiesto la información prosódica y su utilización en el lenguaje hablado es en la identificación del hablante, por ejemplo véase [Sönmez et al., 1997]. Es importante reconocer también que las modificaciones de la prosodia evidentemente generan importantes modificaciones de otras variables que son modelados explícitamente en los sistemas de RAH actuales. Como ya se pudo apreciar en ejemplos anteriores (Sección 1.2.4), las características espectrales de la voz se ven modifi-

[29]Como se verá en el Capítulo 2, esto ocurre tanto en la integración por bandas, para el análisis espectral o el mel cepstrum, como en el caso de los coeficientes de predicción lineal, dado el orden del modelo que se utiliza generalmente.

cadas globalmente cuando se cambia la entonación[30]. También se pueden observar cambios notorios en la duración de los fonemas, principalmente vocales, en función de determinadas características semánticas, sintácticas y hasta ortográficas que son transmitidas en el mensaje del habla espontánea [Caspers, 1997, Batliner et al., 1997]. Se han obtenido importantes mejoras en el RAH considerando simplemente la velocidad de elocución [Busdhtein, 1996]. Estas y muchas otras modificaciones se realizan tanto a nivel de la frase como a nivel de las palabras o incluso sílabas y fonemas. Por ejemplo, un modelo de entonación basado en varios niveles jerárquicos sumados se utiliza en [Ross y Ostendorf, 1999]. Sin embargo, se tiende a relacionar rasgos prosódicos con información no relevante a nivel fonético y más bien asociada sólo con la frase, su sintaxis [Strangert, 1997] o su semántica [Lieske et al., 1997]. Por último, se debe mencionar la información relativa a la separación entre palabras o frases que ofrecen conjuntamente las curvas de energía y entonación [Rajendran y Yegnanarayana, 1996] y el hecho de que tampoco esto se modela de forma explícita en los reconocedores actuales.

Se han utilizado MOM [Brindöpke et al., 1998, Brindöpke et al., 1999] para modelar la entonación en el alemán. En la Tesis [Ying, 1998] se ha realizado un interesante estudio de los rasgos prosódicos para el inglés y se aplicaron diversos métodos de clasificación para relacionar la acentuación con la energía, la duración y la F_0 de las sílabas. En este trabajo se pretende obtener[31] una relación entre rasgos prosódicos y acentuación con el fin de que luego se incorpore esta información a un sistema de RAH basado en MOM. Los resultados son prometedores pero la integración al RAH no forma parte del citado trabajo. Otros autores ya habían destacado la dificultad para encontrar relaciones entre la entonación y la acentuación en habla continua en inglés [Yaeger-Dror, 1996]. En el caso del holandés se ha logrado una clasificación automática de sílabas acentuadas y no acentuadas con 72.6 % de precisión para el mejor de los casos [Kuijk y Boves, 1999]. Un estudio que se aproxima a nuestra finalidad para el español se realizó en [Almiñana, 1991]. Pero como bien destaca el mismo autor, su procedimiento de estilización no toma a la sílaba como unidad de análisis y no sería aplicable a sistemas de RAH que tienen por finalidad obtener información a este nivel. Estas curvas estilizadas serían más aplicables a sistemas que pretendan aprovechar la información sintáctica o semántica contenida en la curva de entonación [Swerts y Ostendorf, 1997].

[30] Aunque, como ya se ha notado, esta información es muchas veces eliminada en el procesamiento de la voz para RAH.

[31] Al igual que en el Capítulo 4 de la presente Tesis.

Existen también algunos antecedentes en la incorporación de la prosodia al RAH. Ciertos autores aluden a los beneficios potenciales de la incorporación de rasgos prosódicos al RAH pero no proponen ninguna solución concreta [Pols et al., 1996]. Por otra parte, existen trabajos que han incorporado algunos de las rasgos prosódicos para solucionar sólo un grupo reducido de problemas relacionados con el RAH. Entre estos se encuentra por ejemplo [López et al., 1998] donde se utiliza con éxito la entonación para recuperar algunos errores particulares de reconocimiento en dígitos conectados. En este caso tanto como en [Chung y Seneff, 1998], es característico el hecho de que la prosodia se incorpore en base para un análisis posterior al reconocimiento en sí y no como parte del reconocedor mismo. Por ejemplo en [Bartkova y Jouvet, 1999, Wang y Seneff, 1998, Wu et al., 1998, Molloy y Isard, 1998] se utiliza como punto de partida las N frases más probables y una posterior recategorización basada en la prosodia. A la inversa, en [Vereecken et al., 1997] se realiza una segmentación previa basada en rasgos prosódicos y luego se reconoce por partes. Los trabajos de [Lee y Hirose, 1999] y [Buckow et al., 1998] (más recientemente publicados en [Warnke et al., 1999] y [Nöth et al., 2000]) constituyen una excepción ya que se utiliza la prosodia para incorporar en el mismo reconocedor las hipótesis de fin de frase. En el mismo sentido, otros autores han utilizado la prosodia en relación con eventos de disfluencia y detección de pausas [Stolcke et al., 1999, Rajendran y Yegnanarayana, 1996, Hirose y Iwano, 2000], pero en la presente Tesis se propone utilizarla para detectar particularidades dentro del ámbito de la palabra. Muchos trabajos se han orientado a estudiar la información contenida en la entonación y su utilización en reconocedores para el caso de varios lenguajes *tonales*[32] [Chih-Heng et al., 1996, Chiang et al., 1996, Lee y Ching, 1999], donde existe una relación muy directa entre el significado de la palabra y la cadencia tonal utilizada [Hirose y Iwano, 1998, Potisuk et al., 1999]. En este punto, es importante recalcar que la utilización de los rasgos prosódicos, y la información que se codifica en ellos, varía mucho de un lenguaje a otro y las extrapolaciones son frecuentemente inválidas. Se han realizado análisis cruzados entre varios lenguajes no tonales por ejemplo en [Pallier et al., 1997, Campione y Véronis, 1998]. Sin embargo, en esta Tesis se trabaja sobre un corpus de habla en español, por lo que son de gran utilidad varios análisis previos de las lenguas de España [Bonafonte et al., 1997, López et al., 1997, Aguilar et al., 1997, Iparraguirre y Torres, 1996].

[32]Como lo son la mayoría de los orientales.

1.5.3. Objetivos de la Tesis

Para finalizar este capítulo se presentan en forma resumida los objetivos de la Tesis. En la sección anterior se ha descrito el problema y la metodología a seguir para su resolución. Los objetivos se desprenden directamente de aquel planteamiento y tienden la cubrir cada etapa de la investigación.

Objetivo general:

> *Investigar diferentes vías que permitan utilizar la información de los rasgos prosódicos y la acentuación para mejorar el rendimiento de un sistema de reconocimiento automático del habla continua en español, basado en modelos ocultos de Markov.*

Objetivos particulares:

> *Realizar un análisis de los tres rasgos prosódicos más importantes: energía, entonación y duración, con el fin de encontrar sus relaciones con la acentuación.*

> *Investigar la segmentación y clasificación automática de estructuras acentuales a partir de las evidencias acústicas en el habla continua.*

> *Investigar la forma de incorporar la información prosódica y acentual en un sistema de reconocimiento automático del habla basado en modelos ocultos de Markov.*

Capítulo 2

Reconocimiento automático del habla

En este capítulo se hará una descripción detallada de las principales técnicas en que se basó la presente Tesis. El objetivo principal es establecer los fundamentos para el desarrollo de los capítulos posteriores. El capítulo se divide en dos grandes bloques: el análisis de la señal de voz y los modelos ocultos de Markov. Ambos bloques están especialmente orientados y restringidos al reconocimiento automático del habla, con especial énfasis en las técnicas que se utilizaron en esta Tesis. En primer lugar se tratará, como marco general, el análisis por tramos de la señal de voz. A partir de esta particular forma de seguir la dinámica de la voz, se describen los diferentes métodos de análisis. En la segunda parte del capítulo se describe la estructura y entrenamiento de un sistema de reconocimiento automático del habla basado en modelos ocultos de Markov. Inicialmente se trata en forma genérica la versión continua de estos modelos y luego se realiza una ampliación para incluir a los modelos semicontinuos. Para completar esta descripción se incluyen los modelos de palabra y los modelos de lenguaje, construyendo así un modelo compuesto. Las ecuaciones para el entrenamiento y la decodificación se extienden al modelo compuesto utilizado en el reconocimiento automático del habla continua.

2.1. Análisis de la señal de voz

La señal de voz posee una gran variabilidad en el tiempo y, como se anticipó en el Capítulo 1, es necesario descomponerla en intervalos de tiempo que permitan su estudio bajo la hipótesis de estacionariedad. Estos intervalos estarán en relación directa con la máxima velocidad con que el tracto vocal pueda modificar significativamente su morfología. En las aplicaciones prácticas para el reconocimiento automático del habla (RAH) se utilizan intervalos de 10 a 30 ms. A continuación se desarrollan estas ideas y a partir de allí se definen técnicas útiles para el análisis de la señal de voz en el contexto del RAH.

2.1.1. Análisis por tramos

Sea $v(\tau)$ la señal continua de voz para la variable real de tiempo τ. Después de un proceso de muestreo uniforme con período T_v, la señal de voz en la variable natural de tiempo discreto $0 < m \leq N_v$ se representa como $v(mT_v)$ o más simplemente $v(m)$.

Sea la señal $\omega(m; N_\omega)$ una ventana de análisis definida para $0 < m \leq N_\omega$, se dice que esta ventana posee un *ancho* $T_\omega = N_\omega T_v$. De la aplicación de la ventana de análisis temporal se obtienen los *tramos* de voz:

$$v(t; n) = \omega(n; N_\omega) v(tN_d + n); \quad 0 < n \leq N_\omega \tag{2.1}$$

que representaremos en notación vectorial como $\mathbf{v}_t$. Se denomina *paso* del análisis por tramos al tiempo $T_d = N_d T_v$. Dadas las definiciones anteriores la variable de tiempo por tramos $t \in \mathbb{N}$ queda acotada según $0 < t \leq T = (N_v - N_\omega)/N_d + 1 < \infty$.

Si $\mathcal{T}(k)$ es un operador para la transformación de dominio, se realiza el proceso de parametrización de la señal de voz según:

$$x(t; k) = \mathcal{T}(k) \{v(t; n)\}, \quad 0 < k \leq N_x$$

para la que se utilizará la notación vectorial simplificada como $\mathbf{x}_t \in \mathbb{X} = \mathbb{R}^{N_x}$. Se conoce a $\mathbb{X}$ como el *espacio de las características* con dimensión N_x. En esta sección se utilizará $0 < k \leq N_x$ como variable independiente discreta en el dominio transformado.

Ventanas de análisis

Las ventanas de análisis más utilizadas se definen para $0 < m \leq N_\omega$ según:

I) Ventana rectangular:

$$\omega_R(m; N_\omega) = 1$$

II) Ventana de Hanning:

$$\omega_h(m; N_\omega) = \frac{1}{2} - \frac{1}{2}\cos(2\pi m/N_\omega)$$

III) Ventana de Hamming:

$$\omega_H(m; N_\omega) = \frac{27}{50} - \frac{23}{50}\cos(2\pi m/N_\omega)$$

IV) Ventana de Bartlett:

$$\omega_B(m; N_\omega) = \begin{cases} 2m/N_\omega & \text{si } 0 < m \leq N_\omega/2 \\ 2 - 2m/N_\omega & \text{si } N_\omega/2 < m \leq N_\omega \end{cases}$$

V) Ventana de Blackman:

$$\omega_K(m; N_\omega) = \frac{21}{50} - \frac{1}{2}\cos(2\pi m/N_\omega) + \frac{2}{25}\cos(4\pi m/N_\omega)$$

Estas ventanas pueden ser caracterizadas por el tamaño de los lóbulos de la magnitud de su espectro de frecuencias. La ventana rectangular posee el lóbulo central con menor ancho de banda pero la magnitud de los lóbulos laterales decae muy lentamente. La ventana de Blackman posee la mínima amplitud en sus lóbulos laterales pero su lóbulo principal tiene un ancho de banda tres veces mayor al de la rectangular [Kuc, 1988]. Dado este compromiso entre resolución frecuencial y distorsión armónica en el proceso de ventaneo, para señales de voz suele utilizarse la ventana de Hamming que además, ofrece una posición media en cuanto al costo computacional de su aplicación [Deller et al., 1993].

Transformaciones

El operador $\mathcal{T}(k)$ permite obtener un vector de características $\mathbf{x}_t$ para el análisis por tramos de la señal de voz. A continuación se tratarán los operadores más comúnmente utilizados en el RAH:

I) Coeficientes espectrales (CE):

$$\mathbf{x}_t = [u(t;k)] = \mathcal{T}_F(k)\left\{v(t;n)\right\},$$

II) Coeficientes de predicción lineal (CPL):

$$\mathbf{x}_t = [a(t;k)] = \mathcal{T}_L(k)\left\{v(t;n)\right\},$$

III) Coeficientes cepstrales (CC):

$$\mathbf{x}_t = [c(t;k)] = \mathcal{T}_C(k)\left\{v(t;n)\right\},$$

En las diferentes alternativas para los vectores de características se definirán N_{uI}, N_a y N_{cI} que, en el caso general, corresponderán a N_x.

2.1.2. Coeficientes espectrales

Se define la transformada discreta de Fourier (TDF) de $v(m)$ como:

$$u(k) = \sum_{m=1}^{N_v} v(m) e^{-\mathrm{j}(2\pi/N_v)k(m-1)} \tag{2.2}$$

Si se aplica la TDF a los tramos de voz $v(t;n)$ de la ecuación (2.1), es posible obtener la denominada transformada de Fourier de tiempo corto o por tramos:

$$\begin{aligned} u(t;k) &= \sum_{n=1}^{N_v} v(t;n) e^{-\mathrm{j}(2\pi/N_v)(k-1)(n-1)} \\ &= \sum_{n=1}^{N_v} \omega(n;N_v) v(tN_d+n) e^{-\mathrm{j}(2\pi/N_v)(k-1)(n-1)} \end{aligned}$$

Generalmente, dado que $v(t;n) \in \mathbb{R}$, se utiliza el espectro de magnitud $|u(t;k)|$ en $0 < k \leq N_v/2$ y la notación vectorial $\mathbf{u}_t \in \mathbb{R}^{N_u}$ con $N_u = N_v/2$.

Integración por bandas

Para el RAH suele utilizarse el logaritmo de la energía de un número reducido de bandas del espectro, en lugar del espectro completo. Es necesario definir las frecuencias de corte para cada banda y para cada ley de mapeo frecuencial o "escala" de integración se podrá obtener un conjunto diferente de coeficientes. Un ejemplo sencillo es la escala de integración lineal, donde la relación entre ambas frecuencias tiene la forma:

$$F_{lin} \propto f_{Hz}$$

Si se consideran N_{uI} bandas de integración en la primera mitad del espectro, es posible calcular los extremos de cada intervalo mediante:

$$B(k) = \frac{kN_u}{2N_{uI}}; \quad 0 \leq k \leq N_{uI}$$

En el caso más simple se realiza la integración mediante ventanas frecuenciales rectangulares:

$$u_I(t;k) = 2 \sum_{\varkappa=B(k-1)}^{\varkappa=B(k)} \log|u(t;\varkappa)|; \quad 0 < k \leq N_{uI}$$

Cuando se utilizan ventanas de Bartlett o de Hamming el esquema de integración se modifica para no perder la energía en los extremos de cada ventana:

$$u_I(t;k) = 2 \sum_{\varkappa=B(k-1)}^{\varkappa=B(k+1)} \omega_B\left(\varkappa - B(k-1); B(k+1) - B(k-1)\right) \log|u(t;\varkappa)| \tag{2.3}$$

con $0 < k < N_{uI}$.

Diversos estudios acerca de la percepción de tonos puros en el ser humano (ver Sección 1.2.2) han permitido aproximar la relación entre la frecuencia percibida y la real mediante:

$$F_{mel}\left(f_{Hz}\right) = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right),$$

relación que es ampliamente utilizada como escala de integración en el procesamiento de señales de voz.

2.1.3. Coeficientes de predicción lineal

Es posible modelar el tracto vocal mediante un sistema auto-regresivo de la forma:

$$\hat{v}(t;n) = -\sum_{j=1}^{N_a} a(t;j) v(t;n-j) + Gg(t;n) \tag{2.4}$$

donde $v(t;n)$ es la señal a modelar, $\hat{v}(t;n)$ es la señal estimada por el modelo, $g(t;n)$ es la entrada al tracto vocal y N_a es el orden del sistema.

Para este análisis se considera inicialmente una entrada nula y la ecuación anterior puede escribirse usando notación vectorial simplificada como:

$$\hat{v}(t;n) = -(\mathbf{v}_t^n)^{\mathrm{T}} \mathbf{a}_t$$

donde $\mathbf{a}_t$ contiene los N_a coeficientes $a(t;j)$ y $\mathbf{v}_t^n$ contiene las últimas N_a salidas $v(t;n-j)$. El error entre $v(t;n)$ y $\hat{v}(t;n)$ se puede medir mediante la distancia euclídea como:

$$e(t;n)^2 = (v(t;n) - \hat{v}(t;n))^2.$$

Para encontrar el vector $\mathbf{a}_t$ se minimiza la medida del error cuadrático total entre $\hat{v}(t;n)$ y $v(t;n)$:

$$\xi^2 = \sum_n e(t;n)^2 = \sum_n \left(v(t;n) + (\mathbf{v}_t^n)^{\mathrm{T}}\mathbf{a}_t\right)^2$$

a partir de:

$$\nabla\xi^2 = 0$$

se obtiene:

$$\left(\sum_n \mathbf{v}_t^n(\mathbf{v}_t^n)^{\mathrm{T}}\right)\mathbf{a}_t = -\sum_n \mathbf{v}_t^n v(t;n)$$

conocido como sistema de Wiener-Hopf y comúnmente representado como:

$$\mathbf{R}_t\mathbf{a}_t = -\mathbf{r}_t \tag{2.5}$$

donde $\mathbf{r}_t$ es el vector de autocorrelación y $\mathbf{R}_t$ la matriz de autocorrelación para $v(t;n)$. Se puede verificar que $R_{ij} = r_{i-j}$ y así $\mathbf{R}_t$ es una matriz Toeplitz. El método de Levinson-Durbin [Kay y Marple, 1981] aprovecha esta propiedad para simplificar la resolución del sistema.

Resta por definir el orden N_a del sistema. Existen varios métodos para encontrar el orden del sistema de forma que se obtenga un buen compromiso entre el error total y la complejidad de su estructura. Estos métodos se basan en medidas del error en la predicción, por ejemplo, a partir de las ecuaciones (2.4) y (2.5) es posible obtener ([Makhoul, 1975]):

$$E(N_a) = r_0 + \mathbf{r}_t^{\mathrm{T}}\mathbf{a}_t$$

y encontrando el modelo más simple cuyo $E(N_a)$ sea mínimo se puede determinar el orden apropiado para la estimación. Otros métodos más elaborados

[Akaike, 1974] utilizan criterios basados en la teoría de la información. Asumiendo una distribución gaussiana en la señal se puede medir el error según:

$$I(N_a) = \log E(N_a) + \frac{2N_a}{N_e}$$

donde N_e es el número efectivo de muestras en la señal, que para el caso de una ventana de Hamming $N_e = 0{,}4N_\omega$. En general, para el modelado de señales de voz en RAH se encuentra un buen compromiso para un orden N_a entre 10 y 16 [Young et al., 2000].

2.1.4. Coeficientes cepstrales

En base a la TDF, se define el cepstrum real de $v(m)$ como:

$$c(m) = \mathcal{T}_F^{-1}\left\{\log\left|\mathcal{T}_F\left\{v(m)\right\}\right|\right\},$$

Esta definición se puede extender para un análisis por tramos. Reemplazando según la TDF (2.2) y su inversa (TDFI), se obtiene:

$$\begin{aligned} c(t;k) &= \mathcal{T}_F^{-1}\left\{\log\left|\sum_{n=1}^{N_v} v(t;n)e^{-j(2\pi/N_v)(\varkappa-1)(n-1)}\right|\right\} \\ &= \frac{1}{N_u}\sum_{\varkappa=1}^{N_u}\log\left|u(t;\varkappa)\right| e^{j(2\pi/N_u)(\varkappa-1)(k-1)} \end{aligned} \tag{2.6}$$

Finalmente, si se considera que el argumento de la TDFI es una secuencia real y par, puede simplificarse su cómputo mediante una transformada coseno (TC):

$$\begin{aligned} c(t;k) &= \frac{1}{N_u}\sum_{\varkappa=1}^{N_u}\log\left|u(t;\varkappa)\right|\cos\left((2\pi/N_u)(\varkappa-1)(k-1)\right) \\ &= \frac{2}{N_u}\sum_{\varkappa=2}^{N_u/2-1}\log\left|u(t;\varkappa)\right|\cos\left((2\pi/N_u)(\varkappa-1)(k-1)\right) \end{aligned} \tag{2.7}$$

La señal de voz y el cepstrum

Siguiendo la idea del modelo para el tracto vocal presentada en la ecuación (2.4), es posible considerar que la señal de voz para fonemas sonoros es generada mediante la convolución:

$$\hat{v}(t;n) = g(t;n) * h(t;n)$$

donde la entrada al sistema es el tren de pulsos glóticos $g(t;n)$ y $h(t;n)$ es la respuesta al impulso del tracto vocal. Cuando se pasa al dominio frecuencial mediante $\mathcal{T}_F$ y se aplica el logaritmo, resulta:

$$\hat{v}(t;\varkappa) = \log|g(t;\varkappa)| + \log|h(t;\varkappa)|$$

Cuando nuevamente se transforma esta señal mediante la TDFI se obtiene:

$$\hat{v}(t;k) = \mathcal{T}_F^{-1}\{\log|g(t;\varkappa)|\} + \mathcal{T}_F^{-1}\{\log|h(t;\varkappa)|\}$$

Generalmente, la señal del pulso glótico varía muy lentamente en relación a la otra componente, digamos, con período $1/F_0$. Cuando se realiza la primera transformación, claramente se puede observar que $g(t;\varkappa)$ es modulada por $h(\varkappa)$ a razón de F_0. Es así como la segunda transformación, luego de haber aplicado el logaritmo a la magnitud, deja en las primeras muestras la información relacionada con $h(t;k)$ y a partir de $1/F_0$ lo relativo al pulso glótico $g(t;k)$. Normalmente, en RAH se utiliza la primera parte del cepstrum y se descarta lo relativo al pulso gótico.

Entonación

Se han descrito muchos métodos para estimar la frecuencia fundamental (F_0) en señales de voz [Hess, 1991]. Además de aquellos basados en CC, existen métodos basados en la correlación cruzada, en CE y en CPL [Deller et al., 1993]. Siguiendo el razonamiento anterior, en relación a la forma en que el cepstrum real separa la información relativa al pulso glótico, se puede observar que la simple detección del pico correspondiente al pulso glótico en el cepstrum real constituye un método para determinar F_0 en los

fonemas sonoros. Los estudios en este sentido fueron iniciados por Michael Noll, quien reunió un conjunto de reglas sencillas para eliminar los principales artefactos generados al aplicar el método en voz continua [Noll, 1967]. Existen tres aspectos centrales a considerar:

- La ausencia del pulso glótico en fonemas sordos
- El fenómeno de duplicación de entonación en la estimación
- Los picos y ausencias aisladas de la F_0 estimada

En pos de resolver varios problemas prácticos del método, deben tomarse en cuenta las siguientes reglas:

1. Antes de la detección del pico correspondiente al pulso glótico conviene realizar una ponderación del cepstrum real, en forma tal que se reduzca la magnitud de las primeras componentes y se aumente la de las últimas:

$$c_p(t;k) = |c(t;k)|\,(k\upsilon + \zeta); \quad \upsilon, \zeta > 0$$

2. La búsqueda del máximo pico debe realizarse en el intervalo de 2 a 15 ms.

3. La amplitud del pico encontrado debe superar un umbral previamente fijado en forma empírica (de acuerdo a alguna estimación de la energía total en la señal de voz).

4. Conviene reducir el umbral requerido a la mitad por cada pico que se encuentre en dos tramos consecutivos que presenten una variación del período menor a 1 ms.

5. En caso de encontrar una ausencia de F_0 entre tramos que sí la tienen, debe considerarse que el tramo posee la F_0 promedio de las de su entorno.

6. Si la frecuencia de F_0 del tramo actual es superior a 1.6 veces la del tramo anterior, entonces conviene buscar un pico 0.5 ms alrededor de la mitad de período del detectado.

Este conjunto reducido de reglas aplicadas a los CC sigue siendo hasta la actualidad el mejor método conocido para la determinación de la entonación en habla limpia [Shimamura y Kobayashi, 2001]. Por otro lado, dado que los CC son los más utilizados en los sistemas actuales de RAH, resulta inicialmente atractivo utilizarlos para la estimación de la entonación, sin incrementar significativamente el costo computacional.

Coeficientes cepstrales en escala de mel

Para combinar las propiedades del cepstrum y los resultados acerca de la percepción de tonos puros en el ser humano, se propuso integrar la representación espectral de la señal según la escala de mel antes de aplicar la TC [Davis y Mermelstein, 1980]. Siguiendo estas ideas se pueden definir los coeficientes cepstrales en escala de mel (CCEM) a partir de las ecuaciones (2.3) y (2.6):

$$\begin{aligned} c_{mel}(t;k) &= \frac{2}{N_{uI}} \sum_{i=2}^{N_{uI}} u_I(t;i) \cos\left((2\pi/N_{uI})(i-1)(k-1)\right) \\ &= \frac{4}{N_{uI}} \sum_{i=2}^{N_{uI}} \sum_{\varkappa=B(i-1)}^{\varkappa=B(i+1)} \omega_B\left(\varkappa - B(i-1); B(i+1) - B(i-1)\right) \\ &\qquad \times \log \left| \sum_{n=1}^{N_v} v(t;n) e^{-j(2\pi/N_v)(\varkappa-1)(n-1)} \right| \\ &\qquad \times \cos\left((2\pi/N_{uI})(i-1)(k-1)\right) \end{aligned}$$

Los resultados experimentales han favorecido ampliamente esta combinación. Como detalle de aplicación práctica debe mencionarse que en general para RAH no se utilizan todos los $c_{mel}(t;k)$ sino que se desecha toda la información relacionada con el pulso glótico. De forma similar que para los CPL, suelen utilizarse los primeros $N_{cI} = 13$ CCEM, a partir de una integración según $N_{uI} = 24$ bandas.

Relación entre CC y CPL

Para completar lo relativo a CC, se describe a continuación una aproximación que resulta útil para su cálculo [Huang et al., 1990]. Denotando

por $\mathcal{Z}\{\cdot\}$ al operador de la transformada Z [Oppenheim y Schafer, 1989], es posible escribir:

$$\frac{\mathrm{d}\mathcal{Z}\left\{v(t;n)\right\}}{\mathrm{d}z^{-1}} = \mathcal{Z}\left\{v(t;n)\right\}\frac{\mathrm{d}\mathcal{Z}\left\{c(t;k)\right\}}{\mathrm{d}z^{-1}}$$

ya que $\log\left(\mathcal{Z}\left\{v(t;n)\right\}\right) = \mathcal{Z}\left\{c(t;k)\right\}$. Considerándose que una estimación del espectro de la señal de voz puede obtenerse a partir del modelo autoregresivo de la ecuación (2.4):

$$\mathcal{Z}\left\{v(t;n)\right\} \approx \frac{G}{\mathcal{Z}\left\{a(t;j)\right\}}$$

ahora se obtiene:

$$-\mathcal{Z}\left\{ja(t;j)\right\} \approx \mathcal{Z}\left\{kc(t;k)\right\}\mathcal{Z}\left\{a(t;j)\right\}.$$

Invirtiendo la transformada Z y teniendo en cuenta que el producto del término de la derecha quedará como una convolución en el dominio no transformado:

$$\hat{c}(t;k) = -a(t;k) - \frac{1}{k}\sum_{j=2}^{k}(k-j+1)c(t;k-j+1)a(t;j)$$

con $a(t;i) = 0$ para $i > p$ y $k \geq 2$, ya que de la ecuación (2.7) se puede ver que $c(t;1) \propto \sum_{\varkappa}\log|u(t;\varkappa)|$.

2.1.5. Coeficientes de energía, delta y aceleración

Cuando se confecciona el vector de características para RAH, es práctica corriente considerar algunas otras variables que llevan información importante del tramo de voz considerado. Una de estas variables consiste en una medida de la energía que se define simplemente como:

$$\epsilon(t) = \log\sum_{n=1}^{N_v} v(t;n)^2 \tag{2.8}$$

También suele agregarse una estimación de las derivadas temporales de todos los elementos calculados. Para un vector de características $x(t;k)$ dado, se obtienen los *coeficientes delta* mediante la regresión:

$$\Delta x(t;k) = \frac{\sum_{j=1}^{N_J} j\,(x(t+j;k) - x(t-j;k))}{2\sum_{j=1}^{N_J} j^2}$$

donde N_J es utilizado para suavizar la estimación a través de los tramos (generalmente $1 \leq N_J \leq 2$). Los coeficientes de aceleración $\Delta^2 x(t;k)$ se obtienen por aplicación directa de la ecuación anterior a los $\Delta x(t;k)$.

2.2. Modelos ocultos de Markov

Los modelos ocultos de Markov (MOM) fueron introducidos conceptualmente en el Capítulo 1 y en esta sección se tratarán formalmente. Para comenzar se definirán los MOM continuos y se deducirán las fórmulas para la estimación de sus parámetros. A continuación se harán las extensiones necesarias para cubrir la estructura y el entrenamiento de los MOM semicontinuos. Finalmente se tratarán los modelos de lenguaje y su incorporación en lo que denominamos modelos compuestos para el RAH.

2.2.1. Estructura del modelo

Un MOM continuo (MOMC) queda definido mediante una estructura algebraica:

$$\Theta = \langle \mathcal{Q}, \mathbb{O}, \mathbf{A}, \mathcal{B} \rangle$$

donde:

$\mathcal{Q}$ es el conjunto de estados posibles,

$\mathbb{O}$ es el espacio observable,

$\mathbf{A}$ es la matriz de probabilidades de transición de estados y

$\mathcal{B}$ es el conjunto de distribuciones de probabilidades de observación.

El conjunto de estados posibles se define como:

$$\mathcal{Q} = \{q \in [1 \dots |\mathcal{Q}|]\}\,; \quad |\mathcal{Q}| < \infty$$

donde $|\mathcal{Q}| \in \mathbb{N}$ es la cardinalidad del conjunto. Para el espacio observable se tiene:

$$\mathbb{O} = \left\{\mathbf{o} \in \mathbb{R}^{N_o}\right\}; \quad N_o = N_x$$

donde $N_o \in \mathbb{N}$ es su dimensión, que coincide con la dimensión del espacio de las características $\mathbb{X}$, que en el contexto del los MOM también se denominará espacio de las evidencias acústicas.

Sean $q_{t-1}, q_t \in \mathcal{Q}$ dos estados cualquiera de modelo Θ, donde $t \in [1 \dots T] \subset \mathbb{N}$ tal como se definió en la Sección 2.1.1, entonces se define la matriz de probabilidades de transición de estados como:

$$\mathbf{A} = [a_{ij} = \Pr(q_t = j \,|q_{t-1} = i)] \quad \forall i, j \in \mathcal{Q}$$

donde $a_{ij} \geq 0 \;\; \forall i, j$ y $\sum_{j=1}^{|\mathcal{Q}|} a_{ij} \stackrel{\circ}{=} 1 \;\; \forall i \in \mathcal{Q}$.

Siendo $\mathbf{x}_t \in \mathbb{X}$ una evidencia acústica para el modelo Θ, se define el conjunto de distribuciones de probabilidad de observación como:

$$\mathcal{B} = \{b_j(\mathbf{x}_t) = \Pr(\mathbf{x}_t \,|q_t = j)\} \quad \forall j \in \mathcal{Q}$$

en donde para cada estado j se modela la distribución de probabilidades mediante la mezcla:

$$b_j(\mathbf{x}_t) = \sum_{k=1}^{N_c} c_{jk} b_{jk}(\mathbf{x}_t) \quad \forall j \in \mathcal{Q};\; N_c < \infty \tag{2.9}$$

siendo en este caso:

I) $b_{jk}(\mathbf{x}_t)$: todas funciones gaussianas de densidad de probabilidad multidimensional con la forma

$$\mathcal{N}(\mathbf{x}_t, \boldsymbol{\mu}_{jk}, \mathbf{U}_{jk}) = \frac{1}{(2\pi)^{N_x} |\mathbf{U}_{jk}|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(\mathbf{x}_t - \boldsymbol{\mu}_{jk})^{\mathrm{T}} \mathbf{U}_{jk}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_{jk})\right]},$$

II) $c_{jk} \in \mathbb{R}^{+0}$: las constantes de peso relativo para cada distribución normal que satisfacen

$$\sum_{k=1}^{N_c} c_{jk} \stackrel{\circ}{=} 1 \;\; \forall j \in \mathcal{Q},$$

III) $\boldsymbol{\mu}_{jk} \in \mathbb{R}^{N_x}$: los vectores de medias,

IV) $\mathbf{U}_{jk} \in \mathbb{R}^{N_x \times N_x}$: las matrices de covarianza y

V) se cumple que

$$\int_{-\infty}^{+\infty} b_j(\mathbf{x}_t)\,\mathrm{d}\mathbf{x}_t \stackrel{\circ}{=} 1 \quad \forall j \in \mathcal{Q}.$$

2.2.2. La secuencia más probable

Dada la secuencia de evidencias acústicas:

$$\mathbf{X}^T = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T; \quad \mathbf{x}_t \in \mathbb{X}$$

y

$$\mathbf{q}^T = q_1, q_2, \ldots, q_T; \quad q_t \in \mathcal{Q}$$

una secuencia cualquiera de exactamente T estados, se calcula la probabilidad de que el modelo Θ haya generado la secuencia de evidencias acústicas $\mathbf{X}^T$ mediante:

$$\Pr\left(\mathbf{X}^T|\Theta\right) = \sum_{\forall \mathbf{q}^T} \Pr\left(\mathbf{X}^T, \mathbf{q}^T|\Theta\right) \tag{2.10}$$

Asumiendo la independencia estadística de las evidencias acústicas en $\mathbf{X}^T$:

$$\begin{aligned}
\Pr\left(\mathbf{X}^T|\Theta\right) &= \sum_{\forall \mathbf{q}^T} \left\{\Pr\left(\mathbf{X}^T|\mathbf{q}^T,\Theta\right)\Pr\left(\mathbf{q}^T|\Theta\right)\right\} \\
&= \sum_{\forall \mathbf{q}^T} \left\{\prod_{t=1}^{T}\Pr\left(\mathbf{x}_t|q_t,\Theta\right)\prod_{t=2}^{T}\Pr\left(q_t|q_{t-1},\Theta\right)\right\} \\
&= \sum_{\forall \mathbf{q}^T} \left\{\prod_{t=1}^{T} b_{q_t}\left(\mathbf{x}_t\right)\prod_{t=2}^{T} a_{q_{t-1}q_t}\right\} \\
&= \sum_{\forall \mathbf{q}^T} \left\{b_{q_1}\left(\mathbf{x}_1\right)\prod_{t=2}^{T} b_{q_t}\left(\mathbf{x}_t\right)a_{q_{t-1}q_t}\right\}
\end{aligned}$$

que puede simplificarse en:

$$\Pr\left(\mathbf{X}^T|\Theta\right) = \sum_{\forall \mathbf{q}^T}\prod_{t=1}^{T} b_{q_t}\left(\mathbf{x}_t\right)a_{q_{t-1}q_t} \tag{2.11}$$

haciendo $a_{01} = 1$.

Una buena aproximación para $\Pr\left(\mathbf{X}^T|\Theta\right)$ es considerar la función de máximo en lugar de la sumatoria sobre las secuencias $\mathbf{q}^T$:

$$\Pr\left(\mathbf{X}^T|\Theta\right) \approx \max_{\forall \mathbf{q}^T}\left\{\Pr\left(\mathbf{X}^T|\mathbf{q}^T,\Theta\right)\Pr\left(\mathbf{q}^T|\Theta\right)\right\}$$

El algoritmo de Viterbi optimiza la búsqueda de esta máxima probabilidad. Para esto se define la variable de probabilidad acumulada:

$$\lambda_t(j) \triangleq \max_{\forall \mathbf{q}^{t-1}}\left\{\Pr\left(\mathbf{q}^{t-1}, q_t = j, \mathbf{X}^t|\Theta\right)\Pr\left(\mathbf{q}^{t-1}|\Theta\right)\right\}; \quad \forall j \in \mathcal{Q} \tag{2.12}$$

con $\lambda_0(j) = 1\ \forall j \in \mathcal{Q}$, y calculable por inducción mediante la recursión:

$$\begin{aligned}
\lambda_t(j) &= \max_{\forall i \in \mathcal{Q}} \{\lambda_{t-1}(i) \Pr(q_t = j, \mathbf{x}_t | q_{t-1} = i, \Theta)\} \\
&= \max_{\forall i \in \mathcal{Q}} \{\lambda_{t-1}(i) \Pr(q_t = j | q_{t-1} = i, \Theta) \Pr(\mathbf{x}_t | q_t = j, \Theta)\} \\
&= \max_{\forall i \in \mathcal{Q}} \{\lambda_{t-1}(i) \Pr(q_t = j | q_{t-1} = i, \Theta)\} \Pr(\mathbf{x}_t | q_t = j, \Theta) \\
&= \max_{\forall i \in \mathcal{Q}} \{\lambda_{t-1}(i) a_{ij}\} b_j(\mathbf{x}_t) \qquad (2.13)
\end{aligned}$$

de forma que:

$$\Pr\left(\mathbf{X}^T | \Theta\right) \approx \max_{\forall j \in \mathcal{Q}} \{\lambda_T(j)\}.$$

Para encontrar la secuencia de estados $\tilde{\mathbf{q}}^T$ asociada a la máxima probabilidad se define:

$$\xi_t(j) \triangleq \arg \max_{\forall i \in \mathcal{Q}} \{\lambda_{t-1}(i) a_{ij}\}$$

y a partir de:

$$\tilde{q}_T = \arg \max_{\forall i \in \mathcal{Q}} \{\lambda_T(i)\}$$

por recursión inversa:

$$\tilde{q}_t = \xi_{t+1}(\tilde{q}_{t+1}); \quad t = T-1, T-2, \ldots, 1 \qquad (2.14)$$

2.2.3. Reestimación de los parámetros

Dada una secuencia de evidencias acústicas $\mathbf{X}^T$, el entrenamiento consiste en maximizar la función de densidad de probabilidad $p(\mathbf{X}^T|\Theta)$, que posee la forma de (2.11). El método para el entrenamiento se fundamenta en la definición de una función auxiliar que guía el proceso de optimización permitiendo obtener una nueva estimación de los parámetros del modelo a

partir de la estimación anterior. La definición que se utiliza en este caso está basada en la teoría de la información[1] y tiene la siguiente forma:

$$\mathcal{O}(\Theta, \tilde{\Theta}) \triangleq \frac{1}{p(\mathbf{X}^T|\Theta)} \sum_{\forall \mathbf{q}^T} p(\mathbf{X}^T, \mathbf{q}^T|\Theta) \log p(\mathbf{X}^T, \mathbf{q}^T|\tilde{\Theta}) \tag{2.15}$$

donde Θ es la estimación inicial que se posee para el modelo y $\tilde{\Theta}$ es la nueva estimación. La normalización mediante $\mathbf{X}^T$ permite aplicar la función auxiliar a múltiples secuencias de entrenamiento (como se verá en la Sección 2.2.7).

El algoritmo de *maximización de la esperanza* es un caso particular del método de máxima verosimilitud que posee menor costo computacional [Duda et al., 1999]. Este algoritmo se basa en iterar haciendo en cada paso $\tilde{\Theta}$ igual a aquel Θ que haya maximizado la función auxiliar $\mathcal{O}$ en el paso anterior. Como requisito de convergencia, si en cualquier paso del algoritmo se verifica $\mathcal{O}(\Theta, \tilde{\Theta}) \geq \mathcal{O}(\Theta, \Theta)$, entonces debe cumplirse que $\Pr(\mathbf{X}^T|\tilde{\Theta}) \geq \Pr(\mathbf{X}^T|\Theta)$. Para el caso de la función auxiliar seleccionada en (2.15), puede encontrarse en [Huang et al., 1990] una demostración sencilla de que esta propiedad se cumple.

Para aplicar este algoritmo a la estimación de los parámetros del MOMC se debe obtener primero la ecuación completa para $\mathcal{O}(\Theta, \tilde{\Theta})$. A partir de (2.9) y (2.11) se puede escribir:

$$p\left(\mathbf{X}^T, \mathbf{q}^T|\Theta\right) = \sum_{k_1=1}^{N_c} \sum_{k_2=1}^{N_c} \cdots \sum_{k_T=1}^{N_c} \left\{ \prod_{t=1}^{T} b_{q_t k_t}\left(\mathbf{x}_t\right) a_{q_{t-1} q_t} \right\} c_{q_1 k_1} c_{q_2 k_2} \cdots c_{q_T k_T}$$

y así es posible redefinir (2.11) como:

$$\Pr\left(\mathbf{X}^T|\Theta\right) = \sum_{\forall \mathbf{q}^T} \sum_{\forall \mathbf{c}^T} \prod_{t=1}^{T} b_{q_t k_t}\left(\mathbf{x}_t\right) a_{q_{t-1} q_t} c_{q_t k_t} = \sum_{\forall \mathbf{q}^T} \sum_{\forall \mathbf{c}^T} p\left(\mathbf{X}^T, \mathbf{q}^T, \mathbf{c}^T|\Theta\right)$$

donde las $\mathbf{c}^T$ son las secuencias de la forma $c_{q_1 k_1}, c_{q_2 k_2}, \ldots, c_{q_T k_T}$.

[1] Número de Kullback-Leibler.

Para poder desarrollar completamente la función auxiliar de (2.15) queda por obtener:

$$\log p\left(\mathbf{X}^T, \mathbf{q}^T, \mathbf{c}^T \left| \tilde{\Theta} \right.\right) = \sum_{t=1}^{T} \log \tilde{b}_{q_t k_t}\left(\mathbf{x}_t\right) + \sum_{t=1}^{T} \log \tilde{a}_{q_{t-1} q_t} + \sum_{t=1}^{T} \log \tilde{c}_{q_t k_t}$$

y así la expresión de la función auxiliar queda convenientemente separada en:

$$\begin{aligned} \mathcal{O}(\Theta, \tilde{\Theta}) &= \frac{1}{p(\mathbf{X}^T|\Theta)} \sum_{\forall \mathbf{q}^T} \sum_{\forall \mathbf{c}^T} p\left(\mathbf{X}^T, \mathbf{q}^T, \mathbf{c}^T \left| \Theta \right.\right) \\ &\quad \times \left\{ \sum_{t=1}^{T} \log \tilde{b}_{q_t k_t}\left(\mathbf{x}_t\right) + \sum_{t=1}^{T} \log \tilde{a}_{q_{t-1} q_t} + \sum_{t=1}^{T} \log \tilde{c}_{q_t k_t} \right\} \\ &= \mathcal{O}_b(\Theta, \tilde{b}_{jk}) + \mathcal{O}_a(\Theta, \tilde{a}_{ij}) + \mathcal{O}_c(\Theta, \tilde{c}_{jk}) \end{aligned}$$

con:

$$\mathcal{O}_b(\Theta, \tilde{b}_{jk}) = \sum_{j=1}^{|\mathcal{Q}|} \sum_{\forall \mathbf{c}^T} \sum_{t=1}^{T} p\left(q_t = j, k_t = k \left| \mathbf{X}^T, \Theta \right.\right) \log \tilde{b}_{jk}\left(\mathbf{x}_t\right) \tag{2.16}$$

$$\mathcal{O}_a(\Theta, \tilde{a}_{ij}) = \sum_{i=1}^{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \sum_{\forall \mathbf{c}^T} p\left(q_{t-1} = i, q_t = j, \mathbf{c}^T \left| \mathbf{X}^T, \Theta \right.\right) \log \tilde{a}_{ij} \tag{2.17}$$

$$\mathcal{O}_c(\Theta, \tilde{c}_{jk}) = \sum_{j=1}^{|\mathcal{Q}|} \sum_{\forall \mathbf{c}^T} \sum_{t=1}^{T} p\left(q_t = j, k_t = k \left| \mathbf{X}^T, \Theta \right.\right) \log \tilde{c}_{jk} \tag{2.18}$$

Probabilidades de transición

En primer lugar considérese la función auxiliar (2.17), con la que se obtendrá la fórmula de reestimación para los a_{ij}. En este caso hay que tener en cuenta que la optimización está condicionada a:

$$\sum_{j=1}^{|\mathcal{Q}|} \tilde{a}_{ij} \stackrel{\circ}{=} 1 \quad \forall i \in \mathcal{Q}$$

Es por esto que conviene utilizar los multiplicadores de Lagrange escribiendo:

$$\nabla_{\tilde{a}} \left(\mathcal{O}_a(\Theta, \tilde{a}_{ij}) - \sum_{i=1}^{|\mathcal{Q}|} \ell_i \left(\sum_{j=1}^{|\mathcal{Q}|} \tilde{a}_{ij} - 1 \right) \right) = 0$$

Reemplazando (2.17) en esta ecuación y haciendo las derivadas parciales con respecto a los $\tilde{a}_{ij}$ se tiene:

$$\sum_{i=1}^{|\mathcal{Q}|} \sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \sum_{\forall \mathbf{c}^T} \left\{ p\left(q_{t-1} = i, q_t = j, \mathbf{c}^T \left| \mathbf{X}^T, \Theta \right.\right) \frac{1}{\tilde{a}_{ij}} \right\} - \ell_i = 0 \qquad (2.19)$$

que puede maximizarse considerando individualmente todos los términos de la sumatoria sobre los i.

Es necesario obtener primero los multiplicadores de Lagrange ℓ_i; multiplicando en ambos términos por los $\tilde{a}_{ij}$:

$$\sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \sum_{\forall \mathbf{c}^T} p\left(q_{t-1} = i, q_t = j, \mathbf{c}^T \left| \mathbf{X}^T, \Theta \right.\right) = \sum_{j=1}^{|\mathcal{Q}|} \tilde{a}_{ij} \ell_i$$

y así:

$$\begin{aligned} \ell_i &= \sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \sum_{\forall \mathbf{c}^T} p\left(q_{t-1} = i, q_t = j, \mathbf{c}^T \left| \mathbf{X}^T, \Theta \right.\right) \\ &= \sum_{t=1}^{T} \sum_{\forall \mathbf{c}^T} p\left(q_{t-1} = i, \mathbf{c}^T \left| \mathbf{X}^T, \Theta \right.\right) \end{aligned}$$

Volviendo a (2.19) ahora se obtiene:

$$\begin{aligned}\tilde{a}_{ij} &= \frac{\sum_{t=1}^{T}\sum_{\forall \mathbf{c}^T} p\left(q_{t-1}=i, q_t=j, \mathbf{c}^T \left| \mathbf{X}^T, \Theta\right.\right)}{\sum_{t=1}^{T}\sum_{\forall \mathbf{c}^T} p\left(q_{t-1}=i, \mathbf{c}^T \left| \mathbf{X}^T, \Theta\right.\right)} \\ &= \frac{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_{t-1}=i, q_t=j \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}}{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_{t-1}=i, \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}} \end{aligned} \tag{2.20}$$

Probabilidades de observación

Considérese ahora (2.18), para cuya optimización existe la restricción:

$$\sum_{k=1}^{N_c} \tilde{c}_{jk} = 1 \quad \forall j.$$

Este es un caso muy similar al de los a_{ij} y la fórmula de reestimación se deduce a partir de:

$$\nabla_{\tilde{c}} \left(\mathcal{O}_c(\Theta, \tilde{c}_{kj}) - \sum_{j=1}^{|\mathcal{Q}|} \ell_j \left(\sum_{k=1}^{N_c} \tilde{c}_{jk} - 1 \right) \right) = 0,$$

se reemplaza aquí (2.18) y nuevamente se obtienen las derivadas parciales, se despejan los multiplicadores de Lagrange y la fórmula de reestimación queda:

$$\tilde{c}_{jk} = \frac{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t=j, k_t=k \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}}{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t=j \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}} \tag{2.21}$$

Para completar la estimación de las probabilidades de observación resta deducir la fórmula de reestimación para los $b_{jk}(\mathbf{x}_t)$, que estaban definidos en función de los vectores de medias $\boldsymbol{\mu}_{jk}$ y las matrices de covarianzas $\mathbf{U}_{jk}$. Anulando $\nabla_{\tilde{b}}\mathcal{O}_b(\Theta, \tilde{b}_{jk})$, se puede derivar primero con respecto a los $\tilde{\boldsymbol{\mu}}_{jk}$ y obtener:

$$0 = \frac{\partial \mathcal{O}_b(\Theta, \tilde{b}_{jk})}{\partial \tilde{\boldsymbol{\mu}}_{jk}} = \sum_{j=1}^{|\mathcal{Q}|} \sum_{\forall \mathbf{c}^T} \sum_{t=1}^{T} p\left(q_t = j, k_t = k \left| \mathbf{X}^T, \Theta \right.\right) \tilde{\mathbf{U}}_{jk}^{-1} (\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})$$

desde donde se despejan los $\tilde{\boldsymbol{\mu}}_{jk}$ quedando:

$$\tilde{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)} \mathbf{x}_t}{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}} \tag{2.22}$$

De forma similar, a partir de $\nabla_{\tilde{b}}\mathcal{O}_b(\Theta, \tilde{b}_{jk}) = 0$ y derivando con respecto a los $\tilde{\mathbf{U}}_{jk}^{-1}$:

$$\begin{aligned} 0 &= \frac{\partial \mathcal{O}_b(\Theta, \tilde{b}_{jk})}{\partial \tilde{\mathbf{U}}_{jk}^{-1}} = \\ &= \sum_{j=1}^{|\mathcal{Q}|} \sum_{\forall \mathbf{c}^T} \sum_{t=1}^{T} p\left(q_t = j, k_t = k \left| \mathbf{X}^T, \Theta \right.\right) \frac{1}{2} \tilde{\mathbf{U}}_{jk}^{-1} - (\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})^{\mathrm{T}} \end{aligned}$$

de donde se despeja:

$$\tilde{\mathbf{U}}_{jk}^{-1} = \frac{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)} (\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})^{\mathrm{T}}}{\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left|\Theta\right.\right)}{p\left(\mathbf{X}^T \left|\Theta\right.\right)}} \tag{2.23}$$

Interpretaciones de las fórmulas de reestimación

Para llegar a una interpretación conceptual de estas fórmulas de reestimación es útil definir:

I) La variable α:

$$\alpha_t(i) \triangleq \Pr\left(\mathbf{x}_1, \ldots, \mathbf{x}_t, q_t = i \,|\Theta\right) \tag{2.24}$$

calculable de forma inductiva a partir de $\alpha_1(i) = b_i(\mathbf{x}_1)$ mediante $\alpha_t(j) = \sum_{\forall i \in \mathcal{Q}} \alpha_{t-1}(i) a_{ij} b_j(\mathbf{x}_t)$. Así se puede reescribir (2.11) como $\Pr\left(\mathbf{X}^T \,|\Theta\right) = \sum_{\forall i \in \mathcal{Q}} \alpha_T(i)$.

II) La variable β:

$$\beta_t(i) \triangleq \Pr\left(\mathbf{x}_{t+1}, \ldots, \mathbf{x}_T, q_t = i \,|\Theta\right) \tag{2.25}$$

que puede calcularse por inducción comenzando con $\beta_T(i) = 1/|\mathcal{Q}|$ y haciendo $\beta_t(j) = \sum_{\forall i \in \mathcal{Q}} a_{ji} b_i(\mathbf{x}_{t+1}) \beta_{t+1}(i)$. Ahora se pueden reescribir (2.11) como $\Pr\left(\mathbf{X}^T \,|\Theta\right) = \sum_{\forall i \in \mathcal{Q}} b_i(\mathbf{x}_1) \beta_1(i)$.

III) Las variables γ:

$$\gamma_t(i) \triangleq \Pr\left(q_t = i \,\middle|\, \mathbf{X}^T, \Theta\right) = \frac{\alpha_t(i)\beta_t(i)}{\sum\limits_{\forall q \in \mathcal{Q}} \alpha_T(q)} \tag{2.26}$$

que puede interpretarse como la cantidad de veces que el estado i es visitado en el instante de tiempo t, para observar la secuencia de evidencias acústicas $\mathbf{X}^T$.

$$\begin{aligned} \gamma_t(i,j) &\triangleq \Pr\left(q_{t-1} = i, q_t = j \,\middle|\, \mathbf{X}^T, \Theta\right) \\ &= \frac{\alpha_{t-1}(i) a_{ij} \sum\limits_{k=1}^{N_c} c_{jk} b_{jk}\left(\mathbf{x}_t\right) \beta_t(j)}{\sum\limits_{\forall q \in \mathcal{Q}} \alpha_T(q)} \end{aligned} \tag{2.27}$$

equivalente a pensar en la cantidad de veces que se ha llegado al estado j a partir del i, bajo las mismas condiciones anteriores.

IV) La variable ψ:

$$\begin{aligned}\psi_t(j,k) &\triangleq \Pr\left(q_t = j, k_t = k \,\middle|\, \mathbf{X}^T, \Theta\right) \\ &= \frac{\sum_{i\in\mathcal{Q}} \alpha_{t-1}(i) a_{ij} c_{jk} b_{jk}\left(\mathbf{x}_t\right) \beta_t(j)}{\sum_{\forall q\in\mathcal{Q}} \alpha_T(q)} \end{aligned} \tag{2.28}$$

interpretable como la cantidad esperada de veces en que se llegó al estado j en el tiempo t utilizando la gaussiana k, cuando se entrenaba el modelo Θ con la secuencia de evidencias acústicas $\mathbf{X}^T$.

Mediante estas definiciones pueden reescribirse las ecuaciones (2.20), (2.21), (2.22) y (2.23), respectivamente, como:

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} \qquad \tilde{c}_{jk} = \frac{\sum_{t=1}^{T} \psi_t(j,k)}{\sum_{t=1}^{T} \gamma_t(i)}$$

$$\tilde{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} \psi_t(j,k)\mathbf{x}_t}{\sum_{t=1}^{T} \psi_t(j,k)} \qquad \tilde{\mathbf{U}}_{jk}^{-1} = \frac{\sum_{t=1}^{T} \psi_t(j,k)(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})^{\mathrm{T}}}{\sum_{t=1}^{T} \psi_t(j,k)}$$

Escritas en esta forma, las fórmulas de reestimación se conocieron originalmente como parte del algoritmo de reestimación de *Baum-Welch*. En el trabajo original las probabilidades de observación eran discretas, con lo que se simplifican los $\tilde{c}_{jk}$, $\tilde{\boldsymbol{\mu}}_{jk}$ y $\tilde{\mathbf{U}}_{jk}^{-1}$ en un $\tilde{b}_j(x_k) = \sum_t \gamma_t(i)\delta(x_k, o_t) / \sum_t \gamma_t(i)$.

Por otro lado, si se realiza la búsqueda de la secuencia más probable $\tilde{\mathbf{q}}^T$ mediante el algoritmo de Viterbi (2.14) y se redefinen (2.26) y (2.27) de forma que solamente tomen valores 0 o 1 ($\gamma_t(i) = 1$ cuando $\tilde{q}_t = i$ y

$\gamma_t(i,j) = 1$ cuando $\tilde{q}_{t-1} = i \wedge \tilde{q}_t = j$), entonces al aplicar las fórmulas de reestimación y buscar la secuencia más probable sucesivamente se obtiene el denominado algoritmo de entrenamiento de Viterbi, que posee un costo computacional mucho menor al de Baum-Welch y tiene buen rendimiento en las aplicaciones prácticas de RAH.

Extensiones para modelos semicontinuos

Los MOM semicontinuos (MOMSC) surgen para reducir el número total de parámetros a estimar durante el entrenamiento. En los MOMC las probabilidades de observación $b_{jk}(\cdot)$ podían estar representadas arbitrariamente por cualquier distribución $\mathcal{N}(\cdot)$. Ahora, los MOMSC, podrán compartir un conjunto fijo de gaussianas conservando para cada estado la posibilidad de asignar diferentes pesos c_{jk} en la mezcla. Esto es conocido también como enlazado de parámetros. Se redefine (2.9) simplificando la dependencia entre los parámetros de $\mathcal{N}(\cdot)$ y el estado j:

$$b_j(\mathbf{x}_t) = \sum_{k=1}^{N_c} c_{jk} b_k(\mathbf{x}_t)$$

siendo en este caso:

$$b_k(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t, \boldsymbol{\mu}_k, \mathbf{U}_k) = \frac{1}{(2\pi)^{N_x} |\mathbf{U}_k|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(\mathbf{x}_t - \boldsymbol{\mu}_k)^{\mathrm{T}} \mathbf{U}_k^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_k)\right]}$$

La función auxiliar para la optimización (2.16) ahora se simplifica y queda:

$$\mathcal{O}_b(\Theta, \tilde{b}_k) = \sum_{\forall \mathbf{c}^T} \sum_{t=1}^{T} p\left(k_t = k \left| \mathbf{X}^T, \Theta \right.\right) \log \tilde{b}_k(\mathbf{x}_t)$$

y al igual que antes, derivando e igualando a cero, se obtienen:

$$\tilde{\boldsymbol{\mu}}_k = \frac{\sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left| \Theta \right.\right)}{p\left(\mathbf{X}^T \left| \Theta \right.\right)} \mathbf{x}_t}{\sum_{j=1}^{|\mathcal{Q}|} \sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \left| \Theta \right.\right)}{p\left(\mathbf{X}^T \left| \Theta \right.\right)}} \tag{2.29}$$

$$\tilde{\mathbf{U}}_k^{-1} = \frac{\sum_{j=1}^{|\mathcal{Q}|}\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \,|\Theta\right)}{p\left(\mathbf{X}^T \,|\Theta\right)} (\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_k)(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_k)^{\mathrm{T}}}{\sum_{j=1}^{|\mathcal{Q}|}\sum_{t=1}^{T} \frac{p\left(\mathbf{X}^T, q_t = j, k_t = k \,|\Theta\right)}{p\left(\mathbf{X}^T \,|\Theta\right)}} \tag{2.30}$$

que en comparación con (2.23) y (2.22) simplemente se han incorporado las sumatorias sobre j, calculando así la probabilidad sobre todos los estados en $p\left(\mathbf{X}^T, q_t = j, k_t = k \,|\Theta\right)$.

2.2.4. Concatenación de modelos

A partir del modelo genérico Θ es posible construir un conjunto con modelos de fonemas para el RAH:

$$\mathcal{F}_\Theta = \left\{ {}^F\Theta_\varphi \right\}; \quad \varphi \in \mathcal{F}$$

donde $\mathcal{F} = [1 \ldots |\mathcal{F}_\Theta|]$ es el conjunto de los fonemas para el reconocimiento. Un modelo de palabra se define como la concatenación de varios modelos de fonemas. El último estado de cada fonema se une directamente —con probabilidad de transición 1— al primero del siguiente conformando palabras:

$${}^W\Theta_w = {}^F\Theta_{\varphi_1}\, {}^F\Theta_{\varphi_2} \ldots {}^F\Theta_{\varphi_{N_w}}; \quad \varphi_f \in \mathcal{F} \tag{2.31}$$

a partir de un diccionario de pronunciaciones o transcripciones fonéticas:

$$\mathcal{W}_\varphi = \{(w; \varphi_1, \varphi_2, \ldots, \varphi_{N_w})\}; \quad N_w < \infty; \quad w \in \mathcal{W}$$

donde $\mathcal{W} = [1 \ldots |\mathcal{W}_\varphi|]$ es el conjunto de palabras para el reconocimiento. Estos modelos compuestos (MC) pueden ser vistos como un MOM de más estados y son tratados formalmente como se describió antes. Si el conjunto de estados de un MOM se pueden obtener mediante el funcional $\mathcal{Q}(\Theta)$, la cantidad de estados de un modelo de palabra es:

$$\left|\mathcal{Q}\left({}^{W}\Theta_{w}\right)\right| = \sum_{f=1}^{N_w} \left|\mathcal{Q}\left({}^{F}\Theta_{\varphi_f}\right)\right| \tag{2.32}$$

Ahora se puede construir el conjunto de modelos del vocabulario de reconocimiento:

$$\mathcal{W}_{\Theta} = \left\{{}^{W}\Theta_{w}\right\}; \quad w \in \mathcal{W}$$

2.2.5. Modelado estadístico del lenguaje

Sean $M, N \in \mathbb{N}; M, N < \infty$ y sea:

$$\mathbf{w}^{M} = w_1, w_2, \ldots, w_M; \quad w_m \in \mathcal{W} \tag{2.33}$$

una secuencia ordenada de M palabras a reconocer. Para cada palabra w_m en la secuencia, se define su historia de orden N como:

$$\mathbf{h}_m^N = w_{m-1}, w_{m-2}, \ldots, w_{m-N+1}; \quad w_{m-j} \in \mathcal{W}.$$

El modelo de lenguaje (ML) puede ser aproximado mediante la utilización de las denominadas n-gramáticas:

$$\Pr\left(\mathbf{w}^{M}\right) = \prod_{m=1}^{M} \Pr\left(w_m \left|\mathbf{h}_m^m\right.\right) \approx \prod_{m=1}^{M} \Pr\left(w_m \left|\mathbf{h}_m^N\right.\right) \triangleq G^N\left(\mathbf{w}^{M}\right) \tag{2.34}$$

La probabilidad de una palabra w_m, dada su historia $\mathbf{h}_m^N$, puede ser estimada simplemente mediante sus frecuencias de ocurrencia:

$$\Pr\left(w_m \left|\mathbf{h}_m^N\right.\right) \approx \frac{\mathcal{C}\left(w_m, \mathbf{h}_m^N\right)}{\mathcal{C}\left(\mathbf{h}_m^N\right)}$$

donde $\mathcal{C}(\cdot)$ es una función que cuenta las ocurrencias de una determinada secuencia de palabras en el corpus de entrenamiento.

Sin embargo, en muchos casos prácticos algunas historias $\mathbf{h}_m^N$ nunca aparecen en el corpus de entrenamiento. Es por esto que resulta necesario considerar el *suavizado* de las gramáticas. Por medio de estas técnicas, es posible estimar las probabilidades de las palabras cuyas historias de orden N nunca aparecen en el corpus de entrenamiento. Existen muchas técnicas útiles para el suavizado de gramáticas [Jelinek, 1999]. Un primer método sencillo es el denominado suavizado por interpolación lineal [Rabiner y Juang, 1993]. Dado un $K \in \mathbb{N}, 0 \leq K \leq N-1$ y la historia:

$$\mathbf{h}_m^K \ / \ \mathcal{C}\left(\mathbf{h}_m^K\right) > 0$$

se estiman las probabilidades para las historias inexistentes mediante:

$$I_m^K = \sum_{k=0}^{K} \iota_k \Pr\left(w_m \left| \mathbf{h}_m^k \right.\right) \tag{2.35}$$

con $0 \leq \iota_k \leq 1$ y $\sum \iota_k = 1$. Las historias $\mathbf{h}^1$ corresponden a una unigramática y la probabilidad para el caso de las historias $\mathbf{h}^0$ se define como:

$$\Pr\left(w_m \left| \mathbf{h}_m^0 \right.\right) \triangleq \frac{1}{|\mathcal{W}|} \quad \forall w_m \in \mathcal{W}.$$

Una de las técnicas más utilizadas para la estimación y suavizado de gramáticas es la denominada *back-off* [Potamianos y Jelinek, 1998]:

$$\Omega_m^K = \begin{cases} \dfrac{\mathcal{C}(w_m, \mathbf{h}_m^K) - \vartheta}{\mathcal{C}(\mathbf{h}_m^K)} & \text{si } \mathcal{C}\left(w_m, \mathbf{h}_m^K\right) > 0 \\ \varsigma(\mathbf{h}_m^K)\Omega_m^{K-1} & \text{si } \mathcal{C}\left(w_m, \mathbf{h}_m^K\right) = 0 \end{cases} \tag{2.36}$$

donde se fija empíricamente $\vartheta = 0{,}5$.

Para encontrar las probabilidades $\varsigma(\mathbf{h}_m^K)$ se debe considerar primeramente que:

$$\sum_{w_m / \mathcal{C}(w_m, \mathbf{h}_m^K) > 0} \frac{\mathcal{C}\left(w_m, \mathbf{h}_m^K\right) - \vartheta}{\mathcal{C}\left(\mathbf{h}_m^K\right)} + \sum_{w_m / \mathcal{C}(w_m, \mathbf{h}_m^K) = 0} \varsigma\left(\mathbf{h}_m^K\right) \Omega_m^{K-1} = 1$$

de esta forma:

$$\varsigma\left(\mathbf{h}_m^K\right)\left(\sum_{w_m/\mathcal{C}(w_m,\mathbf{h}_m^K)=0}\Omega_m^{K-1}\right)=\left(1-\sum_{w_m/\mathcal{C}(w_m,\mathbf{h}_m^K)>0}\frac{\mathcal{C}\left(w_m,\mathbf{h}_m^K\right)-\vartheta}{\mathcal{C}\left(\mathbf{h}_m^K\right)}\right)$$

y así:

$$\varsigma\left(\mathbf{h}_m^K\right)=\frac{1-\sum_{w_m/\mathcal{C}(w_m,\mathbf{h}_m^K)>0}\Omega_m^K}{1-\sum_{w_m/\mathcal{C}(w_m,\mathbf{h}_m^K)>0}\Omega_m^{K-1}}$$

2.2.6. Decodificación en el modelo compuesto

El MC es una estructura en red con todos los modelos de palabra conectados a partir de las probabilidades del ML. También es posible ver al MC como un gran MOM; si $|\mathcal{Q}|_{(m)}$ es el último estado del modelo de palabra ${}^{W}\Theta_{w_m}$ y $1_{(n)}$ el primero de ${}^{W}\Theta_{w_n}$, entonces se define la probabilidad de transición entre las dos palabras del MC como:

$$a_{|\mathcal{Q}|_{(m)},1_{(n)}}\triangleq G_{mn}^{(2)} \tag{2.37}$$

quedando así definida la estructura del MC ${}^{C}\Theta$ para una frase completa[2] o, si se quiere, para cualquier frase posible dado el conjunto de palabras $\mathcal{W}$ y el ML que la relaciona.

En la extensión del algoritmo de Viterbi se requiere incorporar las probabilidades del ML en el proceso de búsqueda sobre el MC. Dadas las palabras $w_m, w_n \in \mathcal{W}$, se utilizará la siguiente notación:

$i_{(m)}, j_{(m)}$: estados pertenecientes al modelo de la palabra w_m,

$q_{(m)t}$: estado de ${}^{W}\Theta_{w_m}$ en el tiempo t,

[2]La diferencia de esta concatenación de modelos en relación a (2.31) radica en que la probabilidad de transición entre dos modelos de palabra queda definida por el ML mientras que la probabilidad de transición entre fonemas era siempre 1.

$\mathbf{q}_{(m)}^{T}$: secuencia de T estados en ${}^{W}\Theta_{w_m}$ y

$G_{mn}^{(2)}$: probabilidad de que se emita w_n con una historia $\mathbf{h}_n^2 = w_m$ (ver ecuación (2.34))

Considerando un ML de bi-gramática es posible redefinir la probabilidad acumulada de la ecuación (2.12) como:

$$\Lambda_t(j_{(n)}) \triangleq \max_{\forall \mathbf{q}_{(n)}^{t-1}} \left\{ \Pr\left(\mathbf{q}_{(n)}^{t}, q_{(n)t} = j_{(n)}, \mathbf{X}^t \,\middle|\, {}^{W}\Theta_{w_m} \right) \right\}$$

con las inicializaciones:

$$\Lambda_0(j_{(n)}) = 1 \quad \forall w_n \in \mathcal{W}, \forall j_{(n)} \in \mathcal{Q}\left({}^{W}\Theta_{w_n}\right)$$

y cuando comienza cada palabra[3]:

$$\Lambda_{t-1}(j_{(n)} = 1) = \max_{\forall w_m \in \mathcal{W}} \left\{ \Lambda_{t-1}\left(i_{(m)} = \left| \mathcal{Q}\left({}^{W}\Theta_{w_m}\right) \right| \right) G_{mn}^{(2)} \right\}.$$

Luego, es posible expandir esta probabilidad acumulada como:

$$\Lambda_t\left(j_{(n)}\right) = \max_{\forall \mathbf{q}_{(n)}^{t-1}} \left\{ \Pr\left(\mathbf{q}_{(n)}^{t-1}, \mathbf{x}^t \,\middle|\, {}^{W}\Theta_{w_n} \right) \Pr\left(q_{(n)t} = j_{(n)}, \mathbf{X}^t \,\middle|\, \mathbf{q}_{(n)}^{t-1}, {}^{W}\Theta_{w_n} \right) \right\}$$

y calcularla por inducción mediante:

$$\begin{aligned}
\Lambda_t\left(j_{(n)}\right) &= \max_{\forall i_{(n)}} \left\{ \Lambda_{t-1}(i_{(n)}) \Pr\left(q_{(n)t} = j_{(n)}, \mathbf{x}_t \,\middle|\, q_{(n)t-1} = i_{(n)}, {}^{W}\Theta_{w_n} \right) \right\} \\
&= \max_{\forall i_{(n)}} \left\{ \Lambda_{t-1}(i_{(n)}) \Pr\left(q_{(n)t} = j_{(n)} \,\middle|\, q_{(n)t-1} = i_{(n)}, {}^{W}\Theta_{w_n} \right) \right. \\
&\qquad \left. \times \Pr\left(\mathbf{x}_t \,\middle|\, q_{(n)t} = j_{(n)}, {}^{W}\Theta_{w_n} \right) \right\} \\
&= \max_{\forall i_{(n)}} \left\{ \Lambda_{t-1}(i_{(n)}) \Pr\left(q_{(n)t} = j_{(n)} \,\middle|\, q_{(n)t-1} = i_{(n)}, {}^{W}\Theta_{w_n} \right) \right\} \\
&\quad \times \Pr\left(\mathbf{x}_t \,\middle|\, q_{(n)t} = j_{(n)}, {}^{W}\Theta_{w_n} \right)
\end{aligned}$$

[3]Obsérvese que en la transición entre dos palabras el modelo no emite y por lo tanto no cambia el índice de tiempo t.

$$\Lambda_t\left(j_{(n)}\right) \quad = \quad \underset{\forall i_{(n)}}{\text{máx}} \left\{ \Lambda_{t-1}(i_{(n)}) a_{i_{(n)} j_{(n)}} \right\} b_{j_{(n)}}\left(\mathbf{x}_t\right)$$

Para obtener la secuencia más probable a partir de las probabilidades acumuladas se define:

$$\Xi_t(j_{(n)} \triangleq \arg \underset{\forall i_{(n)}}{\text{máx}} \left\{ \Lambda_{t-1}(i_{(n)}) a_{i_{(n)} j_{(n)}} \right\}$$

con la salvedad de que:

$$\Xi_t\left(j_{(n)} = \left|\mathcal{Q}\left({}^{W}\Theta_{w_n}\right)\right|\right) = \arg \underset{\forall i_{(m)}=1}{\text{máx}} \left\{ \Lambda_t(i_{(m)}) G^{(2)}_{mn} \right\}$$

Ahora, por recursión inversa:

$$\tilde{q}_{(n)t} = \Xi_{t+1}\big(\tilde{q}_{(n)t+1}\big); \quad t = T-1, T-2, \ldots, 1$$

comenzando por:

$$\tilde{q}_{(n)T} = \arg \underset{\forall i_{(n)}, \forall w_n}{\text{máx}} \left\{ \Lambda_T(i_{(n)}) \right\}.$$

y con las restricciones:

$$\tilde{q}_{(n)T} \stackrel{\circ}{=} |\mathcal{Q}|_{(n)} \qquad \wedge \qquad \tilde{q}_{(n)1} \stackrel{\circ}{=} 1_{(n)}$$

La secuencia resultante está restringida por este algoritmo a una secuencia de palabras válidas ya que los fonemas están concatenados en palabras (2.31) y no hay conexiones hacia afuera de las palabras que no sean a través de las conexiones impuestas por el ML (siempre desde el último estado de una palabra hacia el primero de otra). Por lo tanto, dado que en esta secuencia quedan especificados tanto el número de estado como la palabra

a la que cada uno pertenece, se puede extraer directamente de ella la transcripción reconocida. Estas ecuaciones son la base del denominado algoritmo de *decodificación* para RAH. Se agregan además mejoras de índole práctico como el escalado o la aritmética logarítmica para reducir los errores introducidos por la precisión limitada en el cómputo [Rabiner y Juang, 1993]. Otras mejoras ampliamente utilizadas son las técnicas de podado, que reducen significativamente el espacio de la búsqueda en el algoritmo de Viterbi. Por ejemplo, en el algoritmo de *beam search* se utiliza una probabilidad Φ_Λ como umbral de podado y no se consideran los caminos que acumulan una probabilidad Φ_Λ veces menor que el máximo para cada tiempo t. Puede consultarse una revisión acerca de estos métodos en [Ney y Ortmanns, 1999].

2.2.7. Entrenamiento del modelo compuesto

Es necesario dar respuesta a tres cuestiones importantes para encontrar las fórmulas de reestimación en el MC. La primera tiene que ver con la relación entre el entrenamiento de los MOM de cada fonema y la estimación de las probabilidades del ML. La segunda cuestión se plantea al considerar múltiples secuencias —es decir, muchas frases— de entrenamiento, ya que las fórmulas de reestimación siempre se dedujeron a partir de una única secuencia de evidencias acústicas. La tercera cuestión tiene que ver con la forma en que los diferentes MOMSC, que forman el MC, van a compartir sus parámetros y las modificaciones que esto demanda en las fórmulas de reestimación.

La solución práctica más empleada para la primera cuestión es muy simple y consiste en estimar las probabilidades asociadas con el ML separadamente (por ejemplo mediante (2.35) o (2.36)), dejándolas fijas durante las reestimaciones de todos los restantes parámetros del MC [Young et al., 2000].

Para extender las fórmulas de reestimación a múltiples secuencias de evidencias acústicas, considérese que existen N_X secuencias de entrenamiento:

$$\mathbf{X} = \mathbf{X}_1^{T_1}, \mathbf{X}_2^{T_2}, \ldots, \mathbf{X}_{N_X}^{T_{N_X}}$$

Asumiendo la independencia estadística entre las diferentes secuencias, la ecuación (2.10) debe reescribirse como:

$$\Pr\left(\mathbf{X}\,|\Theta\right) = \prod_{n=1}^{N_X} \sum_{\forall \mathbf{q}_n^{T_n}} \Pr\left(\mathbf{X}_n^{T_n}, \mathbf{q}_n^{T_n}\,|\Theta\right)$$

lo cual agrega simplemente una sumatoria sobre todas las secuencias tanto en el numerador como en el denominador de las fórmulas de reestimación. Por ejemplo, para las probabilidades de transición:

$$\tilde{a}_{ij} = \frac{\sum_{n=1}^{N_X} \sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t-1} = i, q_{n,t} = j\,|\Theta\right)}{p\left(\mathbf{X}_n^{T_n}\,|\Theta\right)}}{\sum_{n=1}^{N_X} \sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t-1} = i,|\Theta\right)}{p\left(\mathbf{X}_n^{T_n}\,|\Theta\right)}} \tag{2.38}$$

Durante el proceso de entrenamiento, además de contar con las secuencias de evidencias acústicas $\mathbf{X}$, también se poseen las transcripciones en palabras para cada secuencia:

$$\mathbf{W} = \mathbf{w}_1^{T_1}, \mathbf{w}_2^{T_2}, \ldots, \mathbf{w}_{N_X}^{T_{N_X}}$$

donde cada transcripción $\mathbf{w}_n^{T_n}$ es una secuencia de T_n palabras como en (2.33):

$$\mathbf{w}_n^{T_n} = w_{n,1}, w_{n,2}, \ldots, w_{n,T_n}; \quad w_{n,m} \in \mathcal{W}$$

A partir de una de estas transcripciones y del diccionario fonético $\mathcal{W}_\varphi$ es posible construir un MC con la concatenación de palabras:

$${}^{C}\Theta_n = {}^{W}\Theta_{w_{n,1}}\,{}^{W}\Theta_{w_{n,2}} \ldots {}^{F}\Theta_{w_{n,T_n}}$$

con probabilidades fijas entre las palabras. De forma similar a (2.37), se puede hacer:

$$a_{|\mathcal{Q}|_{(m-1)},1_{(m)}} \triangleq P$$

donde P, en general, es 1.

A partir de cada uno de los MC construidos, deben estimarse todos los parámetros de los MOM que los componen. En este esquema de entrenamiento debe considerarse que el mismo modelo de fonema o palabra aparecerá en distintas partes del MC y en distintos MC para distintas frases. Al considerar que en uno de estos MC existen conjuntos de estados que comparten sus parámetros surge naturalmente la tercera cuestión, acerca de las diversas formas de compartir los parámetros en el MC. Se podrían compartir los parámetros correspondientes a los estados de una misma palabra o de un mismo fonema. También se podrían compartir parámetros de sonidos similares desde el punto de vista de la fonética acústica o bien utilizar métodos automáticos para encontrar qué conjunto de estados conviene que compartan parámetros.

A continuación se va a considerar que los estados que comparten parámetros se agrupan en conjuntos $\mathcal{Q}_{(m)}$. Estos conjuntos de estados se encontrarán previamente definidos según algún criterio[4] y se utilizará una extensión de la notación $i_{(m)}$ y $j_{(m)}$ para indicar que estos estados pertenecen a la clase m. Anteriormente, se utilizaron subíndices similares para indicar la pertenencia de un estado al conjunto de estados de una palabra. Ahora, en un sentido más amplio, una clase m puede corresponderse con cualquier conjunto de estados arbitrariamente agrupados[5]. De forma similar, como cada clase m posee su propia mezcla de gaussianas, se deben definir los conjuntos de mezclas de gaussianas $\mathcal{M}_{(m)}$, cada uno con $N_{c_{(m)}}$ gaussianas[6]. Para indicar la pertenencia de una gaussiana k al conjunto de gaussianas de la clase m se utilizará la notación $k_{(m)}$.

Así como en (2.29) y (2.30) se compartían los parámetros de las mezclas de gaussianas entre los estados de un único modelo, ahora se generaliza la idea de MOMSC hacia los MC con múltiples secuencias. Siguiendo de (2.20) y (2.38), las probabilidades de transición entre los estados $i_{(m)}, j_{(m)} \in \mathcal{Q}_{(m)}$, se reestiman mediante:

[4]En los experimentos de RAH que se detallan en capítulos posteriores, se compartieron los parámetros de las mezclas de gaussianas para: 1) todos los estados de un mismo fonema y 2) todos los estados del modelo de silencio y del modelo de pausa corta al final de cada palabra.

[5]Las palabras como entidades independientes han desaparecido en los MC para el entrenamiento, salvo la situación particular en que las clases m coincidan con las palabras, para lo cual la notación tampoco es contradictoria.

[6]En general $N_{c_{(m)}}$ es el mismo para todas las clases.

$$\tilde{a}_{i_{(m)}j_{(m)}} = \frac{\sum_{n=1}^{N_X}\sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t-1}=i_{(m)}, q_{n,t}=j_{(m)}\,\middle|{}^{C}\Theta_n\right)}{p\left(\mathbf{X}_n^{T_n}\,\middle|{}^{C}\Theta_n\right)}}{\sum_{n=1}^{N_X}\sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t-1}=i_{(m)},\,\middle|{}^{C}\Theta_n\right)}{p\left(\mathbf{X}_n^{T_n}\,\middle|{}^{C}\Theta_n\right)}} \tag{2.39}$$

En el caso del peso de la gaussiana $k_{(m)}$ con que se modela la probabilidad de observación del estado $j_{(m)}$, a partir de (2.21):

$$\tilde{c}_{j_{(m)}k_{(m)}} = \frac{\sum_{n=1}^{N_X}\sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t}=j_{(m)}, k_t=k_{(m)}\,\middle|{}^{C}\Theta_n\right)}{p\left(\mathbf{X}_n^{T_n}\,\middle|{}^{C}\Theta_n\right)}}{\sum_{n=1}^{N_X}\sum_{t=1}^{T_n} \frac{p\left(\mathbf{X}_n^{T_n}, q_{n,t}=j_{(m)}\,\middle|{}^{C}\Theta_n\right)}{p\left(\mathbf{X}_n^{T_n}\,\middle|{}^{C}\Theta_n\right)}} \tag{2.40}$$

Al igual que en (2.28), se pueden simplificar las expresiones definiendo:

$$\psi_{n,t}(j_{(m)}, k_{(m)}) = \Pr\left(q_{n,t}=j_{(m)}, k_{n,t}=k_{(m)}\,\middle|\mathbf{X}_n^{T_n}, {}^{C}\Theta_n\right)$$

Dado que los parámetros de las gaussianas se comparten para una misma clase m, a partir de (2.29):

$$\tilde{\boldsymbol{\mu}}_{k_{(m)}} = \frac{\sum_{n=1}^{N_X}\sum_{\forall j_{(m)}\in\mathcal{Q}_{(m)}({}^{C}\Theta_n)}\sum_{t=1}^{T_n}\psi_{n,t}(j_{(m)},k_{(m)})\mathbf{x}_{n,t}}{\sum_{n=1}^{N_X}\sum_{\forall j_{(m)}\in\mathcal{Q}_{(m)}({}^{C}\Theta_n)}\sum_{t=1}^{T_n}\psi_{n,t}(j_{(m)},k_{(m)})} \tag{2.41}$$

y a partir de (2.30):

$$\tilde{\mathbf{U}}_{k_{(m)}}^{-1} =$$

$$= \frac{\sum_{n=1}^{N_X} \sum_{\forall j_{(m)} \in \mathcal{Q}_{(m)}({}^{C}\Theta_n)} \sum_{t=1}^{T_n} \psi_{n,t}(j_{(m)}, k_{(m)})(\mathbf{x}_{n,t} - \tilde{\boldsymbol{\mu}}_{j_{(m)}k_{(m)}})(\mathbf{x}_{n,t} - \tilde{\boldsymbol{\mu}}_{j_{(m)}k_{(m)}})^{\mathrm{T}}}{\sum_{n=1}^{N_X} \sum_{\forall j_{(m)} \in \mathcal{Q}_{(m)}({}^{C}\Theta_n)} \sum_{t=1}^{T_n} \psi_{n,t}(j_{(m)}, k_{(m)})} \tag{2.42}$$

Capítulo 3

Prosodia y acentuación en el discurso continuo

En este capítulo se presenta una serie de estudios orientados a esclarecer la forma en que se relacionan la acentuación y los rasgos prosódicos en el discurso continuo del español. Este es el primer paso para la incorporación de información acentual en un sistema de reconocimiento automático del habla. Luego de estudiar las relaciones entre acentuación y rasgos prosódicos, resta encontrar un sistema que obtenga la acentuación a partir de la señal de voz y otro que la incorpore a un reconocedor automático del habla. Estas etapas se tratarán en los siguientes dos capítulos.

El presente capítulo se encuentra dividido en tres partes. En la primera se discute acerca de la relación entre acentuación y rasgos prosódicos en palabras aisladas, mencionando algunos antecedentes al respecto. En la segunda parte se describen las frases analizadas y se detallan algunas características de la estructura acentual de sus palabras. En la tercera parte del capítulo se describen las relaciones entre acentuación y rasgos prosódicos para diferentes formas de caracterizar la energía, la frecuencia fundamental y la duración. En particular se hace un análisis más detallado de la curva de frecuencia fundamental considerando sus máximos, mínimos, tendencias en la frase y cadencias en cada sílaba. En este sentido se han evaluado muy diversas alternativas y aquí se presenta una selección de los resultados más importantes.

3.1. La acentuación y su manifestación prosódica

Como se destacó en el Sección 1.3.3, la tipología acentual del español es *libre*, de igual forma que en el inglés, el alemán o el italiano. Sin embargo, en el inglés no se describen reglas ortográficas que permitan saber la acentuación de una palabra a partir de su grafía. De forma similar que para la transcripción fonética, en el español existe un conjunto de reglas que permite saber cuál es la estructura acentual (EA) de una palabra a partir de su grafía.

Se denomina *acentuación prosódica* a la manifestación del acento en los rasgos prosódicos de una emisión de voz. En cambio, se denomina simplemente *acentuación* a la representación del acento en el lenguaje escrito, ya sea de forma explícita —a través de la tilde— o implícita, según lo definen las reglas ortográficas. En el Capítulo 1 se destacó la alta correlación que existe entre ambas acentuaciones cuando las palabras se pronuncian en forma *aislada*. En este caso se puede observar que en la sílaba tónica se encuentran los máximos de frecuencia fundamental (F_0), energía y duración del núcleo vocálico. En las Figuras 3.1 y 3.3 se muestran los espectrogramas de dos ejemplos sencillos para ilustrar las correspondencias entre acentuación y rasgos prosódicos en palabras aisladas. En las Figuras 3.2 y 3.4 se muestran las curvas de energía, F_0 y duración del núcleo vocálico para estos mismos ejemplos.

Sin embargo, este fenómeno no se presenta tan claramente cuando la palabra está inmersa en un discurso de habla continua. Los estudios realizados por [Quilis, 1993] acerca de la realización del acento en el discurso continuo indican que un 36.56 % de las palabras del español pueden ser consideradas como inacentuadas. Dentro de las palabras inacentuadas un 90.23 % son monosilábicas. Sin embargo, estas palabras no serán de interés fundamental para el análisis dado que su tonicidad silábica no se puede comparar en forma relativa dentro de sí mismas sino que sería necesario referirla a la frase. Dentro de las restantes palabras inacentuadas se encuentran varios grupos de interés que fueron analizados por el autor. Ya se citó la relación que Quilis describe entre la función gramatical que cumplen las palabras y el hecho de que sean acentuadas o no. Por ejemplo, la distinción de la preposición *para*, que es inacentuada, y el verbo en segunda persona del singular *para*, que es acentuado.

El trabajo de Quilis constituye el punto de partida para los estudios que aquí se presentan. En este capítulo se profundizará el análisis para revelar algunas otras características de interés que vinculan la acentuación y los rasgos prosódicos en el discurso continuo. En base a los antecedentes,

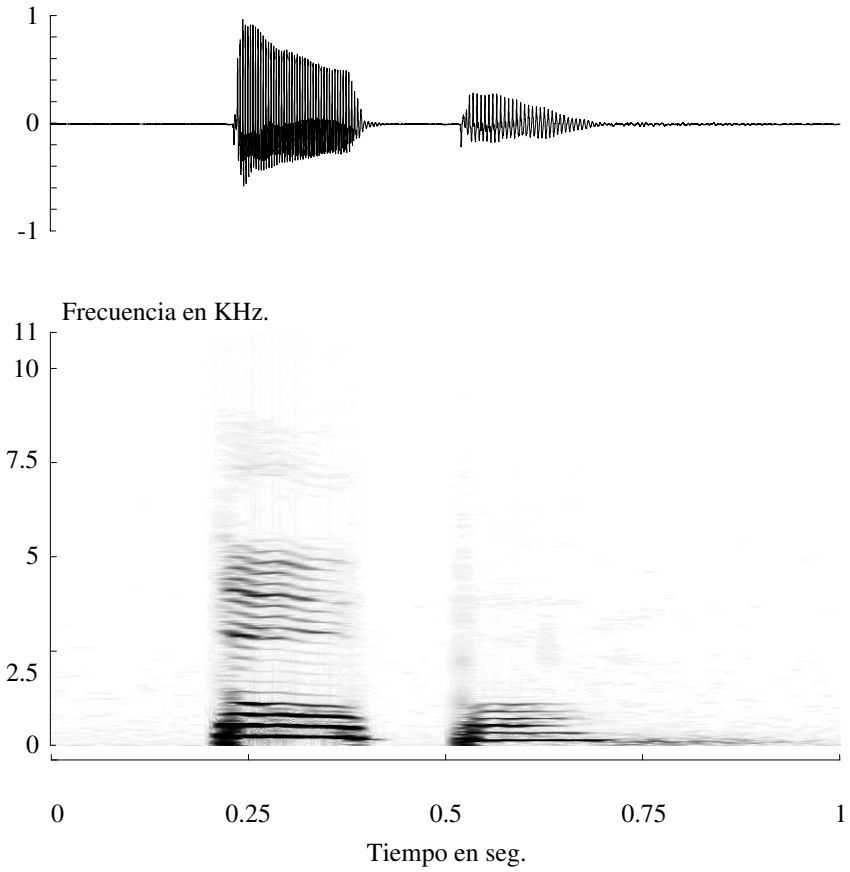

Figura 3.1. Espectrograma para la palabra *topo* /tópo/.

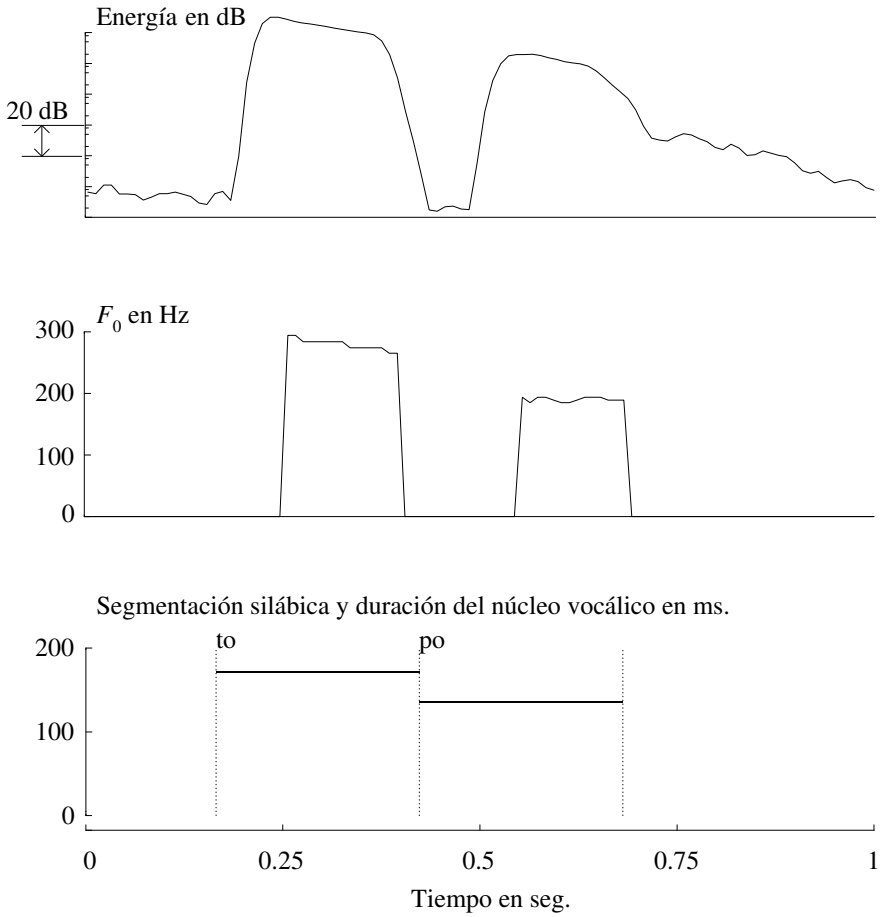

Figura 3.2. Curvas de rasgos prosódicos para la palabra *topo* /tópo/. En la curva de abajo se debe tener en cuenta que el tiempo de segmentación corresponde a la sílaba completa mientras que la duración corresponde solamente al núcleo vocálico de la sílaba.

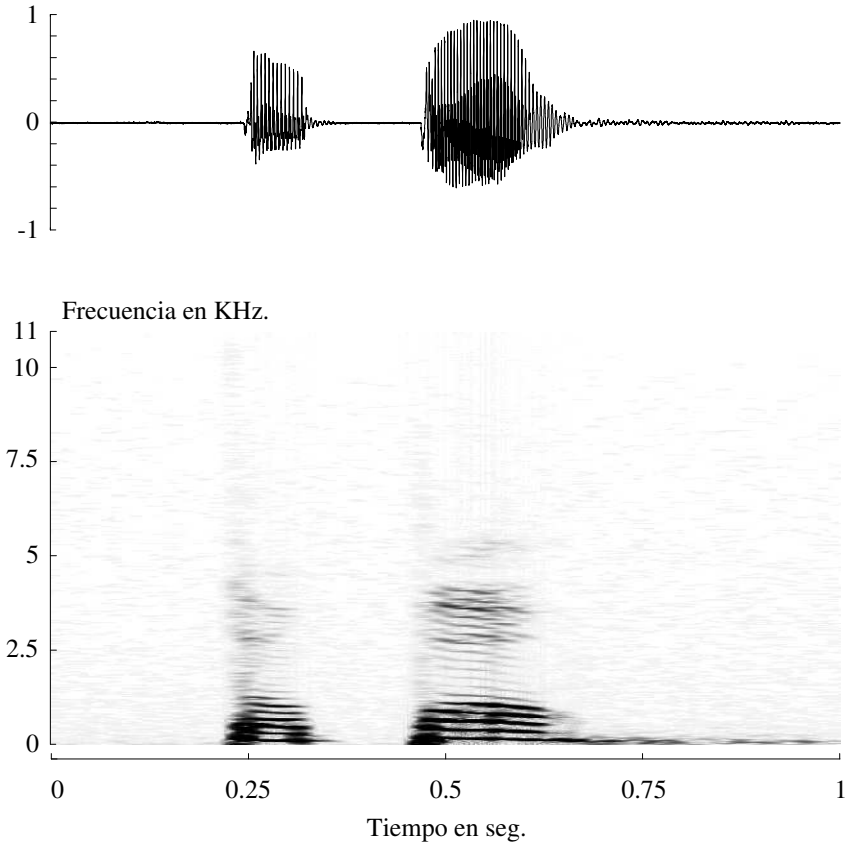

Figura 3.3. Espectrograma para la palabra *topó* /topó/.

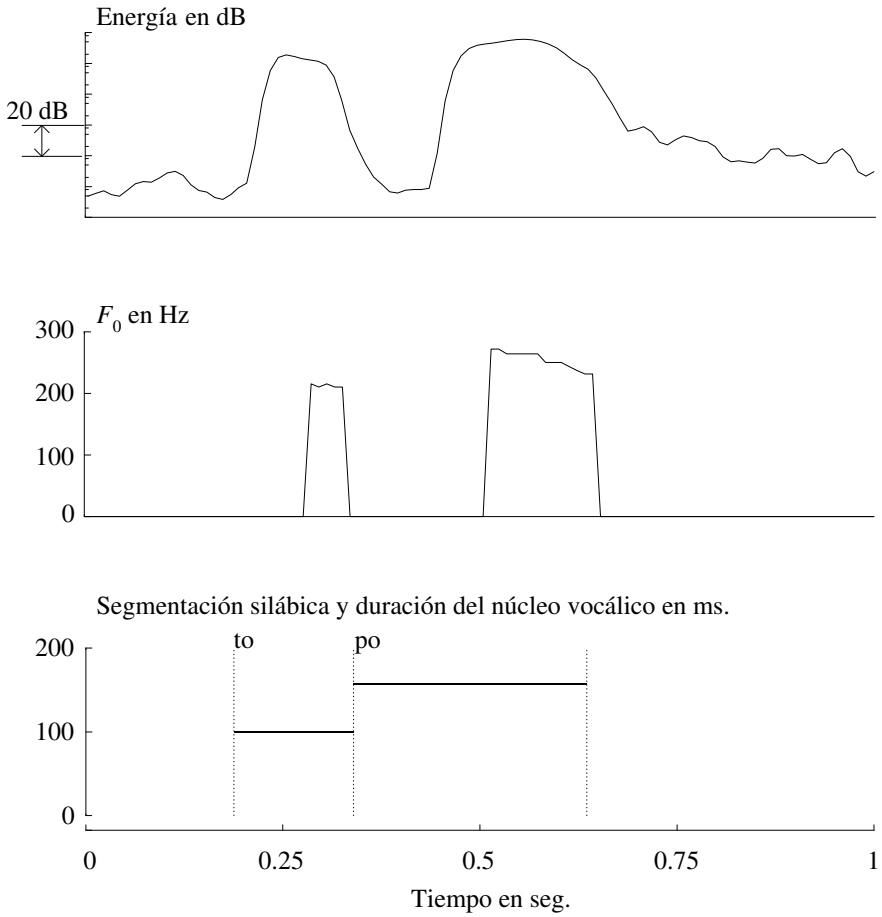

Figura 3.4. Curvas de rasgos prosódicos para la palabra *topó* /topó/. En la curva de abajo se debe tener en cuenta que el tiempo de segmentación corresponde a la sílaba completa mientras que la duración corresponde solamente al núcleo vocálico de la sílaba.

es de esperar que la correspondencia entre la acentuación y su manifestación prosódica se pierda en un grado considerable. Para comenzar se darán algunas características generales de las frases que se analizaron.

3.2. Acentuación

Para los estudios se utilizó un subconjunto de frases del corpus de habla Albayzin (que aquí denominaremos SC1; para más detalles véase el Apéndice A.2). El vocabulario de las frases analizadas contaba con 202 palabras relacionadas con la geografía de España. Descontando las palabras monosilábicas, se analizaron un total de 2929 palabras. La cantidad de sílabas por palabra en las frases analizadas era:

- palabras de 2 sílabas: 1722,
- palabras de 3 sílabas: 463,
- palabras de 4 sílabas: 600 y
- palabras de 5 sílabas: 144.

Toda el corpus de habla se procesó de forma automática a partir de las transcripciones de las frases. Para cada frase se aplicaron las reglas para la separación en sílabas y luego se utilizaron las reglas de acentuación (ver Sección 1.3.3) para obtener la EA de cada palabra. Se tuvieron en cuenta los estudios de Quilis antes mencionados para asignar correctamente las EA para las palabras inacentuadas (ver Sección A.2.3).

3.2.1. Palabras

En la Tabla 3.1 se muestra la distribución de EA para todas palabras analizadas. En esta tabla se incluyen también las palabras inacentuadas, donde se considera que ninguna de sus sílabas es tónica. Es válido aclarar nuevamente que si bien las reglas ortográficas del español sólo permiten un acento por palabra, existen casos especiales, como los adverbios terminados en -*mente*, que poseen dos acentos prosódicos.

También resulta interesante conocer las posiciones relativas del acento dentro de la palabra. Para las palabras analizadas en este estudio se encontró la distribución de la Tabla 3.2.

EA	Cantidad	EA	Cantidad
/AT/	247	/AATA/	197
/TA/	1434	/ATAA/	220
/AAT/	186	/AAATA/	144
/ATA/	202	/AA/	41
/TAA/	46	/AAA/	29
/AAAT/	171	/AAAA/	12

Tabla 3.1. Cantidad de cada tipo de estructura acentual en el corpus de habla analizado.

Comienzan con	Cantidad	Terminan con	Cantidad
/T-/	1480	/-T/	604
/AT-/	669	/-TA/	1977
/AAT-/	383	/-TAA/	266
/AAAT-/	315	/-TAAA/	-

Tabla 3.2. Posición del acento en relación al comienzo y final de la palabra en el corpus de habla analizado (no se listan las 82 palabras inacentuadas).

3.2.2. Frases

Se analizó un total de 600 frases pronunciadas por 6 hablantes femeninos y 6 masculinos. En estas frases había 342 de tipo declarativa y 258 interrogativas. Las frases analizadas tenían entre 3 y 25 palabras. En la Figura 3.5 se presenta la forma en que estas cantidades se distribuyen.

Para tener una mejor idea de las frases analizadas, en la Tabla 3.3 se presentan algunos ejemplos con su separación en sílabas y sus correspondientes EA.

3.3. Relaciones entre prosodia y acentuación

3.3.1. Medición de los rasgos prosódicos

Para la *energía* se utilizó una estimación por tramos como en la ecuación (2.8). En este estudio los tramos de energía se calcularon con un paso de 10 ms y un ancho de ventana de 52 ms (ver Sección 2.1.1).

Con los mismos parámetros para el análisis por tramos se estimó la F_0 mediante la técnica basada en coeficientes cepstrales que se describió en la Sección 2.1.4.

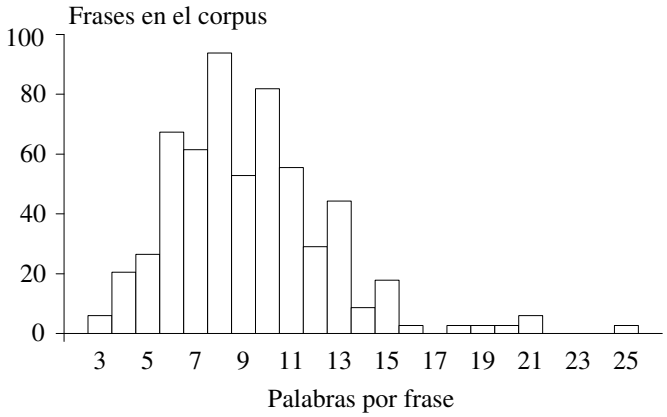

Figura 3.5. Distribución de la cantidad de palabras por frase en el corpus de habla analizado.

Frase 1
Nombre de las tres comunidades de menor extensión Nom–bre+de+las+tres+co–mu–ni–da–des+de+me–nor+ex–ten–sión /TA A A A AAATA A AT AAT/
Frase 2
¿Cuántos ríos con caudal mayor de ochocientos metros cúbicos por segundo pasan por la Comunidad Valenciana? Cuán–tos+rí–os+con+cau–dal+ma–yor+de+o–cho–cien–tos +me–tros+cú–bi–cos+por+se–gun–do+pa–san+por+la +Co–mu–ni–dad+Va–len–cia–na /TA TA A AT AT A AATA TA TAA A ATA TA A A AAAT AATA/
Frase 3
Todos los ríos To–dos+los+rí–os /TA A TA/

Tabla 3.3. Tres ejemplos de las frases analizadas con su separación silábica y sus estructuras acentuales.

Para la *duración* se consideró el núcleo vocálico de cada sílaba distinguiendo también los formados por diptongos. Se entrenó un sistema de reconocimiento automático del habla basado en modelos ocultos de Markov (MOM) y se utilizaron las transcripciones correctas para buscar, en cada frase, la secuencia más probable mediante el algoritmo de Viterbi (todo según los métodos descritos en la Sección 2.2). En la Figura 3.6 se observan las curvas de energía, F_0 y duración para la Frase 1 de la Tabla 3.3.

3.3.2. Máximos prosódicos

Considerando que en las palabras *aisladas* se caracteriza la sílaba tónica por tener mayor energía, F_0 y duración del núcleo vocálico, se analizará cómo se cumplen estas simples reglas para el caso del discurso continuo en la corpus de habla analizado.

Se calculó el porcentaje de coincidencias entre el máximo de alguno de los rasgos prosódicos y la acentuación. Este porcentaje fue calculado dividiendo las veces que el máximo estaba en el lugar correcto por el total de aciertos considerados en las combinaciones. En la Tabla 3.4 se observa, en primer lugar, que en un 17.71 % de los casos no coincide el máximo de ninguno de los tres rasgos prosódicos con la sílaba tónica. Sin embargo, en esta tabla ya es posible observar que tanto los máximos de energía como los de duración coinciden con la acentuación en más ocasiones que los máximos de F_0.

Máx. Ener.	Máx. F_0	Máx. Dur.	Coincidencias %
			17.71
		○	18.03
	○		4.60
	○	○	8.19
○			13.14
○		○	17.26
○	○		6.34
○	○	○	14.68

Tabla 3.4. Porcentaje de coincidencias entre los máximos de energía, frecuencia fundamental y duración, con la acentuación. Los círculos ○ indican qué rasgos prosódicos tuvieron la coincidencia. El porcentaje es relativo al total de coincidencias encontradas con todas las combinaciones. En la primera línea se especifica la cantidad de casos en que ninguno de los máximos estaba en la sílaba tónica.

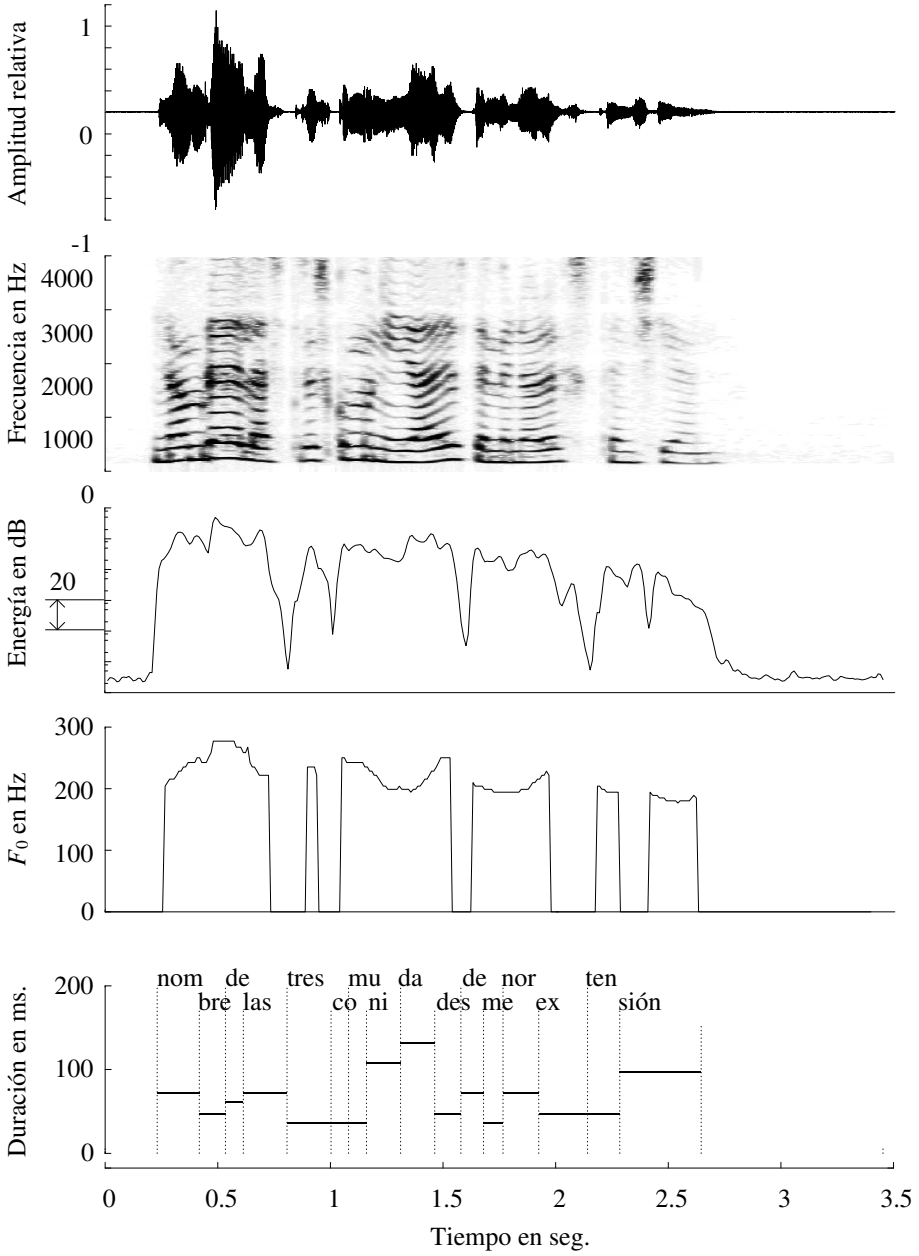

Figura 3.6. De arriba hacia abajo: señal de voz, espectrograma, energía, frecuencia fundamental y duración del núcleo vocálico en la frase *Nombre de las tres comunidades de menor extensión.*

Sílaba →	1	2	3	4	Promedio %
Máx. Energía	56.55	44.94	25.84	71.38	49.68
Máx. F_0	42.23	25.44	27.41	20.78	28.97
Máx. Duración	70.94	30.35	62.14	53.01	54.11

Tabla 3.5. Porcentajes de coincidencia entre los máximos prosódicos y la acentuación. En esta tabla se discrimina según la posición del acento.

Para proveer un análisis más detallado, en la Tabla 3.5 se muestran los aciertos de cada rasgo prosódico por sílaba. Este porcentaje se calculó haciendo la cantidad de aciertos en cada sílaba relativa al total de palabras que poseen acento en la sílaba correspondiente. En esta misma tabla se encuentra una tendencia promedio que nos indica lo representativo del máximo de cada rasgo prosódico en su relación con la acentuación.

Es interesante conocer también qué sucede cuando la posición de la sílaba tónica es considerada según las formas más clásicas: oxítonas, paroxítonas y proparoxítonas (tratadas en la Sección 1.3.3). En las Tablas 3.6 a 3.8 se detallan los porcentajes según esta forma de contar la posición de la sílaba tónica. En particular se puede observar un aumento importante de las correlaciones para los máximos de F_0 en las palabras oxítonas.

Sílaba →	1	2	3	4	Promedio %
Máx. Energía	—	59.51	26.34	81.87	55.91
Máx. F_0	—	47.36	43.01	39.18	43.19
Máx. Duración	—	48.98	70.96	61.98	60.65

Tabla 3.6. Porcentajes de coincidencia entre los máximos prosódicos y la acentuación en palabras oxítonas.

Sílaba →	1	2	3	4	Promedio %
Máx. Energía	57.67	27.31	25.38	61.81	43.04
Máx. F_0	42.67	12.68	12.69	0.69	17.19
Máx. Duración	71.96	27.32	53.81	40.97	48.52

Tabla 3.7. Porcentajes de coincidencia entre los máximos prosódicos y la acentuación en palabras paroxítonas.

Sílaba →	1	2	3	4	Promedio %
Máx. Energía	21.74	45.00	—	—	33.37
Máx. F_0	28.26	12.72	—	—	20.49
Máx. Duración	39.13	12.27	—	—	25.70

Tabla 3.8. Porcentajes de coincidencia entre los máximos prosódicos y la acentuación en palabras proparoxítonas.

3.3.3. Mínimos prosódicos

Resulta de interés saber qué sucede con los mínimos. Se realizaron todos los análisis anteriores considerando los mínimos de energía, F_0 y duración, y todas las combinaciones posibles entre máximos y mínimos. Se confirmó así que los máximos de energía y duración caracterizan a la sílaba tónica. Sin embargo se encontró una correlación mucho más alta entre los mínimos de F_0 y las sílabas tónicas. Esto se puede observar en las Tablas 3.9 a 3.12.

De la misma forma que antes (Tabla 3.5) se consideraron los aciertos por sílaba tónica tomando como referencia al mínimo de F_0. En este caso el promedio para las coincidencias del mínimo de F_0 asciende al 36.76 % (contra los 28.97 % para el máximo de F_0). Los resultados completos se muestran en la Tabla 3.13.

mín. Ener.	mín. F_0	mín. Dur.	Coincidencias %
			34.67
		○	7.42
	○		19.91
	○	○	7.25
○			6.76
○		○	3.20
○	○		12.27
○	○	○	8.47

Tabla 3.9. Porcentaje de coincidencias entre los mínimos de energía, frecuencia fundamental y duración con la acentuación.

mín. Ener.	Máx. F_0	Máx. Dur.	Coincidencias %
			21.62
		○	23.47
	○		7.74
	○	○	16.42
○			9.24
○		○	11.82
○	○		3.20
○	○	○	6.45

Tabla 3.10. Porcentaje de coincidencias entre los mínimos de energía y máximos de frecuencia fundamental y duración con la acentuación.

Máx. Ener.	mín. F_0	Máx. Dur.	Coincidencias %
			11.61
		○	11.82
	○		10.71
	○	○	14.40
○			12.07
○		○	16.56
○	○		7.43
○	○	○	15.38

Tabla 3.11. Porcentaje de coincidencias entre los máximos de energía, mínimos de frecuencia fundamental y máximos de duración con la acentuación.

Máx. Ener.	Máx. F_0	mín. Dur.	Coincidencias %
			25.95
		○	9.80
	○		8.92
	○	○	3.87
○			23.36
○		○	7.04
○	○		15.38
○	○	○	5.65

Tabla 3.12. Porcentaje de coincidencias entre los máximos de energía y frecuencia fundamental y los mínimos duración con la acentuación.

Sílaba →	1	2	3	4	Promedio %
Máx. Energía	56.55	44.94	25.84	71.38	49.68
mín. F_0	66.76	31.40	25.06	23.80	36.76
Máx. Duración	70.94	30.35	62.14	53.01	54.11

Tabla 3.13. Porcentajes de coincidencia entre los máximos y mínimos prosódicos y la acentuación. En esta tabla se discrimina según la posición del acento.

3.3.4. Influencia de las pausas y silencios

Existen importantes variaciones de los rasgos prosódicos cuando una palabra se encuentra antes o después de una pausa. En el caso de las frases que no poseen pausas importantes en el medio, las palabras que se afectan principalmente son la primera y la última. Para verificar la influencia de este efecto se realizaron todas las estadísticas anteriores eliminando de los recuentos a las palabras que se encontraban en los extremos de una frase. Así se analizaron 1984 palabras y se encontró que, en términos generales, los valores de aciertos en máximos no aumentaron significativamente. A continuación se muestras las dos tablas más importantes para este estudio (Tablas 3.14 y 3.15).

3.3.5. Procesamientos alternativos de la curva de entonación

Debido a la correlación tan baja entre la F_0 y la acentuación se hizo otro conjunto de pruebas donde se estudiaron diferentes técnicas de procesamientos a la curva de F_0. En primer lugar se utilizó un ajuste de la curva de F_0 mediante polinomios de orden 6. Los coeficientes para estos polinomios fueron calculados en base al método de cuadrados mínimos generalizado, resuelto por descomposición en valores singulares [Press et al., 1997, Sec. 15.4]. La curva de ajuste resultante posee la forma que marca la tendencia de la entonación que en la frase tiene fundamentalmente una función distintiva (interrogaciones, afirmaciones, exclamaciones, etc.)[1]. Esta curva de ajuste fue restada a la curva original y así se obtuvo la *diferencia de entonación por ajuste* (difF_0). En la Figura 3.7 se puede observar este rasgo prosódico junto con la curva de F_0, el polinomio de interpolación y la segmentación silábica. Los resultados se resumen en las Tablas 3.16 a 3.18.

[1]Se ensayaron polinomios desde orden 2 hasta 25. Como era previsible, los de orden demasiado bajo no respetaban la forma general de la curva y los de orden muy alto poseían problemas de estabilidad en algunas frases. Se eligió finalmente los de orden 6 ya que seguían adecuadamente la función distintiva de la curva de F_0.

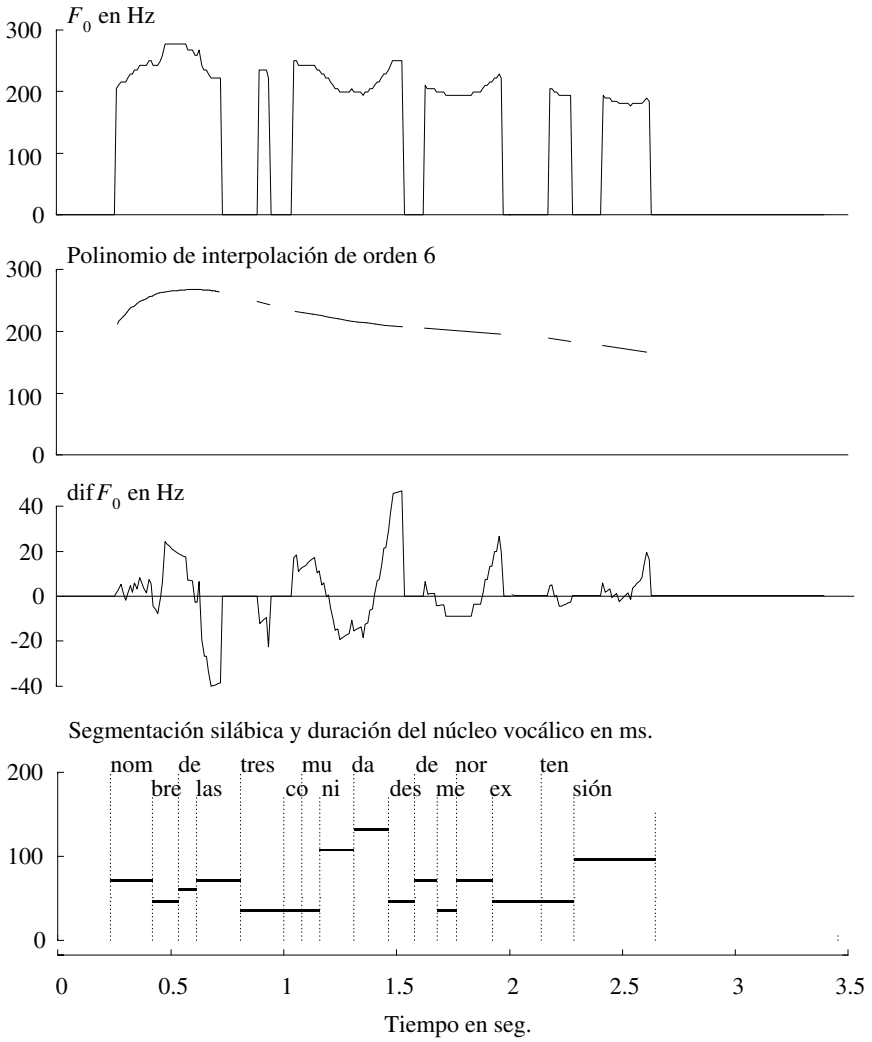

Figura 3.7. Diferencia de entonación por ajuste (difF_0) para la misma frase de la figura anterior: *Nombre de las tres comunidades de menor extensión.*

Máx. Ener.	Máx. F_0	Máx. Dur.	Coincidencias %
			16.03
		○	13.77
	○		5.78
	○	○	10.72
○			13.30
○		○	17.66
○	○		4.73
○	○	○	17.98

Tabla 3.14. Porcentaje de coincidencias entre los máximos de energía, frecuencia fundamental y duración con la acentuación. En este estudio no se consideraron la primera y última palabra de cada frase (contrástese con la Tabla 3.4).

Sílaba →	1	2	3	4	Promedio %
Máx. Energía	55.79	47.75	28.77	83.79	54.03
Máx. F_0	51.32	28.60	28.77	23.71	31.10
Máx. Duración	74.17	31.67	65.26	49.80	44.02

Tabla 3.15. Porcentajes de coincidencia entre los máximos prosódicos y la acentuación. En esta tabla se discrimina según la posición del acento y no se consideran la primera y última palabra de cada frase (contrástese con la Tabla 3.5).

Continuando con este estudio más detallado de la entonación se realizaron análisis de tendencias utilizando como rasgo prosódico representativo a la pendiente de una recta de ajuste para la F_0 en la sílaba de interés. De esta forma pueden distinguirse tres grandes grupos descritos en la Sección 1.3.3 (página 37): las cadencias de F_0, cuya pendiente es negativa; las mesetas de F_0, cuya pendiente se encuentra en un entorno cercano a cero y las anticadencias de F_0, que poseen pendiente positiva. En la Figura 3.8 se muestran las pendientes de las rectas de interpolación para la F_0 en los núcleos vocálicos de las palabras multisilábicas.

Las cadencias de F_0 se analizaron tanto para la curva de entonación como para la curva de diferencia de entonación por ajuste (esta última sin resultados de mayor relevancia). Las Tablas 3.19 a 3.22 muestran estos resultados pudiéndose observar que la asociación entre la F_0 con pendiente positiva (anticadencia) y la acentuación es entre un 15 y un 20 % más acertada que el máximo de F_0, superando así también los análisis realizados en torno a los mínimos de F_0.

Máx. Ener.	Máx. difF_0	Máx. Dur.	Coincidencias %
			17.06
		○	18.14
	○		5.27
	○	○	8.09
○			14.19
○		○	19.36
○	○		5.30
○	○	○	12.59

Tabla 3.16. Porcentaje de coincidencias entre los máximos de energía, máximo de diferencia de entonación por ajuste y duración con la acentuación.

Máx. Ener.	mín. difF_0	Máx. Dur.	Coincidencias %
			12.45
		○	11.89
	○		9.87
	○	○	14.34
○			10.95
○		○	15.45
○	○		8.55
○	○	○	16.50

Tabla 3.17. Porcentaje de coincidencias entre los máximos de energía, mínimos de diferencia de entonación por ajuste y máximos de duración con la acentuación.

Sílaba →	1	2	3	4	Promedio %
Máx. difF_0	33.65	27.38	29.50	30.42	30.24
mín. difF_0	71.22	32.74	20.89	17.47	35.58

Tabla 3.18. Porcentajes de coincidencia entre los máximos y mínimos prosódicos de diferencia de entonación por ajuste y la acentuación. En esta tabla se discrimina según la posición del acento. No se presentan los valores correspondientes a los máximos de energía y duración porque son los mismos que en la Tabla 3.5.

Máx. Ener.	Cad. F_0	Máx. Dur.	Coincidencias %
			15.14
		○	19.11
	○		7.18
	○	○	7.11
○			12.59
○		○	21.69
○	○		6.91
○	○	○	10.25

Tabla 3.19. Porcentaje de coincidencias entre los máximos de energía, cadencias de F_0 y máximos de duración con la acentuación.

Máx. Ener.	Mes. F_0	Máx. Dur.	Coincidencias %
			13.32
		○	17.51
	○		8.99
	○	○	8.72
○			11.79
○		○	20.33
○	○		7.71
○	○	○	11.61

Tabla 3.20. Porcentaje de coincidencias entre los máximos de energía, mesetas de F_0 y máximos de duración con la acentuación.

Máx. Ener.	Anti-cad.F_0	Máx. Dur.	Coincidencias %
			13.25
		○	10.99
	○		9.07
	○	○	15.24
○			11.30
○		○	13.04
○	○		8.20
○	○	○	18.91

Tabla 3.21. Porcentaje de coincidencias entre los máximos de energía, anticadencias de F_0 y máximos de duración con la acentuación.

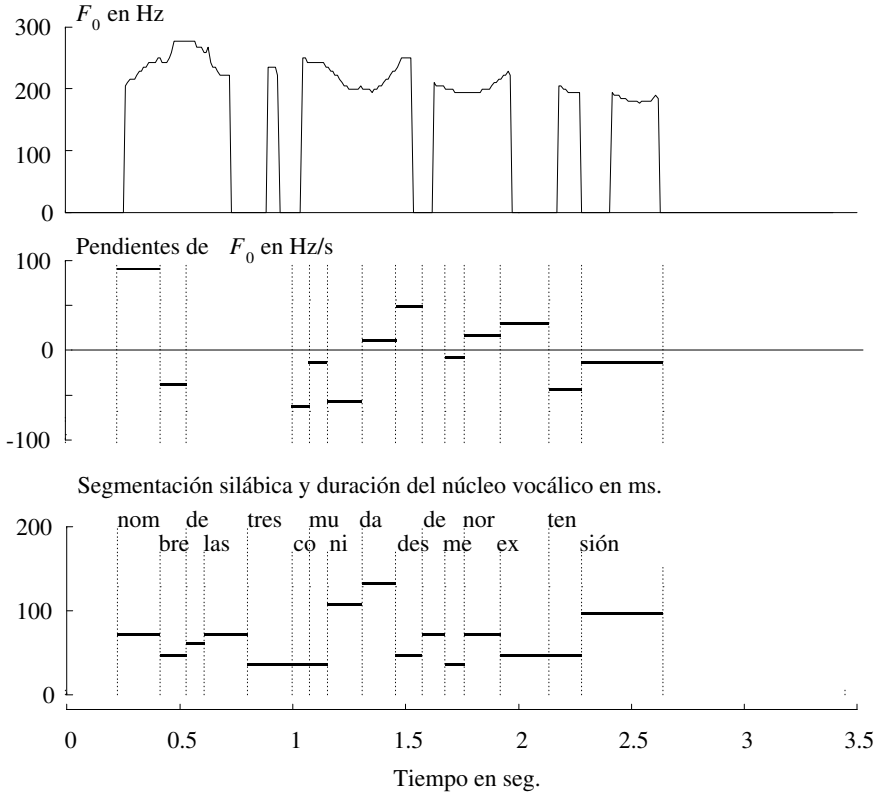

Figura 3.8. Pendientes de F_0 para la misma frase de la figura anterior: *Nombre de las tres comunidades de menor extensión.*

3.3.6. Variaciones en el núcleo vocálico

Se realizó otra serie de pruebas orientadas al estudio de la forma en que los rasgos prosódicos se ven modificados dependiendo de la vocal que forma el núcleo de la sílaba y su relación con la acentuación. En las Tablas 3.23 y 3.24 se muestran los resultados de este estudio.

Para visualizar mejor la forma en que los rasgos prosódicos varían entre las sílabas tónicas y las átonas se presentan a continuación los promedios en forma de gráficas (Figuras 3.9 a 3.11). A pesar de que ciertas tendencias se encuentran bien marcadas en estas estadísticas, debe considerarse que las desviaciones de la media son muy altas.

En todos los casos las vocales acentuadas poseen una duración promedio mayor, pero hay que tener en cuenta que las desviaciones estándar son altas. En el caso de la F_0 promedio se encuentra que en cuatro de las cinco vocales

Sílaba →	1	2	3	4	Promedio %
Cadencia	43.71	27.53	14.36	4.52	22.53
Meseta	48.11	34.52	20.10	12.35	28.77
Anticadencia	55.41	44.79	50.39	48.19	49.69

Tabla 3.22. Porcentajes de coincidencia entre cadencias, mesetas y anticadencias de F_0 y la acentuación. En esta tabla se discrimina según la posición del acento. No se presentan los valores correspondientes a los máximos de energía y duración porque son los mismos que en la Tabla 3.5.

Medidas	/a/	/e/	/i/	/o/	/u/
μ_E	0,63	0,54	0,24	0,50	0,31
σ_E	0,73	0,77	0,68	0,75	0,72
μ_{F_0}	187,99	197,09	173,32	191,13	184,80
σ_{F_0}	62,87	68,42	64,72	76,50	66,29
μ_D	65,84	54,01	62,74	60,62	47,36
σ_D	31,20	27,35	22,67	34,30	17,78

Tabla 3.23. Valores medios (μ) y desviación estándar (σ) para la energía (E) normalizada con el máximo en la palabra, la frecuencia fundamental (F_0 en Hz.) y la duración (D en ms.), del núcleo vocálico en sílabas átonas.

Medidas	/á/	/é/	/í/	/ó/	/ú/
μ_E	1,00	0,51	0,52	0,70	0,30
σ_E	0,71	0,68	0,78	0,74	0,68
μ_{F_0}	82.15	71.60	95.63	64.81	60.75
σ_{F_0}	287.10	166.51	134.14	197.41	90.34
μ_D	49.17	56.43	66.85	53.63	74.09
σ_D	30.27	32.23	32.08	24.93	28.29

Tabla 3.24. Valores medios (μ) y desviación estándar (σ) para la energía (E) normalizada con el máximo en la palabra, la frecuencia fundamental (F_0 en Hz.) y la duración (D en ms.), del núcleo vocálico en sílabas tónicas.

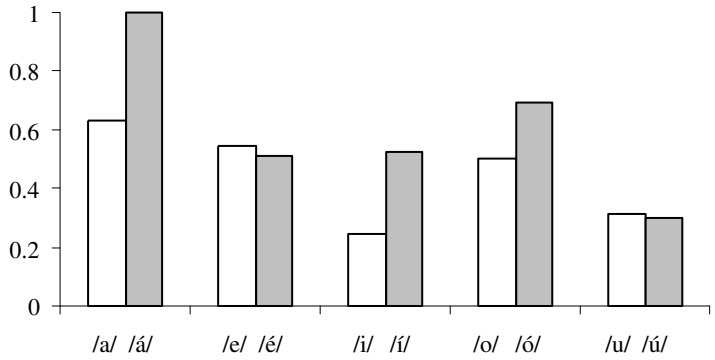

Figura 3.9. Valores medios de energía para los 5 núcleos vocálicos acentuados y no acentuados. Para simplificar el gráfico se han utilizado valores de energía relativos al máximo promedio encontrado.

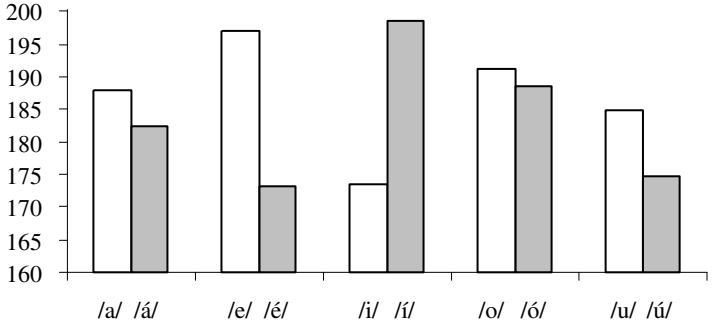

Figura 3.10. Valores medios de frecuencia fundamental (F_0 en Hz.) para los 5 núcleos vocálicos acentuados y no acentuados.

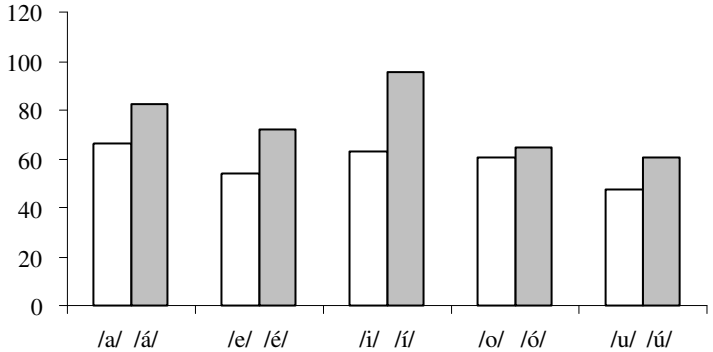

Figura 3.11. Valores medios de duración (en ms.) para los 5 núcleos vocálicos acentuados y no acentuados.

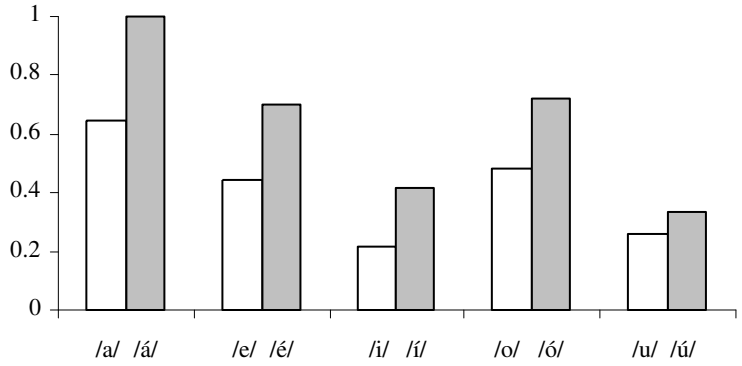

Figura 3.12. Valores medios de energía normalizados por palabra, para los 5 núcleos vocálicos acentuados y no acentuados. Para simplificar el gráfico se han utilizado valores de energía relativos al máximo promedio encontrado.

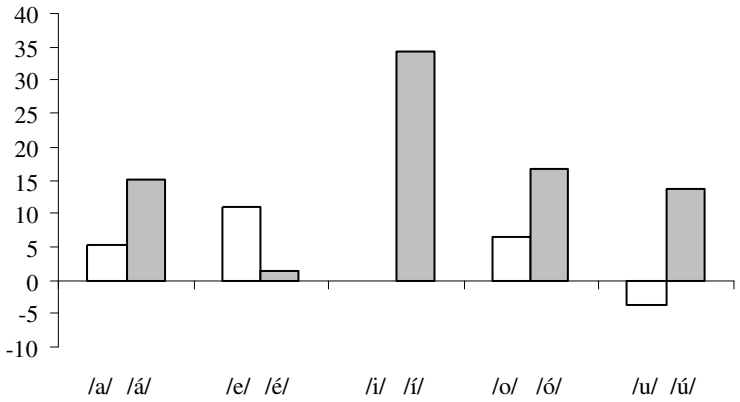

Figura 3.13. Valores medios de las pendientes de F_0 para los 5 núcleos vocálicos acentuados y no acentuados.

es mayor para la sílaba átona. En la energía se puede observar que en dos de los casos la vocal no acentuada es mayor.

Dado que la energía varía considerablemente a lo largo de toda una frase y aún más en las distintas frases, se consideró oportuna una normalización al nivel de palabras. En la Figura 3.12 se observan estos resultados. Ahora, aunque las desviaciones siguen siendo altas, se observa que los promedios de las sílabas tónicas siempre son superiores a los de las sílabas átonas. Finalmente se puede observar este mismo análisis para la cadencia de F_0 en la Figura 3.13. En este caso solamente la /e/ posee la característica de

cadencia (promedio) en la sílaba tónica. Las restantes sílabas tónicas son bien caracterizadas por la anticadencia (promedio).

3.4. Resumen de resultados y discusión

A continuación se presenta en dos tablas un resumen de los resultados más importantes de las secciones anteriores. En la Tabla 3.25 se pueden apreciar, para cada sílaba, las coincidencias entre los máximos de los tres rasgos prosódicos más clásicos y la acentuación. En la Tabla 3.26 se pueden observar las coincidencias para los tres mejores procesamientos de la F_0.

A pesar de que las coincidencias no son definitorias, parece posible imaginar un sistema relativamente simple que pueda obtener las EA de una frase a partir solamente de la emisión sonora. Sin embargo, ésta no es una tarea tan simple. En primer lugar hay que tener en cuenta que para obtener todos estos recuentos se ha tomado, como punto de partida, una buena segmentación de las frases. Para obtener esta segmentación se utilizó un MOM entrenado específicamente para las frases del corpus de habla analizado. Además, durante la obtención de la mejor secuencia por el algoritmo de Viterbi se utilizó la transcripción completa de cada una de las frases. Este es un punto central ya que todos los recuentos se basan en esta segmentación. ¿Qué sucedería si no se contase con las transcripciones de cada frase? ¿Qué sucedería si el MOM no estuviera especialmente adaptado a las frases que se analizaron? Indudablemente las relaciones entre los rasgos prosódicos y la acentuación que con mucho esfuerzo se extrajeron de las frases, quedarían casi totalmente ocultas. Y, si las relaciones que se poseen actualmente ya no son definitorias en cuanto a la determinación de la acentuación, prácticamente no sería posible extraer ninguna información útil acerca de la acentuación si no se posee la transcripción correcta y una buena segmentación para analizar los rasgos prosódicos.

Una de las conclusiones de este estudio, presagiada por otros, es que en el discurso continuo la correspondencia entre acentuación y rasgos prosódicos se pierde en un grado considerable. Queda claro que la tarea de extraer la acentuación a partir de los rasgos prosódicos no se puede realizar a partir de unas reglas sencillas y la señal de voz. En el siguiente capítulo se describirán diversas técnicas orientadas a encontrar un sistema automático que pueda extraer EA a partir la señal de voz.

Posición del acento	→	1	2	3	4
Máximo	1	**837**	265	**184**	39
de energía	2	618	**302**	32	29
	3	25	75	99	14
	4	0	30	68	**237**
Máximo	1	625	**292**	**157**	**146**
de F_0	2	**840**	171	25	44
	3	15	111	105	4
	4	0	98	96	69
Máximo	1	**1050**	**270**	69	3
de duración	2	422	204	14	21
	3	8	114	**238**	97
	4	0	84	62	**176**

Tabla 3.25. Matriz de confusión que indica la cantidad de palabras en las que coincide un determinado rasgo prosódico con la posición del acento. En esta tabla se analizan los máximos para energía, frecuencia fundamental y duración. (No se han representado los máximos más allá de la cuarta sílaba para simplificar la tabla.)

Posición del acento	→	1	2	3	4
Mínimo	1	**988**	**283**	92	94
de F_0	2	478	211	**154**	39
	3	14	111	96	75
	4	0	67	41	**79**
Anticadencia	1	**849**	160	66	88
de F_0	2	621	**304**	74	22
	3	10	141	**177**	20
	4	0	67	66	**150**
Cadencia	1	620	**360**	**157**	76
de F_0	2	**836**	176	116	**118**
	3	24	78	59	79
	4	0	57	51	16

Tabla 3.26. Matriz de confusión que indica la cantidad de palabras en las que coincide un determinado procesamiento en la F_0 con la posición del acento. (Al igual que en la tabla anterior, no se han representado las coincidencias más allá de la cuarta sílaba para simplificar la tabla.)

Capítulo 4

Estimación de estructuras acentuales

En este capítulo se describirán diversos métodos para estimar de forma automática las estructuras acentuales a partir de la señal de voz. Como se pudo ver en el Capítulo 3, no existe un método directo para obtener la acentuación de las palabras de una frase a partir de los rasgos prosódicos. Debido a que en el discurso continuo se pierde de manera significativa la relación entre acentuación y rasgos prosódicos, es necesario utilizar técnicas más sofisticas que puedan extraer relaciones complejas entre los datos. En este sentido es necesario cubrir dos aspectos importantes del problema: las características locales de los segmentos de voz y sus dinámicas a lo largo de una frase. En la primera parte del capítulo se describirán varias técnicas que permiten encontrar las estructuras acentuales a partir de los rasgos prosódicos, en base a una segmentación silábica conocida. En la segunda parte del capítulo se atacará el problema de la segmentación *ciega* de la voz, es decir, un método de segmentación que solamente utiliza la señal de voz. En la última parte del capítulo se describe un método que ataca los dos aspectos del problema en forma conjunta. Este método, basado en modelos ocultos de Markov, realiza la segmentación y clasificación de estructuras acentuales simultáneamente. En el próximo capítulo, estas estructuras acentuales estimadas servirán para mejorar el rendimiento de un sistema de reconocimiento automático del habla.

4.1. Clasificación con segmentación conocida

Así como en el Capítulo 3 se realizaron los estudios de prosodia y acentuación a partir de una segmentación conocida, aquí se utilizará de la misma forma una segmentación conocida para extraer la información prosódica de cada sílaba en una palabra y entrenar sistemas que pueda clasificar las estructuras acentuales (EA).

4.1.1. Clasificación de patrones

Los *árboles de decisión* (AD) y las *redes neuronales artificiales* (RNA) son dos técnicas ampliamente utilizadas para la clasificación de patrones. Los AD generan un conjunto de particiones en el espacio de entrada basándose en una estructura jerárquica de nodos en los que se realizan comparaciones sobre alguna componente del vector de características. Las redes neuronales están formadas por un conjunto de unidades de procesamiento no lineal altamente interconectadas, que procesan en paralelo un conjunto de datos para extraer información. Existen diferentes modelos neuronales para la implementación de clasificadores supervisados y no supervisados. El perceptrón multicapa es un ejemplo clásico de clasificador simple supervisado, mientras que los *mapas autoorganizativos* (MAO) son ejemplos de clasificadores simples no supervisados [Bishop, 1995, Kohonen, 1995].

Mapas autoorganizativos

Diversas áreas del cerebro, especialmente de la corteza cerebral, se hallan organizadas según diferentes modalidades sensoriales. Esta organización de la actividad cortical del cerebro puede describirse mediante mapas ordenados. Por ejemplo, se encuentran los mapas retinoscópicos de la corteza visual, los mapas tonotópicos de la corteza auditiva, los mapas somatotópicos de la corteza somatosensorial y los mapas de retardo interaural. Inspirado en el mapeo ordenado del cerebro, Kohonen introdujo en 1982 un algoritmo de autoorganización para producir mapas ordenados que simulan cortezas biológicas simplificadas, con el objeto de resolver problemas prácticos de clasificación y reconocimiento de patrones [Kohonen et al., 1984]. Los MAO presentan la propiedad de preservación de la vecindad, que los distingue de otros paradigmas de RNA. Estás arquitecturas son entrenadas mediante aprendizaje competitivo, es decir, las neuronas compiten entre ellas para ser activadas, dando como resultado la activación de una sola a la vez. Esta neurona es llamada *neurona ganadora* y a diferencia de otras RNA donde sólo

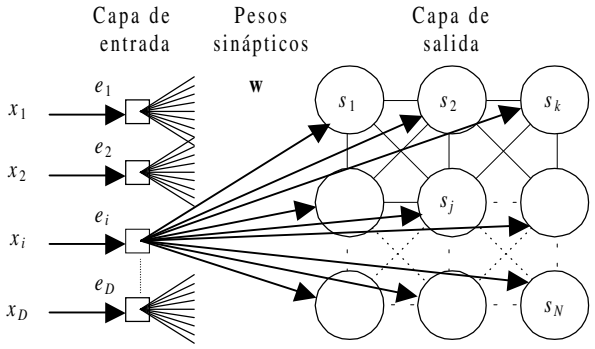

Figura 4.1. Configuración de las neuronas en un mapa autoorganizativo. x_i: patrón de entrada; e_i: neuronas de entrada; s_j: neuronas de salida; $w_{i,j}$ pesos sinápticos.

se permite que aprenda la unidad ganadora, en los MAO todas las unidades vecinas a la ganadora reciben una realimentación procedente de la misma, participando de esta manera en el proceso de aprendizaje. Esta importante característica es también denominada realimentación lateral y puede ser excitatoria, inhibitoria o una combinación de ambas.

En la Figura 4.1 se puede ver la configuración básica de un MAO. Se observan las neuronas de entrada e_i y una red bidimensional de neuronas de salida s_j. Un peso sináptico $w_{i,j}$ conecta a la neurona e_i con la s_j. A cada neurona de entrada e_i se le presenta el i-ésimo elemento de cada patrón de entrada $\mathbf{x}(n) \in \mathbb{R}^D$, siendo n la ocurrencia temporal de este patrón. El arreglo bidimensional de neuronas de salida incluye conexiones entre las neuronas vecinas simulando la realimentación lateral.

Si G es una neurona ganadora durante el entrenamiento de un MAO, las neuronas vecinas que también serán actualizadas quedan en una región determinada por una función de vecindad $\Lambda_G(n)$. Esta región puede tener diferentes formas y es variable con n. El área cubierta comienza siendo máxima y se reduce a medida que avanza el entrenamiento hasta no incluir ninguna neurona vecina a la ganadora.

En la Figura 4.2 se describe el algoritmo de entrenamiento de un MAO. Tanto la velocidad de aprendizaje $\eta(n)$ como el función de vecindad $\Lambda_G(n)$ varían durante el entrenamiento, aunque no existe una base teórica para seleccionarlas. Básicamente se decrementan según dos etapas de aprendizaje: la etapa de ordenamiento topológico y la de convergencia. Una vez que se ha entrenado un MAO, los vectores de pesos $\mathbf{w}_j$, que van desde la salida s_j a todas las entradas, determinan los denominados *centroides* de cada clase. Para más detalles véase [Kohonen, 1990].

`Comienzo`: se asignan valores aleatorios en $[-0{,}5; 0{,}5]$ para los vectores de pesos.
`Repetir`
`Muestreo`: se presenta un patrón de entrada $\mathbf{x}(n)$ elegido de forma aleatoria.
`Prueba de similitud`: se encuentra la neurona ganadora:

$$G(n) = \arg \min_{j=1}^{N} \|\mathbf{x}(n) - \mathbf{w}_j(n)\|$$

`Adaptación`: se ajustan los vectores de pesos sinápticos para la neurona ganadora y sus vecinas:

$$\mathbf{w}_j(n+1) = \begin{cases} \mathbf{w}_j(n) + \eta(n)[\mathbf{x}(n) - \mathbf{w}_j(n)] & \text{si } s_j \in \Lambda_G(n) \\ \mathbf{w}_j(n) & \text{en otro caso} \end{cases}$$

`Hasta` no observar cambios en el mapa de características.

Figura 4.2. Algoritmo de entrenamiento para un mapa autoorganizativo.

Cuantización vectorial con aprendizaje

La cuantización vectorial surge originalmente como un método de compresión, pero también puede ser interpretada como un proceso de clasificación. En la cuantización vectorial se intenta extraer la estructura subyacente a un grupo de patrones para dividir el espacio de entrada en un número finito de regiones y asociar a cada una de ellas un vector característico o centroide. Cada uno de estos centroides está asociado a una etiqueta o número de índice y de esta forma se *cuantiza* la información contenida en los vectores de entrada. En particular, la cuantización vectorial con aprendizaje (CVA) es una técnica que se puede utilizar para ajustar la posición de los centroides y mejorar el rendimiento de un clasificador en las fronteras de las regiones de decisión.

Existen diferentes versiones del algoritmo de CVA en base a una misma idea central. A partir de una apropiada configuración inicial, el algoritmo CVA1 consiste simplemente en acercar o alejar un centroide al patrón de entrada de acuerdo a si fue bien o mal clasificado, respectivamente. El algoritmo completo se describe en la Figura 4.3.

Una optimización para este algoritmo consiste en la adecuada selec-

`Comienzo`: los N_C centroides se hacen igual a los primeros patrones de entrenamiento de cada clase (existen diversos métodos de inicialización).
`Repetir`
`Muestreo`: se presenta un patrón de entrada $\mathbf{x}(n)$ de la clase $x^c(n)$ elegido de forma aleatoria.
`Prueba de similitud`: se clasifica de acuerdo a la mínima distancia euclídea:

$$c = \arg \min_{j=1}^{N_C} \|\mathbf{x}(n) - \mathbf{w}_j(n)\|$$

`Adaptación`: se ajusta el centroide más cercano de acuerdo a:

$$\mathbf{w}_c(n+1) = \mathbf{w}_c(n) + s(n)\eta(n)[\mathbf{x}(n) - \mathbf{w}_c(n)]$$

donde $0 < \eta(n) < 1$ y se define:

$$s(n) = \begin{cases} +1 & \text{si } x^c(n) = c \\ -1 & \text{si } x^c(n) \neq c \end{cases}$$

`Hasta` satisfacer algún criterio de convergencia o máximo de iteraciones (N_I).

Figura 4.3. Algoritmo de entrenamiento para la cuantización vectorial con aprendizaje.

ción de la función de variación para la velocidad de aprendizaje. Si se considera una velocidad de aprendizaje independiente para cada centroide, la ecuación de adaptación de los centroides se puede escribir como $\mathbf{w}_c(n+1) = [1 - s(n)\eta_c(n)]\,\mathbf{w}_c(n) + s(n)\eta_c(n)\mathbf{x}(n)$. En esta ecuación se observa que el valor que toma el centroide en $n+1$ depende del patrón de entrada en n (segundo término) y del antecedentes de todos los anteriores que se guarda en $\mathbf{w}_c(n)$ (primer término). Si la velocidad de aprendizaje es constante existe una diferencia en cómo se considera el patrón de entrenamiento actual y los anteriores. Si simplemente se observa un instante hacia atrás se puede ver que, mientras el patrón actual es pesado con la constante $\eta_c(n)$, el anterior es pesado con $[1 - s(n)\eta_c(n)]\,\eta_c(n-1)$.

Sin embargo, sería deseable que todos los patrones tuvieran la misma importancia en el valor final de $\mathbf{w}_c$. Para solucionar este problema se puede plantear que la velocidad de aprendizaje decrezca de forma que $\eta_c(n) = [1 - s(n)\eta_c(n)]\,\eta_c(n-1)$. De esta igualdad se puede obtener una regla para la variación óptima de la velocidad de aprendizaje:

$$\eta_c(n) = \frac{\eta_c(n-1)}{1 + s(n)\eta_c(n-1)}$$

El método resultante de optimizar la velocidad de aprendizaje se denominarse CVA1 optimizado (CVA1-O) [Kohonen, 1995].

Inducción de reglas mediante árboles de decisión

En este paradigma el algoritmo de aprendizaje busca una colección de reglas que clasifican "mejor" los ejemplos de entrenamiento y se puedan representar como un AD. Estas estructuras pueden pensarse como diagramas de flujo en donde cada nodo representa una prueba y cada rama que sale del nodo representa un resultado posible a dicha prueba. Para una revisión más detallada se puede consultar [Breiman et al., 1984].

Existen AD binarios y n-arios, de acuerdo a la cantidad de particiones realizadas en cada nodo. Dependiendo de las características de la función del nodo y del tamaño del árbol, la frontera final de decisión puede ser muy compleja. Una de las funciones más empleadas es la prueba mediante un cierto umbral para cada atributo, teniendo como resultado la partición del espacio de atributos por medio de hiperplanos paralelos u ortogonales a los ejes coordenados del espacio de atributos.

Dos de los algoritmos de aprendizaje más utilizados son ID3 y CART [Quinlan, 1993]. El algoritmo ID3 genera AD n-arios debido a que particiona el conjunto de ejemplos de entrenamiento en función del mejor atributo. La función heurística que utiliza ID3 para determinar el mejor atributo es una medida de la entropía para cada atributo. El algoritmo CART genera AD binarios ya que para particionar el conjunto de ejemplos en un nodo elige el mejor par atributo-valor de acuerdo con el denominado *criterio de Gini* [Sestito y Dillon, 1994].

Aunque los AD son intuitivamente atractivos y han tenido aplicaciones exitosas, existen algunos problemas que pueden obstaculizar su empleo en casos reales. Entre estos problemas se pueden mencionar la presencia de datos inconclusos, incompletos o ruidosos y el hecho de que raramente se aprovechan en simultáneo todos atributos de los vectores de entrada.

4.1.2. Árboles de redes neuronales autoorganizativas

Para solucionar algunos de los problemas que se presentan con los AD, una alternativa consiste en la implementación híbrida de AD y RNA. Este

tipo de enfoque permite aprovechar las ventajas de la clasificación jerárquica y crear fronteras de decisión más complejas con menos nodos, minimizando los problemas de ruido y de estructuras intrincadas. Los árboles de redes neuronales (ARN) son AD que implementan la tarea de decisión en los nodos mediante una red neuronal. De esta manera la decisión que se toma en cada nodo se basa en reglas más complejas, lo que permite aproximar mejor las fronteras a costa de perder claridad en la interpretación de las reglas resultantes.

La cantidad de particiones que se producen en cada nodo puede ser fija o variable. Cuando la cantidad de clases generadas puede variar para cada nodo, el ARN tiene la posibilidad de adoptar una configuración más adecuada para el problema a resolver. Los ARN realizan una clasificación basada en una combinación de los métodos de clasificación simple y jerárquica. Si se utiliza un MAO en cada uno de los nodos del ARN se aprovecha también el hecho de que estas redes de entrenamiento no supervisado pueden separar los patrones de acuerdo a su distribución natural.

El algoritmo propuesto permite que en las primeras capas o nodos se separen los grupos de patrones más alejados entre sí (o más fácilmente separables) y en las capas finales se haga una separación más fina de los patrones (es decir, los más difícilmente separables). En el caso de árboles n-arios, un problema importante es cómo decidir acerca de la cantidad de particiones a realizar en cada nodo. Para atacar este problema se establecieron criterios basados en los coeficientes de clasificación que se describen a continuación.

Coeficientes de clasificación

Dado un clasificador general se define el conjunto de patrones de entrada como $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_P\}$ con $\mathbf{x}_i \in \mathbb{R}^D$. Los patrones de X pueden ser agrupados en M clases de entrada C_i^I. Para el conjunto de clases de entrada $C^I = \left\{C_1^I, C_2^I, \ldots, C_M^I\right\}$ se cumplen las siguientes hipótesis:

$$X = C_1^I \cup C_2^I \cup \cdots \cup C_M^I \tag{4.1}$$

$$C_i^I \cap C_j^I = \emptyset \quad \forall i \neq j \tag{4.2}$$

$$C_i^I \neq \emptyset \quad \forall i \tag{4.3}$$

De la misma forma en que los patrones de entrada se agrupan según las clases a las que pertenecen realmente, también se pueden agrupar de acuerdo

a las clases C_j^O en que son separados por el clasificador. Estas últimas forman el conjunto de clases de salida $C^O = \{C_1^O, C_2^O, \ldots, C_N^O\}$ y cumplen con las siguientes hipótesis[1]:

$$X = C_1^O \cup C_2^O \cup \cdots \cup C_N^O \tag{4.4}$$
$$C_i^O \cap C_j^O = \emptyset \quad \forall i \neq j \tag{4.5}$$

Debe observarse que, en el caso más general, la cantidad de clases de entrada M no necesariamente debe ser igual a la cantidad de clases de salida N. Esta generalización resulta muy útil cuando el proceso de clasificación se realiza mediante clasificaciones sucesivas. En estas etapas intermedias en general $M \geq N$. No obstante, en el clasificador visto como un solo conjunto generalmente se tiene $M \leq N$.

Una definición importante para el desarrollo posterior es la *matriz de intersección de entrada-salida*:

$$N_{i,j}^{IO} = \left|\left(C_i^I \cap C_j^O\right)\right|; \quad 1 \leq i \leq M; \;\; 1 \leq j \leq N.$$

donde $|\cdot|$ es el operador de cardinalidad. Esta matriz contiene en su i,j-ésima celda la cantidad de patrones de la clase de entrada C_i^I clasificados como pertenecientes a la clase de salida C_j^O.

A continuación se analizan las limitaciones en la utilización del coeficiente de reconocimiento clásico como criterio para el desarrollo de la topología de un ARN.

Coeficiente de reconocimiento clásico

El coeficiente de reconocimiento que se utiliza generalmente para medir el rendimiento de un clasificador en el reconocimiento de patrones se puede definir según:

$$cr = \frac{\sum_{i=1}^{M} \max_{j=1}^{N} \left(N_{i,j}^{IO}\right)}{|X|} \tag{4.6}$$

[1] En este caso no se requiere $C_j^O \neq \emptyset \quad \forall j$.

siempre que se cumplan:

$$M = N \tag{4.7}$$

$$j_{i_1} \neq j_{i_2} \quad \forall i_1 \neq i_2; \quad 1 \leq i \leq M \tag{4.8}$$

siendo $j_i = \arg \max_{j=1}^{N} \left(N_{i,j}^{IO} \right)$.

Este coeficiente tiene algunas propiedades que suelen hacer confusa su interpretación. Si se cumplen las hipótesis (4.1 - 4.3), (4.4 - 4.5) y (4.7 - 4.8), el máximo que puede alcanzar cr es 1 cuando $\forall i \exists j / N_{i,j}^{IO} = \left| C_i^I \right|$. Vale aclarar que para cada i existe un único j por la restricción impuesta en la (4.8). Sin embargo, el mínimo que puede alcanzar cr no es cero ya que éste depende del número de clases de salida M. Cuando el clasificador se encuentra en un máximo de confusión, distribuye igualmente cada clase de entrada en las clases de salida. Por lo tanto el mínimo para cr es $1/M$, ya que el máximo en cualquier clase de salida es $\left| C_i^I \right| / M$. Esto es particularmente confuso ya que un clasificador con dos clases de salida no podría tener nunca un $cr \leq 0{,}5$ (rendimiento menor al 50 %).

Por otro lado, este coeficiente de reconocimiento no es aplicable cuando $M \neq N$. Además, (4.8) restringe su aplicabilidad cuando se hacen agrupaciones intermedias de varias clases de entrada en una clase de salida para ser luego separadas por otro clasificador. Cuando se relaja (4.8), el coeficiente no permite discernir en qué medida patrones de las misma clase de entrada son concentrados en la misma clase de salida y patrones de distintas clases de entrada son distribuidos en distintas clases de salida. Para poder eliminar (4.7) y (4.8), se definen dos coeficientes que miden estas concentraciones y dispersiones por separado.

Coeficiente de concentración interclase

Para medir en qué grado un clasificador agrupa patrones pertenecientes a una clase de entrada en una misma clase de salida se define, en primer lugar, el coeficiente de concentración interclase para la clase de entrada C_i^I en las N clases de salida C_j^O como:

$$cc_i = \frac{N \max_{j=1}^{N} \left(N_{i,j}^{IO} \right) - \sum_{j=1}^{N} N_{i,j}^{IO}}{(N-1) \sum_{j=1}^{N} N_{i,j}^{IO}} \tag{4.9}$$

donde $\sum_j N_{i,j}^{IO} = \left| C_i^I \right| \neq 0 \; \forall i$ por (4.3). El coeficiente cc_i posee las siguientes propiedades:

I) $cc_i = 1 \Leftrightarrow \exists j^* / N_{i,j^*}^{IO} = \left| C_i^I \right|$ (si $\exists j^*$ entonces es único por (4.5)),

II) $cc_i = 0 \Leftrightarrow N_{i,j_1}^{IO} = N_{i,j_2}^{IO} \quad \forall 1 \leq j_1, j_2 \leq N$,

III) $cc_i \in [0, 1] \;\; \forall i$ y

IV) cc_i es monótono decreciente con $\max_{j=1}^{N} \left(N_{i,j}^{IO} \right)$

Se define el coeficiente de concentración interclase para un clasificador como el promedio de los cc_i ponderados por la cantidad de patrones de la clase de entrada correspondiente:

$$cc = \frac{\sum_{i=1}^{M} \left| C_i^I \right| cc_i}{\sum_{i=1}^{M} \left| C_i^I \right|}$$

sustituyendo según la ecuación (4.9) y simplificando se obtiene:

$$cc = \frac{\sum_{i=1}^{M} N \max_{j=1}^{N} \left(N_{i,j}^{IO} \right) - \sum_{i=1}^{M} \sum_{j=1}^{N} N_{i,j}^{IO}}{(N-1) \sum_{i=1}^{M} \sum_{j=1}^{N} N_{i,j}^{IO}} \tag{4.10}$$

donde $\sum_i \sum_j N_{i,j}^{IO} = |X| \neq 0$ por (4.3).

Asumiendo las hipótesis (4.7) y (4.8), y a partir de (4.6) y (4.10) se puede deducir que $cr = cc(M-1)/M + 1/M$. Así, se puede ver que cuando los máximos de cada clase de entrada se encuentran en distintas clases de salida y la cantidad de clases de salida es igual a la de clases de entrada, el

coeficiente de reconocimiento puede expresarse como una versión escalada y desplazada del coeficiente de concentración intraclase.

Hay que destacar que, si bien el coeficiente de concentración mide la capacidad con que un clasificador agrupa patrones de una misma clase de entrada en una única clase de salida, no es capaz de detectar cuando todos los patrones de entrada son llevados a una misma clase de salida. Para esto se define a continuación el coeficiente de dispersión intraclase.

Coeficiente de dispersión intraclase

Para medir la capacidad que posee un clasificador para llevar patrones de distintas clases de entrada a distintas clases de salida se define, en primer lugar, el coeficiente de dispersión intraclase para la clase de salida C_j^O en las M clases de entrada C_i^I como:

$$cd_j = \begin{cases} \dfrac{M \max\limits_{i=1}^{M}\left(N_{i,j}^{IO}\right) - \sum\limits_{i=1}^{M} N_{i,j}^{IO}}{(M-1)\sum\limits_{i=1}^{M} N_{i,j}^{IO}} & \text{si } \left|C_j^O\right| \neq 0 \\ 0 & \text{si } \left|C_j^O\right| = 0 \end{cases} \tag{4.11}$$

donde $\sum_i N_{i,j}^{IO} = \left|C_j^O\right|$. Este coeficiente posee las siguientes propiedades:

I) $cd_j = 1 \Leftrightarrow \exists i^*/N_{i^*,j}^{IO} = 0 \;\; \forall i \neq i^*$ (si $\exists i^*$, es único por (4.2)),

II) $cd_j = 0 \Leftrightarrow N_{i_1,j}^{IO} = N_{i_2,j}^{IO} \;\; \forall 1 \leq i_1, i_2 \leq M$,

III) $cd_j \in [0,1] \;\; \forall j$ y

IV) cd_j es monótono creciente con $\max\limits_{i=1}^{N}\left(N_{i,j}^{IO}\right)$.

De forma similar que para el coeficiente de concentración, se define el coeficiente de dispersión intraclase para un clasificador como el promedio de los cd_j ponderados por la cantidad de patrones en cada clase de salida:

$$cd = \frac{\sum\limits_{j=1}^{N}\left|C_j^O\right| cd_j}{\sum\limits_{j=1}^{N}\left|C_j^O\right|}$$

Sustituyendo según (4.11) y simplificando se obtiene:

$$cd = \frac{\sum_{j=1}^{N} M \max_{i=1}^{M} \left(N_{i,j}^{IO} \right) - \sum_{j=1}^{j=N} \sum_{i=1}^{i=M} N_{i,j}^{IO}}{(M-1) \sum_{j=1}^{j=N} \sum_{i=1}^{i=M} N_{i,j}^{IO}} \tag{4.12}$$

donde nuevamente $\sum_j \sum_i N_{i,j}^{IO} = |X|$.

El coeficiente de dispersión intraclase no mide el grado en que patrones de entrada de una misma clase son derivados a diferentes clases de salida ya que esto es cuantificado por el coeficiente de concentración interclase.

Para ilustrar el comportamiento de los coeficientes definidos se muestran algunos ejemplos sencillos de aplicación.

1. Clasificador ideal:

$$N^{IO} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix} \rightarrow \begin{cases} cc = 1 \\ cd = 1 \\ cr = 1 \end{cases}$$

   En este caso los tres coeficientes de clasificación llegan a su valor máximo, indicando una clasificación perfecta.

2. Clasificador totalmente confundido:

$$N^{IO} = \begin{bmatrix} 10 & 10 & 10 \\ 10 & 10 & 10 \\ 10 & 10 & 10 \end{bmatrix} \rightarrow \begin{cases} cc = 0 \\ cd = 0 \\ cr = 1/3 \end{cases}$$

   El ejemplo muestra como cc y cd marcan más fehacientemente la deficiencia del clasificador.

3. Concentración sin dispersión:

$$N^{IO} = \begin{bmatrix} 0 & 0 & 10 \\ 0 & 0 & 10 \\ 0 & 0 & 10 \end{bmatrix} \rightarrow \begin{cases} cc = 1 \\ cd = 0 \\ cr = \text{no aplicable} \end{cases}$$

   La matriz N^{IO} corresponde a un clasificador que clasificó a todos los patrones de entrada como pertenecientes a una misma clase de salida. Este es un ejemplo típico de máxima concentración y mínima dispersión, como los coeficientes lo indican.

4. Máxima dispersión:

$$N^{IO} = \begin{bmatrix} 10 & 0 & 0 & 0 & 10 \\ 0 & 10 & 0 & 10 & 0 \\ 0 & 0 & 10 & 0 & 0 \end{bmatrix} \rightarrow \begin{cases} cc = 1/2 \\ cd = 1 \\ cr = \text{no aplicable} \end{cases}$$

Algoritmo de entrenamiento

El algoritmo de entrenamiento tiene por finalidad encontrar la estructura del AD y entrenar cada uno de los nodos clasificadores. La totalidad de los patrones de entrenamiento se presenta inicialmente al nodo que se encuentra en la raíz del árbol y a los nodos de los niveles siguientes les llega un subconjunto de patrones que ha sido derivado jerárquicamente de abajo (raíz) hacia arriba (hojas).

Considerando un nodo en particular se debe decidir, en primer lugar, si se justifica o no realizar una tarea de clasificación. Así se distingue entre dos tipos de nodos: nodos clasificadores y nodos terminales. Para declarar que un nodo es terminal o clasificador se deben tener en cuenta dos características de su conjunto de patrones de entrada: el grado de homogeneidad en clases y el número de patrones que posee. Si bien esta última característica no presenta ninguna dificultad en cuanto a su medición objetiva la medida de la homogeneidad en clases no es tan trivial. Por esta razón se define el coeficiente de concentración para el conjunto de patrones de entrada como:

$$pc = \frac{M \max\limits_{i=1}^{M} (|C_i|) - |X|}{(M-1)|X|} \tag{4.13}$$

del cual, en forma similar a *cc* y *cd*, se pueden enunciar las propiedades:

I) $pc = 1 \Leftrightarrow \exists i^* / |C_i| = 0 \quad \forall i \neq i^*$ (si $\exists i^*$, es único por (4.5))

II) $pc = 0 \Leftrightarrow |C_{i_1}| = |C_{i_2}| \quad \forall 1 \leq i_1, i_2 \leq M$

III) $pc \in [0, 1]$ y

IV) pc es monótono creciente con $\max\limits_{i=1}^{M} |C_i|$

Para determinar el tipo de nodo en base a las características mencionadas se comparan sus medidas con dos umbrales: el umbral de concentración

mínima de patrones de entrada (u_{pc}) y el umbral de cantidad mínima de patrones de entrada (u_X). Si se encuentra un nodo clasificador entonces debe entrenarse el MAO correspondiente. La dimensión de entrada en esta red está determinada por la dimensión de los patrones y es la misma para todo el árbol. La dimensión o cantidad de clases de salida junto con los nodos terminales definen la topología final del árbol.

Para determinar la cantidad apropiada de clases de salida se utiliza un proceso de crecimiento de nodo basado en los coeficientes *cc* y *cd* y dos umbrales de capacidad de clasificación mínima u_{cc} y u_{cd}. Se adopta inicialmente una configuración con dos clases de salida ($N = 2$), se entrena la red y se evalúa su rendimiento en la clasificación. En el caso en que no se supere alguno de los umbrales se incrementa N en uno y se repite el entrenamiento y prueba. Este proceso culmina cuando ambos coeficientes superan sus correspondientes umbrales o cuando N alcanza el máximo permitido N_{max}. En este último caso, se elige la mejor de todas las configuraciones entre 2 y N_{max} y se considera concluido el entrenamiento de ese nodo. Este algoritmo de crecimiento de nodo se repite para todos los nodos de cada nivel del árbol. En la Figura 4.4 se describe el algoritmo completo.

Las exigencias en cuanto a concentración y dispersión varían de acuerdo al nivel de profundidad en el proceso de clasificación. En las primeras etapas de la clasificación se propone una mayor exigencia en cuanto a la concentración. La separación basada en detalles más finos se realiza progresivamente en niveles posteriores, en los que se exige mejor dispersión en la clasificación. Así se pasa gradualmente desde la no supervisión a la supervisión y se logra progresivamente la concordancia entre las clases de salida y las de entrada.

Cuando el árbol ha sido entrenado se procede al etiquetado de los nodos terminales. La elección de la etiqueta asignada a cada nodo terminal se realiza en base al máximo de la matriz $N_{i,j}^{IO}$ del nodo clasificador que le dio origen. Luego, los nodos terminales se unen —de acuerdo a su etiqueta— en otro nivel de nodos artificiales que poseen las etiquetas de todas las clases. De esta forma en el ARN en su conjunto cumple $M = N$.

Funcionamiento del ARN entrenado

Para realizar la clasificación de un patrón se necesita propagarlo a través del ARN. La propagación del patrón puede realizarse en forma secuencial o en forma paralela. Cuando se propaga un patrón en forma secuencial se describe un camino a través del árbol mediante un simple algoritmo: se comienza por el nodo raíz, se miden las distancias del patrón a cada uno de los centroides del MAO correspondiente y se elige como nodo siguiente

```
Para cada nivel del árbol
  Para cada nodo del nivel
    NodoTerminal = (pc > u_pc) ∨ (|X| < u_X)
    Si ¬(NodoTerminal)
      N = N_min
      Mientras ¬(NodoEntrenado)
        Crear Nodo
        Entrenar Nodo
        Probar Nodo
        Si (N = N_max)
          NodoEntrenado = verdadero
          Buscar MejorN
          Si (MejorN <> N_max)
            Destruir Nodo
            N = MejorN
            Crear Nodo
            Entrenar Nodo
          Sino
            NodoEntrenado = (cc > u_cc) ∧ (cd > u_cd)
        Si ¬(NodoEntrenado)
          Destruir Nodo
          N = N + 1
      FinMientras
    FinPara
  Actualizar los umbrales
FinPara
```

Figura 4.4. Algoritmo de entrenamiento para un árbol de redes neuronales.

aquel indicado por el centroide que está más cerca del patrón. Los dos últimos pasos se repiten hasta que se llega a un nodo terminal y se clasifica al patrón según la etiqueta de este último nodo.

En la propagación paralela se miden las distancias entre el patrón y todos los centroides del ARN simultáneamente y luego se sigue el camino formado por los nodos activados a partir del nodo raíz, hasta llegar a un nodo terminal. En [Milone et al., 1998a] se pueden encontrar más detalles acerca de los ARN y un conjunto de experimentos con baterías de prueba de dominio público. Todos estos experimentos se contrastan con otros cla-

sificadores mostrando las ventajas del método. En [Milone et al., 1998b] se presentan pruebas para el reconocimiento de fonemas.

4.1.3. Resultados

Para generar los patrones de entrenamiento y prueba se utilizó un subconjunto de frases del corpus de habla Albayzin (que aquí denominamos SC2; para más detalles véase el Apéndice A.3). Con este subconjunto de 1000 frases se entrenó un sistema de reconocimiento automático del habla (RAH) basado en modelos ocultos de Markov (MOM) y con las mismas frases se realizó la segmentación buscando la secuencia más probable mediante el algoritmo de Viterbi (Sección 2.2.6). Los modelos del sistema de RAH fueron MOM semicontinuos, con 3 estados para los fonemas y el silencio y 1 estado para una pausa corta al final de cada palabra. Las características de la voz utilizadas en este sistema de RAH fueron coeficientes cepstrales en escala de mel (CCEM) con coeficientes de energía y delta (un total de 26 elementos). La ventana de análisis fue de 25 ms y el paso del análisis de 10 ms, con ventana de Hamming.

Para cada una de las frases se obtuvieron las EA y las curvas de energía, frecuencia fundamental (F_0) y duración del núcleo vocálico en cada sílaba. A partir de estas frases se generaron 6860 patrones de entrenamiento y 4570 para las pruebas de validación. Cada patrón de entrada se corresponde con una palabra y consiste en un vector con los valores de los rasgos prosódicos para cada una de las sílabas. Dado que los patrones de entrada deben tener dimensión fija, los elementos que están más allá de la cantidad de sílabas de la palabra se hacen cero. Como clase de salida se asigna un código que representa a la EA correcta de cada palabra. Como ejemplo de esta configuración, en la Tabla 4.1 se muestran los patrones de entrada y salida para algunas palabras suponiendo que se tomen como rasgos prosódicos los máximos de energía y F_0.

Resultados con CVA1-O

Debido a que los métodos CVA poseen una topología fija que es definida antes de comenzar el entrenamiento, se han evaluado diversas alternativas con el objetivo de encontrar la estructura con el número de centroides (N_C) más apropiado. Un parámetro que también debe considerarse en el entrenamiento es la cantidad de veces que se ajustan los centroides a partir de los patrones (N_I). La elección de estos parámetros de entrenamiento no es obvia, principalmente porque son muy dependientes de la estructura de los

ϵ_1	ϵ_2	ϵ_3	ϵ_4	ϵ_5	F_{01}	F_{02}	F_{03}	F_{04}	F_{05}	EA
1.00	0.92	0.00	0.00	0.00	0.85	1.00	0.00	0.00	0.00	/TA/
0.63	1.00	0.00	0.00	0.00	1.00	0.84	0.00	0.00	0.00	/TA/
0.88	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	/TA/
0.69	0.39	1.00	0.00	0.00	0.69	0.74	1.00	0.00	0.00	/ATA/
1.00	0.95	0.00	0.00	0.00	1.00	0.88	0.00	0.00	0.00	/AT/
1.00	0.31	0.12	0.00	0.00	0.97	1.00	1.00	0.00	0.00	/TAA/
1.00	0.92	0.00	0.00	0.00	1.00	0.97	0.00	0.00	0.00	/TA/
0.80	0.82	1.00	0.63	0.00	0.89	0.00	0.80	1.00	0.00	/AATA/

Tabla 4.1. Ejemplo de patrones de entrada con sus correspondientes clases de salida. Se tomaron 5 sílabas de 8 palabras consecutivas en los datos de entrenamiento. En las primeras 5 columnas se encuentra el valor de la energía, normalizado con el máximo en la palabra. En las siguientes 5 columnas se encuentran los valores de frecuencia fundamental para cada sílaba, también normalizados con el máximo en la palabra.

N_I	$\rightarrow$	100	500	1000	2000	4000	8000	10000
N_C	32	61.82	61.86	61.79	61.79	61.79	61.82	61.84
$\downarrow$	64	60.90	64.05	64.79	64.73	64.75	65.38	65.21
	128	66.81	67.57	67.64	68.40	68.60	69.10	69.65
	256	71.23	71.05	71.42	71.73	73.17	73.13	73.54
	512	72.84	72.08	71.36	72.47	73.15	73.41	73.85
	1024	72.74	71.29	72.14	72.10	72.84	73.17	73.22
	2048	74.25	74.00	73.11	73.00	74.07	74.46	74.49

Tabla 4.2. Resultados de clasificación de estructuras acentuales mediante cuantización vectorial con aprendizaje. En las columnas se muestran los resultados con diferentes números de iteraciones N_I en el entrenamiento. Las distintas filas indican la cantidad de centroides N_C utilizados en el clasificador.

patrones de entrada y la complejidad del problema de clasificación. Sin embargo, se pueden considerar algunos casos extremos como referencia. Por ejemplo, no parece apropiado tener tantos centroides como patrones de entrenamiento. En el otro extremo, salvo para problemas muy simples, no es suficiente con poseer tantos centroides como clases a discriminar. A partir de estos criterios empíricos se ha realizado la búsqueda de los parámetros óptimos y los resultados se muestran en la Tabla 4.2. Los porcentajes que se observan son el resultado de la utilización de los centroides obtenidos para la clasificación de los patrones del conjunto de prueba.

Máxima dimensión de salida	6
Número de iteraciones de entrenamiento de cada nodo	750
Umbral de concentración de patrones (u_{pc}) inicial	0.7
Umbral de concentración de patrones (u_{cc}) final	0.9
Umbral de capacidad de concentración (u_{cc}) inicial	0.3
Umbral de capacidad de concentración (u_{cc}) final	0.9
Umbral de capacidad de dispersión (u_{cd}) inicial	0.8
Umbral de capacidad de dispersión (u_{cc}) final	0.2
Resultado con energía y F_0 (cr)	85.65 %
Resultado con energía, F_0 y duración (cr)	89.98 %

Tabla 4.3. Resultados de clasificación de estructuras acentuales mediante árboles de redes neuronales. Se incluyen en esta tabla los parámetros con que fue entrenado el árbol de redes neuronales y los resultados de clasificación sobre el conjunto de prueba. Estos resultados se muestran para un entrenamiento con energía y frecuencia fundamental solamente y con los tres rasgos prosódicos juntos.

Resultados con ARN

En el caso de los ARN la topología se optimiza en el mismo algoritmo de entrenamiento (tanto la estructura interna de cada nodo como la del árbol en su conjunto). Sin embargo, como se vio anteriormente, existen algunos parámetros que regulan el crecimiento del árbol. La forma en que estos parámetros deben variar durante el crecimiento del árbol ha sido analizada y verificada experimentalmente en [Milone et al., 1998a]. Por lo tanto sólo se realizaron dos experimentos con diferente cantidad de iteraciones en el entrenamiento de cada nodo. Los resultados fueron muy similares debido a que en el ARN el resultado final no es tan dependiente del entrenamiento de los nodos como de la estructura de árbol generada. En la Tabla 4.3 se muestran los parámetros de configuración del algoritmo de entrenamiento y los resultados obtenidos con los datos de prueba. También se muestra en esta tabla el resultado para un experimento donde se incluyó la duración del núcleo vocálico en cada sílaba.

4.1.4. Discusión

Cuando se compara el ARN con el mejor caso de CVA1-O se encuentra una diferencia realmente importante a favor del ARN. Además, hay que considerar que habiendo 4570 patrones de prueba el número de 2048 centroides para el CVA1-O es algo excesivo. Si se considera que para el ARN se han

utilizado solamente 750 iteraciones en el entrenamiento, sería más razonable compararlo con clasificadores de la región central de la Tabla 4.2, donde las diferencias a favor del ARN son aún más significativas.

Una consideración muy importante a la hora de realizar comparaciones entre diferentes arquitecturas es el hecho de que mientras el ARN adapta su topología al problema en cuestión, otros métodos necesitan que se especifique una configuración inicial, generalmente basada en la experiencia del usuario y refinada mediante prueba y error. El ARN adapta su topología localmente a través de múltiples pruebas, como se desprende del algoritmo de crecimiento. Estas pruebas se realizan de manera jerárquica y automática en cada nodo lo que da lugar a un ahorro importante del costo computacional. Hay que destacar que el resultado del algoritmo de crecimiento no es sensible a los umbrales que deben fijarse de antemano. Como regla general, es suficiente con seguir simplemente los alineamientos de la Tabla 4.3 para asignar el inicio y el fin de cada umbral a lo largo de los niveles.

Dado que los cómputos realizados para la generación de un ARN son sencillos, esta arquitectura es considerablemente más veloz que otras estructuras neuronales. La forma jerárquica en que se organiza la información permite que la clasificación de cada patrón de prueba sea sustancialmente más rápida. No se necesitan más de 6 medidas de distancias por nivel[2], mientras que en el método de CVA se requieren tantas medidas de distancia como centroides existan.

La principal fuente de las ventajas de este método está en la combinación de diferentes paradigmas de clasificación. El algoritmo planteado combina las ventajas del aprendizaje no supervisado con las del aprendizaje supervisado. Por un lado, durante el crecimiento y definición de la topología del árbol se utiliza información acerca de la identidad de los patrones. En cambio, para la tarea de clasificación en cada nodo el MAO no usa información acerca de la identidad de los patrones de entrenamiento. Otra de las combinaciones de paradigmas de clasificación que se encuentran en este algoritmo es la de los clasificadores simples y los jerarquizados. Mientras que la estructura general responde a los métodos de clasificación jerarquizada, en cada nodo se utiliza un típico clasificador simple.

Finalmente, debe destacarse el hecho de que estos clasificadores son estáticos y no pueden modelar la información temporal contenida en la señal. De hecho, la segmentación silábica correcta siempre se ha supuesto conocida a priori. Pero la segmentación automática no es una tarea simple. En la siguiente sección se proponen nuevas técnicas para solucionar este problema.

[2]Para el ARN utilizado en los experimentos

4.2. El problema de la segmentación

La segmentación de la voz consiste en dividir una emisión en diferentes trozos de acuerdo con algún criterio. Es común que se segmente la voz para separarla en fonemas pero también suele ser de interés la segmentación según sílabas o unidades de nivel superior, como la palabra [Reddy, 1966, Svendsen y Soong, 1987, Hemert, 1991].

En el caso más simple, el problema de la segmentación de voz consiste en encontrar los límites precisos que definen a cada segmento o unidad fonética. Cada segmento presenta dos límites o marcadores que miden el tiempo, a partir del inicio de la emisión, en que se encuentran el principio y el final del segmento en cuestión. Una emisión puede tener muchos segmentos y así la ubicación correcta de todos sus límites puede ser un problema complejo. Más aún si se consideran todas las variaciones asociadas con los distintos lenguajes, como generalmente ocurre en los problemas relacionados con el habla.

Para la segmentación de voz se han utilizado varias técnicas. En primer lugar está la segmentación manual, en la que generalmente un experto lingüista genera la segmentación en base a espectrogramas, curvas de energía, entonación y otros estudios utilizados para el análisis de la voz. Esta técnica posee la ventaja de que la experiencia del lingüista asegura un muy buen resultado en la segmentación. Sin embargo los costos en tiempo y recursos que lleva este proceso manual son altísimos, lo que lo hace sólo aplicable a estudios muy especializados. La segunda técnica aplicable a la segmentación viene de la mano de los sistemas de RAH basados en MOM. Como se explicó anteriormente, se entrena un sistema de RAH convencional y mediante el algoritmo de Viterbi se puede obtener la secuencia más probable de estados que determina la segmentación. Sin embargo, para realizar esta operación es necesario contar con la transcripción correcta de la emisión de voz [Brugnara et al., 1993].

También existen otros métodos alternativos que no están necesariamente ligados con las técnicas del procesamiento de la voz sino que más bien son métodos de aplicación general. Entre ellos se puede mencionar a las RNA [Lee y Ching, 1999, Vorstermans et al., 1996, Jeong y Jeong, 1996], el modelado estadístico [Gallwitz et al., 1998, Pauws et al., 1996] y el filtrado paramétrico [Li y Gibson, 1996]. En cualquier caso el problema de la segmentación automática aún sigue sin ser resuelto totalmente y menos aún en aplicaciones de tiempo real.

4.2.1. Computación evolutiva

Los diferentes métodos de computación evolutiva han brindado en la última década una solución a muchos problemas, principalmente en la búsqueda y optimización de soluciones. Por ejemplo, se han aplicado con buenos resultados en la segmentación de imágenes [Bhandarkar y Zhang, 1999]. La analogía en que se basa la computación evolutiva estriba en reconocer el mecanismo esencial del proceso evolutivo en la naturaleza e imitarlo para el diseño y optimización de sistemas artificiales.

La computación evolutiva abarca un número cada vez mayor de métodos basados en la misma idea original. Entre muchos otros se destacan: los algoritmos genéticos [Goldberg, 1997], la programación genética [Koza, 1992] y la programación evolutiva [Michalewicz, 1992]. Una revisión y comparativa de éstos y otros métodos de computación evolutiva puede verse en [Bäck et al., 1997]. Los componentes fundamentales del mecanismo de la evolución biológica son los cromosomas —material genético de un individuo biológico—, donde se guardan sus características únicas. Los cambios en el material genético de las especies permiten el proceso de adaptación. El proceso de evolución se ve afectado por: la selección natural, la recombinación de material genético y la mutación; fenómenos que se presentan durante la reproducción de las especies. La competencia entre los individuos por los recursos naturales limitados y por la posibilidad de procreación o reproducción permite que sólo los mejor adaptados sobrevivan. Esto significa que, en términos generales, el material genético de los mejores individuos sobrevive y se reproduce.

Los métodos de computación evolutiva manipulan una población de soluciones potenciales codificadas en cadenas o vectores que las representan. Los operadores artificiales de selección, cruza y mutación son aplicados para buscar los mejores individuos (mejores soluciones) a través de la simulación del proceso evolutivo natural. Cada solución potencial se asocia con un valor de aptitud, que mide qué tan buena es comparada con las otras soluciones de la población. Este valor de aptitud es la simulación del papel que juega el ambiente en la evolución natural darwiniana. Este paradigma se resume en la Figura 4.5.

Para comenzar se crea la población completamente al azar. En la configuración inicial hay que tener en cuenta que la distribución de valores debe ser uniforme para cada rango representado por los cromosomas. Luego se decodifica el genotipo en el fenotipo de esta población inicial y se evalúa la aptitud de cada individuo: se le asigna un valor numérico a su “capacidad de supervivencia” o bien, en el espacio de soluciones del problema, se mide que

```
Crear Población
Evaluar Población
Mientras MejorAptitud < AptitudRequerida
  Seleccionar Progenitores
  Reproducir Progenitores
  Evaluar nueva Población
FinMientras
```

Figura 4.5. Algoritmo básico de computación evolutiva.

tan bien resuelve el problema cada individuo. A continuación se entra en el bucle de optimización o búsqueda. Este ciclo termina cuando se encuentra una solución adecuada para el problema —cuando la aptitud para el mejor determina que su fenotipo es suficientemente bueno como solución— o se cumple un número máximo de iteraciones.

Durante el proceso evolutivo artificial se aplican varios operadores. Mediante un proceso de tipo estocástico se genera una nueva población de individuos tomando en cuenta la aptitud de cada uno. Básicamente, durante la selección se decide cuáles individuos serán padres de una nueva generación. Los operadores más elementales que se aplican a los cromosomas progenitores son las cruzas y las mutaciones. Las cruzas son intercambios de genes: el proceso consiste en intercambiar segmentos de los cromosomas de las parejas seleccionadas en forma aleatoria. Cuando un cromosoma sufre una mutación el alelo de uno de sus genes cambia en forma aleatoria. Finalmente la población nace y se decodifica el genotipo en fenotipo para evaluar su aptitud. La nueva población puede remplazar completamente a la población anterior o solamente a los peores individuos. Al volver al principio del ciclo evolutivo se verifican las condiciones de finalización y mientras ninguna se cumpla el proceso se repite nuevamente.

Cuando se pretende resolver un problema mediante computación evolutiva es necesario determinar un conjunto de especificaciones clave:

- *Representación de los individuos*: lo primero es determinar cómo se representa una solución del problema mediante cromosomas. Además hay que especificar cómo se obtiene una solución del problema a partir del material genético.

- *Función de aptitud*: el objetivo en este caso es encontrar una medida de la capacidad de supervivencia de un individuo, sus posibilidades de

procrear y transferir la información de sus genes a la próxima generación. En el dominio de las soluciones, se debe poder medir qué tan buena es cada solución en relación a las demás.

- *Mecanismo de selección*: data toda una población evaluada según la aptitud se debe elegir a los individuos que serán padres de la próxima generación. Los diferentes operadores de selección actúan asignando una alta probabilidad a los mejores pero sin dejar a los peores sin ninguna posibilidad de ser elegidos.

- *Operadores de reproducción y variación*: los operadores básicos son las cruzas y mutaciones. Existen muchos otros operadores que actualmente se utilizan pero sin embargo estos dos son los más elementales y, de una u otra forma, se encuentran presentes en todos los algoritmos de computación evolutiva.

4.2.2. Algoritmo evolutivo para la segmentación de voz

Marcadores de segmentación

Considerando la señal de voz según lo descrito en la Sección 2.1.1, la segmentación da como resultado un conjunto $\Phi = \{E_m\}$ donde cada segmento E_m contiene vectores de características $x(t;k)$ con determinado grado de pertenencia. Sobre esta definición general se harán dos restricciones. La primera es considerar que la segmentación es totalmente exclusiva, es decir, cada vector de características puede pertenecer a sólo un segmento $x(t;k) \in E_{j_1} \Leftrightarrow x(t;k) \notin E_{j_2} \forall j_2 \neq j_1$. Esto permite describir la pertenencia sin un *grado* de pertenencia asociado a cada vector. La segunda restricción está en que el orden temporal según el que aparecen los vectores de características en los segmentos no puede ser invertido. Las dos restricciones se pueden expresar conjuntamente mediante $x(t_1;k) \in E_{j_1} \wedge x(t_2;k) \in E_{j_2} \Leftrightarrow t_1 < t_2 \forall j_1 < j_2$.

Dadas estas restricciones, se puede representar la segmentación mediante el vector de los marcadores del primer elemento de cada segmento $\boldsymbol{\phi} = \left[M_1, M_2, \ldots, M_{N_\phi}\right]$ con $N_\phi = |\Phi| + 1$ ya que se incluyen los marcadores inicial y final y además $1 \leq M_1 < M_2 < \ldots < M_{N_\phi} \leq T + 1$. Como se verá luego, es conveniente dejar abierta la posibilidad de que la primera marca sea mayor a 1 y la última menor que $T + 1$. Estrictamente $\mathbf{x}_t$ no está definido en $t = T + 1$ pero si será válido el marcador para la definición de la función de aptitud.

Representación de los individuos

El primer aspecto a resolver en el diseño del algoritmo de computación evolutiva es la codificación del problema en un alfabeto finito. Tradicionalmente se han empleado cadenas binarias —los denominado algoritmos genéticos puros— pero actualmente se están empleando esquemas más flexibles [Michalewicz, 1992, Merelo et al., 2000].

En el material genético de cada individuo de la población se deberá codificar un conjunto de marcadores de segmentación. Esta codificación tomará como punto de partida la segmentación lineal de la emisión de voz. En principio se trabajará en la base de que se conoce el número de segmentos $|\Phi|$. Luego se discutirá un método para eliminar esta restricción. La partición lineal consiste en asignar los marcadores de cada segmento según $M_j = M_1 + \frac{M_{N_\Phi} - M_1}{N_\Phi - 1}(j-1)$ con $1 < j < N_\Phi$, donde los marcadores inicial y final pueden no necesariamente ser 1 y T. De hecho, se implementó un detector de inicio y finalización de la emisión basado en el análisis por ventanas de la energía según la ecuación (2.8), lo cual permite reducir el espacio de búsqueda para la segmentación.

A partir de esta segmentación lineal se pueden definir los desplazamientos de los marcadores como $\Delta\boldsymbol{\phi} = \left[\Delta M_2, \Delta M_3, \ldots, \Delta M_{N_\phi - 1}\right]$, que será un vector más conveniente para la evolución (ver Figura 4.6). El vector de desplazamientos $\Delta\boldsymbol{\phi}$ no incluye al desplazamiento para el primer y último marcador debido a que quedan fijos. Los desplazamientos para los marcadores ΔM_j son números enteros que están en un rango determinado por las máximas longitudes posibles para los segmentos. En el caso de la segmentación de fonemas es suficiente que este rango permita hasta 50 ms de desplazamiento. Sin embargo, para la segmentación de sílabas, el rango puede llegar a los 200 ms.

De esta forma queda definida la codificación del material genético de cada individuo como un vector de enteros, con rango acotado y conocido, que posee los desplazamientos que deben realizarse a partir de los marcadores de la segmentación lineal, sin incluir el primero y el último. El método para obtener los marcadores a partir de la información codificada en el material genético de cada individuo es $M_j = M_1 + \frac{M_{N_\phi} - M_1}{N_\phi - 1}(j-1) + \frac{\Delta M_j}{T_d}$ con $1 < j < N_\phi$. En esta ecuación aparece el paso de las ventanas de análisis T_d para convertir el tiempo de los desplazamientos de cada marcador en índices de tiempo en el análisis por tramos.

Quedan por resolver algunas cuestiones relacionadas con el proceso mismo de evolución. Dado que evoluciona una codificación de las soluciones del

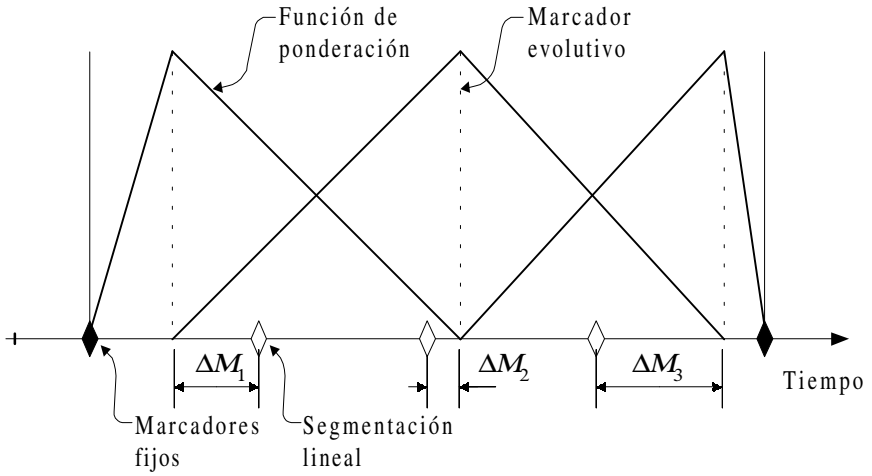

Figura 4.6. Marcadores de segmentación y funciones de ponderación. En este ejemplo se pueden observar los marcadores evolutivos (líneas de punto) y la segmentación lineal (♢). A partir de los marcadores se esquematiza también la función de ponderación $\alpha(\cdot)$.

problema y no las soluciones en sí mismas, es posible que durante la evolución el material genético dé como resultado fenotipos no válidos (soluciones incoherentes). En este problema en particular y dada la codificación elegida, existen dos casos en que las soluciones no son válidas. El primero es cuando al decodificar los marcadores no se respeta su orden natural y se producen solapamientos. El segundo caso es cuando uno o más marcadores están fuera de los límites de tiempo de la emisión, posibilidad que existe independientemente del primer caso dado que los marcadores inicial y final no forman parte de la evolución.

El problema se puede resolver de muchas formas [Michalewicz, 1992]. Por ejemplo, se podría elegir una codificación que no permita estos errores genéticos luego de la aplicación de los diferentes operadores. También se podrían diseñar operadores que no permitan la generación de cromosomas erróneos a partir de cromosomas válidos. En cualquier caso se trata de adaptaciones del algoritmo de computación evolutiva al problema en mano.

Una técnica más sencilla que no implica una modificación importante en la idea de la computación evolutiva es la operación de verificación y reparación combinada con la penalización de aptitud. Para realizar la verificación del solapamiento simplemente se debe comprobar la inecuación $M_{j_1} < M_{j_2} \forall j_1 < j_2$ con $1 < j_1, j_2 < N_\phi$. La verificación se completa comprobando que ningún marcador se encuentra fuera de los límites determina-

dos por los marcadores inicial y final. Todo se puede resumir ampliando los rangos en la expresión anterior a $1 \leq j_1, j_2 \leq N_\phi$.

Función de aptitud

Es necesario obtener una medida de qué tan buena es la solución que ofrece cada individuo. La función de aptitud trabaja en el dominio del problema, sobre el fenotipo de cada individuo.

Se define el vector propio de un segmento como:

$$\varphi_j(k) = \frac{1}{A_j(\cdot)} \sum_{t=M_j}^{M_{t+1}-1} \alpha(\cdot) x(t;k)$$

siendo $A_j(\cdot) = \sum_{i=M_j}^{M_{j+1}-1} \alpha(\cdot)$ con $0 < j < N_\phi - 1$. El vector propio cumple la función de representar a todo el segmento ya que se obtiene mediante un promedio ponderado de todos sus vectores de características.

La función de ponderación $\alpha(\cdot)$ tiene por objetivo asignar diferente peso a los vectores de características según se encuentren más cerca o más lejos del límite del segmento. Como función de ponderación se puede definir, por ejemplo, $\alpha(d, N) = e^{-\frac{d}{N}}$ o bien una relación lineal $\alpha(d, N) = 1 - \frac{d}{N}$, siendo d la distancia al marcador y N el número total muestras a ponderar. En el caso de que se adopte la relación lineal y $1 \leq d \leq N$, se puede demostrar que $A(d, N) = \sum_{d=1}^{N} 1 - \frac{d}{N} = \frac{1}{2}(N+1)$.

Para distinguir entre el vector propio de un segmento ponderado como anterior o posterior a un marcador, se utilizarán los superíndices '–' y '+', respectivamente. A continuación se presentan las ecuaciones de los vectores propios de un segmento según su posición relativa al marcador:

$$\varphi_j^-(k) = \frac{\sum_{t=M_j}^{M_{j+1}-1} \alpha(M_{j+1} - t, N_{M_{j+1}}) x(t;k)}{\sum_{t=M_j}^{M_{j+1}-1} \alpha(M_{j+1} - t, N_{M_{j+1}})}$$

y

$$\varphi_j^+(k) = \frac{\sum_{t=M_j}^{M_{j+1}-1} \alpha(t - M_j + 1, N_{M_{j+1}}) x(t;k)}{\sum_{t=M_j}^{M_{j+1}-1} \alpha(t - M_j + 1, N_{M_{j+1}})}$$

con $N_{M_j} = M_j - M_{j-1} + 1$.

La distancia euclídea entre dos vectores propios en torno al marcador M_j es:

$$\delta_j^E = \sum_{k=1}^{N_x} \left(\varphi_{j-1}^-(k) - \varphi_j^+(k)\right)^2; \quad 1 < j < N_\phi - 1 \tag{4.14}$$

A partir de esta expresión se define la función de aptitud como el promedio $\Gamma_\phi = \frac{1}{N_\phi - 2} \sum_{j=2}^{N_\phi - 1} \delta_j^E$. Reemplazando según las consideraciones tomadas hasta el momento se obtiene:

$$\Gamma_\phi = \frac{1}{N_\phi - 2} \sum_{j=2}^{N_\phi - 1} \sum_{k=1}^{N_x} \left[\frac{2}{N_{M_{j-1}} + 1} \sum_{t=M_{j-1}}^{M_j - 1} \left(1 - \frac{M_j - t}{N_{M_{j-1}}}\right) x(t;k) - \right. \\ \left. - \frac{2}{N_{M_j} + 1} \sum_{t=M_j}^{M_{j+1}-1} \left(1 - \frac{t - M_j + 1}{N_{M_j}}\right) x(t;k) \right]^2 \tag{4.15}$$

Ejemplos para la función de aptitud

Se presentan los siguientes ejemplos con el objetivo de aclarar la forma en que actúa la función de aptitud. Para simplificarlos se considera $\alpha(\cdot) = 1$, $N_x = 1$, $T = 30$ y $N_\phi = 6$. Así la función de aptitud queda:

$$\Gamma_\phi = \frac{1}{4} \sum_{j=2}^{5} \left(\frac{1}{N_{M_{j-1}}} \sum_{t=M_{j-1}}^{M_j - 1} x(t) - \frac{1}{N_{M_j}} \sum_{t=M_j}^{M_{j+1}-1} x(t) \right)^2$$

Dado el vector de características:

`0000001111000000000001111000000`

se evalúa la aptitud de las siguientes segmentaciones:

1. segmentación ideal $\boldsymbol{\phi}_I = [7, 11, 21, 25]$:

   `000000/1111/0000000000/1111/000000` $\rightarrow \quad \Gamma_{\phi_I} = 1{,}00,$

2. uno en posición ideal y uno lineal $\boldsymbol{\phi}_{IL} = [7, 11, 19, 25]$:

   `000000/1111/00000000/001111/000000` $\rightarrow \quad \Gamma_{\phi_{IL}} = 0{,}72,$

3. segmentación lineal $\boldsymbol{\phi}_L = [7, 13, 19, 25]$:

   `000000/111100/000000/001111/000000` $\rightarrow \quad \Gamma_{\phi_L} = 0{,}44,$

4. segmentación incorrecta $\boldsymbol{\phi}_X = [7, 11, 17, 25]$:

   `000000/11110000/00/00001111/000000` $\rightarrow \quad \Gamma_{\phi_X} = 0{,}25.$

Selección

Existen varias formas de realizar la selección de los progenitores. Al igual que en la naturaleza, la selección no está relacionada directamente con la aptitud de un individuo sino a través de operadores probabilísticos. Desde el punto de vista del algoritmo de búsqueda, la selección lleva a cabo la tarea de concentrar el esfuerzo computacional en las regiones del espacio de soluciones que se presentan como más prometedoras [Salomon, 1998]. Los operadores de selección utilizados en la computación evolutiva generalmente encuentran un compromiso entre estos dos extremos. Tres operadores elementales de selección son: la rueda de ruleta, la selección por ventanas y la competencia [Goldberg, 1997].

En los siguientes experimentos se utilizó el método de competencias, según el cual se eligen completamente al azar $v > 1$ individuos, se los hace competir por aptitud y queda seleccionado el ganador. Generalmente se utilizan valores de v entre 2 y 5 dependiendo del tamaño de la población. Este método es uno de los más utilizados debido a lo simple y eficiente de su implementación.

Reproducción

La reproducción es el proceso mediante el cual se obtiene la nueva población a partir de los individuos seleccionados y los operadores de variación. Existen varias alternativas para realizar la reproducción, en el caso más sencillo se obtienen todos los individuos de la nueva población a partir de variaciones (cruzas y mutaciones) de los progenitores. Es posible también transferir directamente a la población nueva los padres seleccionados en la población anterior y completar los individuos faltantes mediante variaciones.

Una variante adicional en la reproducción que no se extrae directamente de la evolución biológica pero que es utilizada con muy buenos resultados es el *elitismo*. En esta estrategia se busca el mejor individuo de la población anterior e independientemente de la selección y variación se lo copia exactamente en la nueva población. De esta manera se resguardada la mejor solución a través de las generaciones.

Operadores de variación

La *mutación* trabaja alterando alelos de genes con una probabilidad p_m muy baja, por ejemplo $p_m = 0{,}001$. Las mutaciones son típicamente realizadas con una probabilidad uniforme en toda la población y el número de mutaciones por individuo puede ser fijado de acuerdo a esta probabilidad y la cantidad de individuos. En los casos más simples se da la posibilidad de mutar sólo un alelo por individuo o se distribuye uniformemente sobre todo el cromosoma. Cuando se utiliza elitismo es posible asegurar la mejor solución de cada generación lo que permite utilizar probabilidades de mutación más altas. Una revisión comparativa y combinación de diferentes métodos de mutación puede verse en [Chellapilla, 1998].

En el algoritmo de segmentación evolutiva se elige al azar un gen y se lo muta mediante $\Delta M_{j^*}(G+1) = \Delta M_{j^*}(G) + R \sqcup (-1, 1)$, donde j^* es el gen elegido para la mutación, G es el número de la generación actual y R es el rango en que se produce la alteración. La función $\sqcup(a, b)$ devuelve un número real al azar entre a y b con una distribución uniforme. Existe un control para que el resultado no salga del rango previsto para los desplazamientos de los marcadores.

La *cruza* en un operador que actúa sobre dos cromosomas para obtener otros dos. Existen dos tipos de cruzas: cruzas simples y cruzas múltiples. En las cruzas simples se elige un punto de cruza al azar y se intercambia el material genético correspondiente a las partes del cromosoma que separa este punto. En la cruza múltiple puede cortarse el cromosoma en más de

dos partes para realizar el intercambio. También en este caso los puntos son elegidos al azar. Para el problema de segmentación de voz se utiliza la cruza simple. El punto de cruza se elige al azar pero los dos cromosomas se cortan en el mismo lugar. Esto asegura que la longitud de los cromosomas se mantenga después de la cruza. Sin embargo, sería de interés para aplicaciones de tiempo real poder tener cromosomas con diferentes números de segmentos y así elegir un punto de cruza diferente para cada uno de los dos cromosomas que intervienen. Se presentan más detalles de este algoritmo evolutivo en [Milone et al., 2002].

4.2.3. Algoritmo de segmentación con detector de máximos

La ecuación (4.14), que mide la distancia euclídea entre dos vectores propios, puede utilizarse como una medida del cambio en los vector de características a cada lado de un marcador. Si estas distancias no se integran sobre toda la frase como se hizo en la función de aptitud (4.15), entonces pueden utilizarse como medida de los cambios a nivel *local* para cada tramo de análisis. Se puede esperar que en las posiciones de la frase en donde esta medida sea máxima se encuentren los límites que separan dos estructuras acústicas relevantes. En base a esta idea se desarrolla a continuación un método de segmentación ciega de voz. En este caso no existe un conjunto de marcadores predefinidos ni se necesita medir la aptitud como en el caso de la segmentación evolutiva. Ahora el conjunto de marcadores surgirá a través de un proceso iterativo de optimización.

Redefinición de la distancia entre segmentos

Es necesario realizar unos cambios en la definición original, ya que ahora no se poseen marcadores. Para independizar la distancia (4.14) del contexto es necesario fijar la cantidad de vectores de características que se consideran a cada lado de un tramo de voz dado. Así, se redefinen los nuevos vectores propios para cada t:

$$\varphi_t^{-\Delta M}(k) = \frac{\sum_{\tau=t-1}^{t-\Delta M} \alpha(t-\tau, \Delta M) x(\tau; k)}{\sum_{\tau=t-1}^{t-\Delta M} \alpha(t-\tau, \Delta M)}$$

y

$$\varphi_t^{+\Delta M}(k) = \frac{\sum_{\tau=t}^{t+\Delta M-1} \alpha(\tau - t + 1, \Delta M) x(\tau; k)}{\sum_{\tau=t}^{t+\Delta M-1} \alpha(\tau - t + 1, \Delta M)}$$

Considerando una relación lineal para $\alpha(\cdot)$, se define la distancia euclídea entre los segmentos en torno al tiempo t y con ancho ΔM:

$$\begin{aligned} \delta_t^E(\Delta M) &= \frac{2}{\Delta M + 1} \sum_{k=1}^{N_x} \left[\sum_{\tau=t-1}^{t-\Delta M} \left(1 - \frac{t-\tau}{\Delta M}\right) x(\tau; k) - \right. \\ &\quad \left. - \sum_{\tau=t}^{t+\Delta M-1} \left(1 - \frac{\tau - t + 1}{\Delta M}\right) x(\tau; k) \right]^2 \end{aligned}$$

con $\Delta M < t \leq T - \Delta M$.

Búsqueda de los picos de segmentación

Para segmentar resta definir un algoritmo que detecte los picos de la función $\delta_t^E(\Delta M)$, es decir, aquellos instantes de tiempo en donde se realizan mayores cambios en los vectores de características. La detección de estos máximos se realiza en dos pasos: búsqueda de los candidatos por caída de gradiente y selección de los mejores máximos.

El algoritmo para buscar los mejores candidatos consiste en acumular los gradientes que se encuentran a cada lado de un pico y de esta forma medir su importancia relativa. El algoritmo comienza considerando que existe un candidato en cada instante de tiempo t de la curva $\delta_t^E(\Delta M)$ y en cada paso elimina aquellos candidatos para los que no se cumpla $\delta_{t-1}^E(\Delta M) < \delta_t^E(\Delta M) > \delta_{t+1}^E(\Delta M)$. Cada vez que un candidato no supera esta prueba se elimina de la lista y se acumula su diferencia con el que sea mayor de los que están a su lado. En la Figura 4.7 se resume este algoritmo de detección de picos.

Los candidatos quedan indicados en los elementos $pk_t \neq 0$. La selección definitiva se realiza en dos etapas de filtrado con diferentes tamaños de ventana. En la primera se consideran ventanas de ancho $W_f = 10T_d$ y se eliminan los máximos menores a un 10 % del máximo en la ventana. En la

```
Comienzo: δ_t^2 = δ_t^E(ΔM)^2; pk_t = 1∀t.
Repetir
  Para cada t Si pk_t ≠ 0
    Si (δ_{t+1}^2 ≥ δ_t^2) ∧ (δ_{t+1}^2 − δ_t^2 > δ_{t−1}^2 − δ_t^2)
      pk_{t+1} = pk_{t+1} + δ_{t+1}^2 − δ_t^2
      pk_t = 0
    Si δ_{t−1}^2 ≥ δ_t
      pk_{t−1} = pk_{t−1} + δ_{t−1}^2 − δ_t^2
      pk_t = 0
  FinPara
Hasta no observar cambios en pk_t.
```

Figura 4.7. Algoritmo detector de picos de segmentación.

segunda etapa se consideran ventanas de una ancho menor $(W_f/2)$ y se deja un único máximo por ventana.

4.2.4. Resultados

Las pruebas que se realizaron se dividen en tres partes. En primer lugar se presenta un ejemplo que tiende a mostrar las características más importantes del algoritmo de segmentación evolutiva. Este experimento se realiza en base a una señal creada artificialmente con información que resulta en una segmentación obvia. Los segundos experimentos se realizaron en un archivo de voz y se comparan los resultados con la segmentación realizada por MOM. En los últimos experimentos se segmentaron 600 frases.

Para las primeras pruebas se generó un archivo de 1 segundo con las siguientes señales: silencio $[0, 166)$ ms; ruido blanco $[166, 250)$ ms; silencio $[250, 750)$ ms; seno de 1000 Hz $[750, 833)$ y silencio $[750, 1000]$ ms. En esta señal los segmentos del ruido y la senoidal son fácilmente detectables. El algoritmo de segmentación evolutiva se aplicó con los parámetros se muestran en la Tabla 4.4. En la Figura 4.8 se puede observar la evolución de la aptitud del mejor individuo por generación, en la Figura 4.9 la superficie de aptitud y en la Figura 4.10 el resultado de la segmentación.

En el ejemplo de segmentación de voz se realizaron diversas pruebas con un archivo del corpus de voz Albayzin. Para el primer caso en la segmentación de voz se utilizaron los parámetros de la Tabla 4.5 y se exigieron

Individuos en la población	10
Rango de alelos en ms	400
Probabilidad de cruzas	0.5
Probabilidad de mutaciones	0.5
Generaciones	500
Elitismo	si
Ancho de la ventana de análisis en ms	8
Paso de la ventana de análisis en ms	8
Tipo de análisis (ver Sección 2.1.4)	CCEM

Tabla 4.4. Parámetros utilizados en el ejemplo de ruido y senoidal.

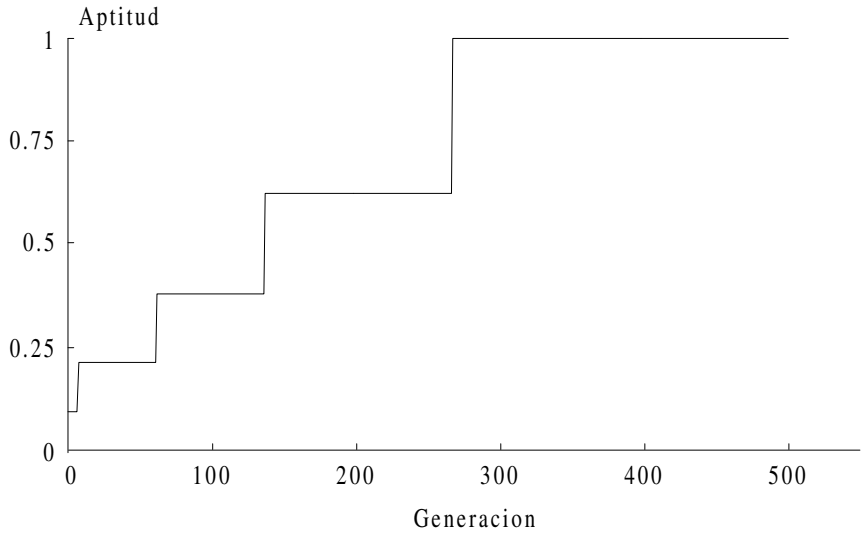

Figura 4.8. Aptitud para el mejor individuo en el ejemplo de ruido y senoidal. En esta curva se puede observar claramente el efecto de la estrategia elitista en la selección de progenitores.

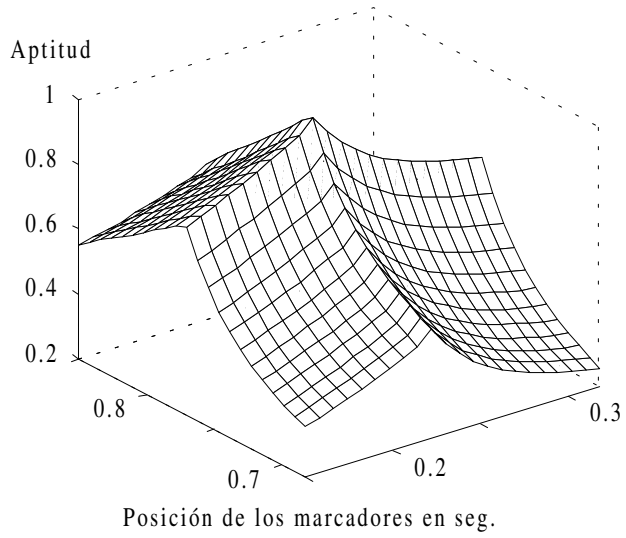

Figura 4.9. Superficie de aptitud para el ejemplo de ruido y senoidal. Se evoluciona la posición de los dos marcadores centrales (los marcadores de inicio y fin se encuentran fijos).

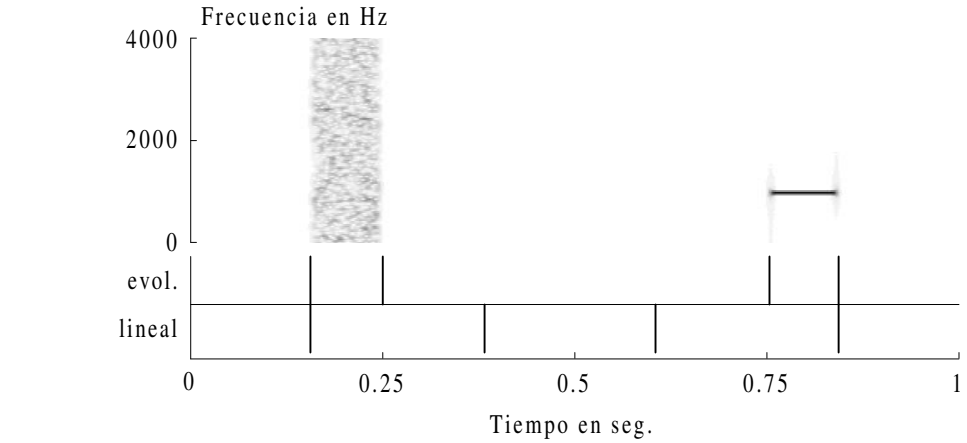

Figura 4.10. Segmentación obtenida en el ejemplo de ruido y senoidal. En la parte superior se observa el espectrograma de la señal del ejemplo. Las líneas de abajo indican la segmentación lineal, a partir de la cual evolucionan los marcadores. Las líneas de arriba (evol.) indican la segmentación realizada por el algoritmo evolutivo.

Individuos en la población	200
Rango de alelos en ms	100
Probabilidad de cruzas	0.5
Probabilidad de mutaciones	0.5
Generaciones	500
Elitismo	si
Ancho de la ventana de análisis en ms	16
Paso de la ventana de análisis en ms	16
Tipo de análisis (ver Sección 2.1.4)	CCEM

Tabla 4.5. Parámetros utilizados en el primer ejemplo con una señal de voz.

tantos segmentos como sílabas tenía la frase.

En la Figura 4.11, con etiqueta 'evol.1', se observa el resultado de la segmentación por sílabas. Se han realizado varias pruebas en las que la convergencia se obtuvo antes de las 100 generaciones. En la parte inferior de la misma gráfica se indica como referencia la segmentación obtenida con MOM. Como se explicó antes (Sección 4.1.3), esta segmentación se obtiene buscando la secuencia más probable mediante el algoritmo de Viterbi. En esta segmentación se provee a los MOM de la transcripción completa de cada frase.

Para el segundo caso de segmentación mediante el algoritmo evolutivo se modificó únicamente el rango de los alelos, que se amplió a 250 ms para poder incluir palabras. Sin embargo, se mantuvo la exigencia de cantidad de segmentos de acuerdo con una segmentación silábica. En la Figura 4.11, con etiqueta 'evol.2', se observa claramente que el método tiende a realizar una segmentación por palabras.

Para las pruebas segmentación local con detección de máximos se utilizó el subconjunto SC1 del corpus de habla Albayzin (Apéndice A.2). Los parámetros utilizados en el algoritmo fueron $\Delta M = 21$ y $W_f = 10T_d$. En la Figura 4.11, se puede apreciar la curva de δ_t^2 y la segmentación realizada por el algoritmo (con etiqueta 'max δ'). Luego se midió el error sobre las 600 frases, contando las veces en que la segmentación resultantes coincidía con la segmentación realizada mediante los MOM. Para la segmentación silábica el error promedio fue de 32.36 % y para la segmentación de palabras 47.57 %. Si se acepta el error de una sílaba por palabra el error promedio se reduce a 7.59 %.

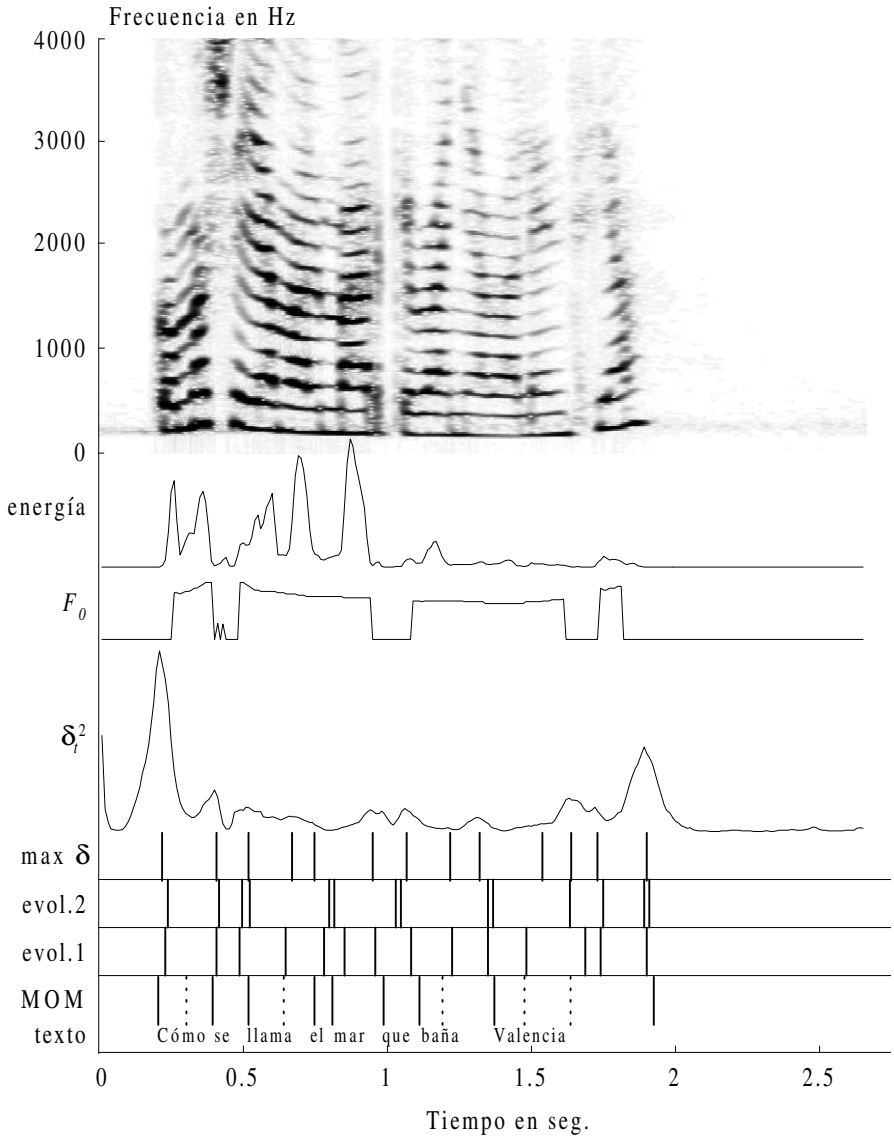

Figura 4.11. Segmentación de la frase *¿Cómo se llama el mar que baña Valencia?* mediante los diferentes métodos evaluados. En la parte superior se observa el espectrograma y las curvas de energía y frecuencia fundamental. A continuación se puede ver la curva δ_t^2, a partir de la cual se obtiene la segmentación silábica por el algoritmo de detección de máximos, indicada a como 'max δ'. Las etiquetas 'evol.2' y 'evol.1' indican las segmentaciones por palabras y por sílabas obtenidas con el algoritmo evolutivo. En la parte inferior se observa la segmentación obtenida mediante MOM y el etiquetado en palabras correspondiente (con líneas de puntos la segmentación silábica).

4.2.5. Discusión

Los primeros resultados muestran que el silencio y la senoidal son segmentados fácilmente, con muy poca carga computacional, y una población mínima que hasta es inusual en métodos de computación evolutiva. La curva de evolución de aptitud (Figura 4.8) muestra claramente el efecto de la estrategia elitista en su comportamiento de ascenso monotónico. La utilización del elitismo permitió elegir una alta probabilidad de mutaciones acelerando la convergencia (sin llegar a una búsqueda al azar). También se puede observar en la curva de evolución que el tiempo total podría ser reducido casi a la mitad. Vale destacar que se podría reducir el análisis de la señal a simplemente el cálculo de la energía por ventanas. El rango de los alelos (400 ms) fue fijado en base a cuánto se tiene que poder desviar el marcador de la segmentación lineal para poder realizar la segmentación ideal.

En el caso de la segmentación por sílabas los marcadores encontrados por el método de segmentación evolutiva coinciden casi exactamente (considerando las ventanas de análisis utilizadas) con los marcadores de la segmentación por MOM (Figura 4.11, etiqueta 'evol.1'). Sin embargo, se puede ver que existe un error por omisión en la primera sílaba y uno por inserción en la penúltima. El primer error puede ser debido a que la emisión de la palabra *cómo* tiene el mismo fonema /o/ en cada sílaba. Además está separado por una /m/, que ofrece una transición suave de las formantes de los sonidos vocálicos de su entorno, que en este caso son iguales. El error en el anteúltimo marcador puede responder a varias causas. En primer lugar debe considerarse que el rango elegido para los alelos apenas alcanza para abarcar a la sílaba /cia/. Por otro lado se puede observar que dado el error de omisión en la primera sílaba el método queda forzado a insertar un marcador (ya que la cantidad total de marcadores es fija). En la misma línea de razonamiento, se puede observar que la pausa que se produce entre /len/ y /cia/ determina una fuerte diferencia entre estas regiones y el método encuentra que la función de aptitud se maximiza haciendo esta separación a costa de unir la palabra *cómo*. En la segmentación 'evol.2', si bien el rango de los alelos es mucho mayor (250 ms), aún se observa el error en *Valencia*.

En la segmentación por palabras (Figura 4.11, etiqueta 'evol.2') se puede ver la forma en que la elección del rango de los alelos condiciona fuertemente los resultados. Esto permitiría seleccionar el rango de los alelos a partir de las longitudes típicas de las unidades a segmentar. Sin embargo, puede que esto no sea tan obvio en el caso del habla. Existen palabras que pueden tener la longitud de tan sólo una sílaba o fonema y, de la misma forma, algunas sílabas pueden tener la longitud de toda una palabra. Este

puede ser el punto más débil del método dado que no utiliza otra información relativa al contexto o a la gramática, como en el caso de los MOM. Otro aspecto que puede constituir una desventaja es el tiempo total necesario para realizar la segmentación. Para dar una idea de estos tiempos, para segmentar un frase de 3.5 segundos en un procesador Pentium Celeron 366 MHz se necesitaron 17.2 segundos. Esta podría ser una limitación importante para un sistema de tiempo real, pero no invalida la aplicación del método a la segmentación de corpus de habla.

Al igual que el rango de los alelos controla el tipo de segmentación en el algoritmo evolutivo, los parámetros ΔM y W_f lo hacen para el método por detección de máximos. En este caso (Figura 4.11, etiqueta 'max δ') se puede observar que nuevamente no se detecta la separación silábica de la palabra *cómo* y se agrega un marcador extra en la palabra *Valencia*. Hay que destacar que en este método no es necesario conocer a priori la cantidad total de marcadores. De los 12 marcadores de la segmentación por MOM, el método por detección de máximos ha encontrado 11 (sin contar el primero y el último de la frase). Esto abre la posibilidad de combinar ambos métodos, uno para la detección de los extremos de la frase y la cantidad de sílabas, y el otro para la segmentación propiamente dicha.

Cuando se realizaron pruebas con los coeficientes espectrales (CE) (definidos en la Sección 2.1.2) se observó que la energía condicionaba fuertemente la posición de los marcadores. En este caso las marcas se ubicaron en las máximas variaciones de energía, no segmentando sílabas sino más bien vocales. Esta influencia de la energía también se observa, aunque en menor medida, para los CCEM. Esto último podría dar lugar a una revisión del algoritmo para obtener una normalización por energías que anule este efecto indeseado. Las pruebas realizadas con coeficientes de predicción lineal (CPL) no difieren mucho de las realizadas con CCEM pero el cálculo de los CPL es algo más lento.

Por último, cabe mencionar una particularidad de los métodos propuestos: en ningún caso hay un proceso de entrenamiento ni parámetros almacenados para su posterior utilización durante la segmentación. Esto, si bien hace que los métodos trabajen con muy poca información de la tarea a realizar, también les da robustez, flexibilidad y aprovecha al máximo su capacidad de autoadaptación.

4.3. Segmentación y clasificación conjunta

Para buscar una solución integrada, que combine un buen rendimiento tanto en la segmentación como en la clasificación, se realizaron diferentes pruebas de estimación de secuencias de estructuras acentuales (SEA) mediante MOM. Las alternativas investigadas se implementan mediante cambios en los tres niveles de un MOM para RAH: procesamiento de la señal, modelado acústico y modelado del lenguaje. En el Capítulo 2 se ha tratado ampliamente la técnica de los MOM y a continuación se describen las adaptaciones que se realizaron para utilizarlos en la estimación de SEA.

4.3.1. Alternativas en el procesamiento de la señal

Para el procesamiento de la señal es necesario redefinir el vector $\mathbf{x}_t$. En la Sección 2.1.1 se definió este vector como:

$$x(t;k) = \mathcal{T}(k)\left\{v(t;n)\right\}, \quad 0 < k \leq N_x$$

donde $\mathcal{T}(k)$ es un operador para la transformación de dominio y $v(t;n)$ los tramos de voz en el tiempo. Estos vectores forman las evidencias acústicas que el MOM modela mediante las mezclas de N_c gaussianas en $\mathbb{R}^{N_x}$ (Sección 2.2.1). En esta sección se describen algunas de las alternativas evaluadas para $\mathcal{T}(k)$. Debe destacarse que en todos los casos $\mathcal{T}(k)$ no puede basarse en una segmentación conocida. Esto hace que queden fuera del estudio las cadencias de F_0 (Secciones 1.3.3 y 3.3.5), ya que para el cálculo de las pendientes se requería conocer de antemano los límites de la sílaba.

Energía y frecuencia fundamental

En este caso se define $\mathbf{x}_t = [\epsilon(t), F_0(t)]$. La energía en función del tiempo ya se definió en la ecuación (2.8):

$$\epsilon(t) = \log \sum_{n=1}^{N_v} v(t;n)^2$$

La $F_0(t)$ se calcula en base al cepstrum real, como se describe en la Sección 2.1.4 (página 77). En el caso de completarse el vector con coeficientes delta y aceleración, se constituye:

$$\mathbf{x}_t = [\epsilon(t), F_0(t), \Delta\epsilon(t), \Delta F_0(t), \Delta^2\epsilon(t), \Delta^2 F_0(t)].$$

Curvas de diferencia por ajuste

Siguiendo las ideas presentadas en la Sección 3.3.5 se incorporaron procesamientos alternativos para la F_0. El primer paso fue considerar un ajuste de la curva de entonación mediante polinomios de grado variable entre 3 y 15. Los coeficientes para estos polinomios fueron calculados en base al método de cuadrados mínimos generalizado, resuelto por descomposición en valores singulares [Press et al., 1997, Sec. 15.4]. Una vez obtenido el polinomio de interpolación, se resta a la curva de entonación original y se utiliza la curva resultante como otra evidencia para los MOM. Esta curva resultante fue denominada *diferencia de entonación por ajuste* (difF_0). En este caso el vector de evidencias acústicas para los MOM queda definido como: $\mathbf{x}_t = [\epsilon(t), \mathrm{dif}F_0(t)]$.

Este análisis de diferencia por ajuste se hizo extensivo a la curva de energía y se realizaron pruebas con *diferencia de energía por ajuste* (difϵ). También se probaron polinomios con grados que iban desde 3 hasta 15. Para completar la descripción, el vector de evidencias acústicas para los MOM queda definido según: $\mathbf{x}_t = [\mathrm{dif}\epsilon(t), \mathrm{dif}F_0(t)]$ aunque también se realizaron experimentos con $\mathbf{x}_t = [\mathrm{dif}\epsilon(t), F_0(t)]$.

Otras alternativas evaluadas

Resta por mencionar la utilización de CCEM, tal como se describieron en el Capítulo 2 y como se utilizan normalmente para el RAH. Dado que las unidades elementales a reconocer en esta aplicación de MOM tienen una longitud mayor (generalmente la de una sílaba), también se experimentó con la variación del ancho (T_ω) y el paso (T_d) de la ventana de análisis. El ancho de ventana fue extendido desde 25 hasta 40, 64 y 100 ms. El paso de análisis fue extendido desde 10 hasta 20, 25 y 50 ms.

4.3.2. Alternativas en el modelado acústico

En el caso del modelo acústico (MA) se probaron diversas alternativas que pueden separarse en dos grupos: las relacionadas con lo que se modela y las relacionadas con cómo se modela.

Alternativas en el objeto de modelado

En los MOM utilizados para el RAH, según se describió en el Capítulo 2, las unidades elementales eran los modelos de fonemas ${}^{F}\Theta_{\varphi}$ y con éstos se construían por concatenación los modelos palabras ${}^{W}\Theta_{w}$ (Sección 2.2.4, página 95). Para utilizar los MOM en la estimación de SEA es necesario reemplazar el modelo de fonema por un modelo de tonicidad silábica. De la concatenación de modelos de tonicidad silábica se obtienen las EA que, en la organización estructural del habla, poseen un nivel jerárquico equivalente al de las palabras.

Las dos alternativas básicas para el modelo de tonicidad silábica son la /A/ para las sílabas átonas y la /T/ para las sílabas tónicas:

$$\mathcal{F}_{\Theta} = \left\{ {}^{F}\Theta_{\mathrm{A}}, {}^{F}\Theta_{\mathrm{T}} \right\} \tag{4.16}$$

En base a estos dos modelos se pueden construir todas las EA del corpus de habla utilizado. Por ejemplo:

$${}^{W}\Theta_{\mathrm{ATA}} = {}^{F}\Theta_{\mathrm{A}}\,{}^{F}\Theta_{\mathrm{T}}\,{}^{F}\Theta_{\mathrm{A}} \tag{4.17}$$

donde se ha obviado la definición de un diccionario ya que su estructura es trivial. Adicionalmente, para algunos experimentos se definió un modelo especial para las palabras monosilábicas ${}^{W}\Theta_{\mathrm{M}} = {}^{F}\Theta_{\mathrm{M}}$ (modelos TAM). Para otros se clasificaron las palabras como acentuadas e inacentuadas, utilizando todos los modelos /A/ en estas últimas [Quilis, 1993] (modelos TA-Q). Sobre esta clasificación se dieron algunos ejemplos en el Capítulo 1 (página 34) y se pueden encontrar más detalles en la Sección A.2.3 (página 219).

En busca de ampliar la estructura de los MA también se realizaron pruebas en donde se formaron modelos para cada una de las vocales y diptongos del núcleo silábico con cada tonicidad. Para estos experimentos se formaron 31 modelos elementales distinguiendo en cada caso:

- la vocal o diptongo que forma el núcleo: /a/, /e/, ..., /ai/, /ie/, ...
- su tonicidad: /T/ y /A/.

Los mejores resultados para estos experimentos se resumen luego, en la Sección 4.3.4, bajo la denominación TA-v.

Alternativas en los parámetros del modelo

En relación a la estructura de los MOM se probaron diferentes configuraciones que incluían variaciones en:

- Cantidad de estados: entre 3 y 15.
- Tipo de MOM: continuos, semicontinuos o discretos[3].
- Cantidad de gaussianas (N_c) en la mezclas que modelan las observaciones en cada estado: 1, 2 y 4.

4.3.3. Alternativas en el modelo de lenguaje

Una vez formadas las EA, con cualquiera de las alternativas mencionadas en la sección anterior, se pueden formar modelos de n-gramáticas e incorporar estas probabilidades en el modelo compuesto. En los experimentos realizados se utilizaron siempre modelos de bi-gramáticas con probabilidades estimadas por el método de *back-off* (ecuación (2.36), página 97).

En torno a esta estructura básica se consideraron dos variantes:

- Modelos de tonicidad silábica con distinción entre vocales (como en TA-v) pero sin formar EA (a nivel de palabras). En este caso no se concatenan sílabas para formar palabras sino que se trata a las frases como una secuencia de sílabas y a partir de la cual se construye una secuencia de modelos independientes. En estos experimentos, que luego se denominan TA*-v, las probabilidades del modelo de lenguaje (ML) se incorporan directamente a nivel de sílabas.
- Diferentes pesos relativos para las probabilidades de los MA y ML. En el momento de incorporar las probabilidades del ML en la búsqueda por el algoritmo de Viterbi (ecuación (2.37), página 98), se multiplica $G^{(2)}_{mn}$ por una constante que, en estos experimentos, tomó los valores: 0.01, 0.5, 1.0 y 5.0.

Dado que las restricciones de tiempo no son importantes, todos los experimentos reportados en este capítulo se realizaron sin utilizar el método de podado.

Para terminar, aunque no en relación directa con el ML, cabe mencionar que se han repetido diversos experimentos separando las frases interrogativas de las de tipo declarativo. En lugar de realizar los experimentos con

[3]Sólo para los primeros se reportan resultados.

342 de tipo declarativa y 258 interrogativas, se realizaron solamente con las declarativas o solamente con las interrogativas para evaluar la influencia que pudiera tener esta distinción. En todos los casos se encontró que las variaciones en los resultados eran mínimas (menores al 0.1 %), por lo que no se reportarán más detalles de estos experimentos.

4.3.4. Resumen de resultados

De la amplia lista de combinaciones posibles para las configuraciones presentadas en la sección anterior, se han seleccionado en la Tabla 4.6 los experimentos con resultados de reconocimiento mayor al 40 %. En todos los casos se utilizaron las 600 frases del subconjunto SC1 del corpus de habla Albayzin, ya citado anteriormente (Apéndice A.2). En estos experimentos las frases se separaron en dos grupos, uno para el entrenamiento y el otro para las pruebas de validación (80 y 20 % respectivamente). Para ilustrar estos resultados, se transcribe a continuación un ejemplo de la estimación de SEA realizada por el modelo TA-Q listado en la Tabla 4.6:

Frase: *Ríos de la Comunidad Autónoma Gallega.*

SEA correcta: /TA A A AAAT ATAA ATA/

SEA estimada: /T TA A AAAT ATAA A A/

4.3.5. Discusión

Las consideraciones realizadas en cuanto a la longitud de los segmentos a reconocer han mostrado sus beneficios con las modificaciones realizadas tanto en el procesamiento de la señal como en los parámetros del modelo. Los mejores resultados se han alcanzado para modelos de 7 estados con T_d y T_ω algo superiores al procesamiento estándar en RAH. Es importante destacar que con un procesamiento sencillo como el de $[\epsilon, \mathrm{dif}F_0]$ se han logrado rendimientos comparables al de $[\epsilon, \mathbf{c}_{mel}, \Delta, \Delta^2]$, que cuenta con mucha más información en el vector de evidencias acústicas e implica MOM más complejos, con más parámetros y mayor costo computacional. Evidentemente, al igual que en el Capítulo 3, la eliminación de la función distintiva de F_0 a nivel de frases ha permitido una mejor extracción de la información relativa a la acentuación. Sin embargo, como se podía esperar, no ocurrió lo mismo para el caso de la energía.

Los dos mejores resultados corresponden a los modelos que consideran a los monosílabos por separado, pero el rendimiento del modelo TA-Q no

Procesamiento ($\mathbf{x}_t$)	MA	$\lvert\mathcal{Q}\rvert$	N_c	GP	T_d	T_ω	Rendimiento
$[\epsilon, F_0, \Delta, \Delta^2]$	TAM	5	–	–	20	64	45.56 %
$[\epsilon, F_0, \Delta, \Delta^2]$	TAM	7	–	–	20	64	53.31 %
$[\epsilon, F_0, \Delta, \Delta^2]$	TAM	7	4	–	20	64	55.16 %
$[\epsilon, \mathrm{dif}F_0]$	TAM	7	–	11	20	64	56.82 %
$[\epsilon, \mathrm{dif}F_0, \Delta]$	TAM	7	–	11	20	64	50.56 %
$[\epsilon, \mathrm{dif}F_0, \Delta, \Delta^2]$	TAM	7	–	11	20	64	50.49 %
$[\mathrm{dif}\epsilon, \mathrm{dif}F_0]$	TAM	7	–	13	20	64	44.59 %
$[\epsilon, \mathbf{c}_{mel}, \Delta, \Delta^2]$	TAM	7	–	–	10	25	53.08 %
$[\epsilon, \mathbf{c}_{mel}, \Delta, \Delta^2]$	TAM	4	4	–	10	25	54.88 %
$[\epsilon, \mathbf{c}_{mel}, \Delta, \Delta^2]$	TAM	15	4	–	10	25	43.39 %
$[\epsilon, \mathbf{c}_{mel}, \Delta, \Delta^2]$	TAM	5	4	–	50	100	53.09 %
$[\epsilon, \mathbf{c}_{mel}, \Delta, \Delta^2]$	TAM	7	4	–	25	100	56.94 %
$[\epsilon, \mathbf{c}_{mel}, \Delta, \Delta^2]$	TA-Q	7	4	–	25	100	54.41 %
$[\epsilon, \mathbf{c}_{mel}, \Delta, \Delta^2]$	TA-v	5	4	–	10	25	50.69 %
$[\epsilon, \mathbf{a}]$	TA-v	7	4	–	10	25	50.74 %
$[\epsilon, \mathbf{c}_{mel}, \Delta, \Delta^2]$	TA*-v	7	4	–	25	100	52.58 %

Tabla 4.6. Resumen de los mejores resultados obtenidos para la estimación de estructuras acentuales con modelos ocultos de Markov. En las columnas se indican: MA: modelos acústicos elementales; $\lvert\mathcal{Q}\rvert$: estados por modelo; N_c: cantidad de gaussianas en la mezcla; GP: grado de los polinomios con que se obtuvo el resultado; T_d: paso en la ventana de análisis en ms; T_ω: ancho en la ventana de análisis en ms; Rendimiento: medido como EA correctamente estimadas en relación a las obtenidas desde la transcripción mediante reglas ortográficas (salvo en el caso TA*-v, donde se cuentan directamente las tonicidades silábicas). En relación con el procesamiento se ha simplificado la notación suprimiendo la t de tiempo: ϵ: energía; F_0: frecuencia fundamental; Δ: coeficientes delta; Δ^2: coeficientes de aceleración; dif: diferencia por ajuste con polinomios de grado 3 a 15; $\mathbf{c}_{mel}$: vector de coeficientes cepstrales en escala de mel; $\mathbf{a}$: vector de coeficientes de predicción lineal; Para los modelos acústicos se ha abreviado: TAM: modelos /T/, /A/ y /M/; TA-Q: modelos /T/ y /A/ con palabras inacentuadas; TA-v: modelos /T/ y /A/ por cada vocal y diptongo; TA*-v: tonicidades silábicas por separado (sin formar EA);

se encuentra muy lejos de estos. Hay que considerar que los resultados del modelo TA-Q proveen más información acerca de las SEA, ya que no sólo se clasifican los monosílabos como tales sino que además se explicita su tonicidad silábica /A/ o /T/ y se contemplan las palabras inacentuadas. En este mismo sentido, el resultado obtenido con los modelos TA-v también proporciona más información útil para una etapa posterior ya que se modelan por separado los diferentes núcleos vocálicos. Sin embargo, en este punto se mezcla en parte el MA tradicional a nivel fonético con el nuevo nivel suprasegmental que se incorpora en este trabajo. En el próximo capítulo se describirá un método que permite incorporar esta información en un sistema de RAH basado en MOM.

Capítulo 5

Reconocimiento del habla con penalización prosódica

Hacia el presente capítulo convergen todos los anteriores dado que es aquí donde se integran los mejores hallazgos en un sistema completo para el reconocimiento automático del habla continua. Los estudios acerca de la naturaleza de la prosodia y la acentuación en el habla continua, detallados en el Capítulo 3, sirvieron de base para que en el Capítulo 4 se diseñe un sistema automático para la estimación de las secuencias de estructuras acentuales de una frase a partir de la emisión de voz. En la primera parte del capítulo se describe con mayor detalle el sistema que se ha utilizado como referencia para la comparación. Este sistema se basa en los desarrollos formales del Capítulo 2 y es probado mediante validación cruzada para estimar las tasas de error. A continuación se hace un análisis de los intervalos de confianza para estas estimaciones y se describen los principios básicos para comparar dos sistemas de reconocimiento. En la segunda parte del capítulo se propone un método para la incorporación de información prosódica a través de los modelos de lenguaje de un reconocedor estándar. El método de los modelos de lenguaje variantes en el tiempo y su implementación práctica a través de modelos de lenguaje con red expandida se describe detalladamente y se realiza un análisis de la influencia de cada una de las constantes que controlan su funcionamiento. En las últimas secciones se presentan y discuten los resultados finales, incluyendo varios experimentos que permiten obtener una mejor idea de los alcances del método.

5.1. Sistema de referencia

La referencia para los experimentos se estableció mediante un sistema de reconocimiento automático del habla (RAH) basado en modelos ocultos de Markov (MOM). Los fundamentos de la estructura, entrenamiento y utilización de los MOM en RAH han sido tratados en el Capítulo 2. En esta sección se especificará la configuración utilizada y los resultados obtenidos, que sirvieron como punto de comparación para los restantes experimentos.

5.1.1. Procesamiento de la señal

El procesamiento de la señal da como resultado un vector de características $\mathbf{x}_t$ consistente en: coeficientes cepstrales en escala de mel (CCEM, Sección 2.1.4) con coeficientes de energía y delta (Sección 2.1.5). Todos los coeficientes se obtienen a partir de un análisis por tramos con $T_d = 10$ ms y $T_\omega = 25$ ms. Dado que las señales han sido muestreadas a razón de $1/T_v = 8000$ Hz, los tramos de voz constan de $N_\omega = 200$ muestras y están solapados en $N_d - N_\omega = 120$ muestras.

Antes de transformar cada tramo de la señal se realizan algunos procesos simples, como la eliminación de la media temporal y el filtrado de preénfasis. Para la eliminación de la media temporal se calcula $\mu_t = \sum_{n=1}^{N_\omega} v(t;n)$ y luego se resta a cada $v(t;n)$. El filtrado de preénfasis se aplica para aplanar el espectro de la señal de habla, que típicamente se ve afectado por una caída de la magnitud con la frecuencia que responde a los efectos de radiación en los labios y la glotis. Adicionalmente, este efecto de "blanqueado" espectral también previene la inestabilidad numérica en posteriores etapas de procesamiento. El filtro de preénfasis es de tipo pasa alto y tiene una estructura $1 - a_1 z^{-1}$, siendo $a_1 = 0{,}97$ para este sistema de referencia.

Después de realizar estos procesos previos, a cada tramo de voz se le aplica una ventana de Hamming, (tal como se definió en la Sección 2.1.1). El primer paso para el cálculo de los CCEM es una transformada rápida de Fourier, donde se completa el tramo de voz con ceros hasta obtener 256 muestras. A continuación se integra con ventanas de Bartlett según 24 bandas en escala de mel. Para obtener los CCEM se aplica la transformada coseno y se utilizan los primeros 12 coeficientes resultantes (como define la ecuación (2.7), página 76). Después del agregado del coeficiente de energía normalizado por cada frase y de los coeficientes delta, el vector de características queda compuesto por $N_x = 26$ coeficientes: $\mathbf{x}_t = [\epsilon(t), \Delta\epsilon(t), \mathbf{c}_{mel}(t), \Delta\mathbf{c}_{mel}(t)]$.

5.1.2. Modelado acústico

El modelo acústico (MA) se define en base a 24 MOM semicontinuos (MOMSC), uno para cada uno de los fonemas básicos:

/a/	/b/	/ĉ/	/d/	/e/	/f/	/g/	/i/
/x/	/k/	/l/	/λ/	/m/	/n/	/ñ/	/o/
/p/	/r/	/r̄/	/s/	/t/	/u/	/y/	/θ/

Cada uno de estos fonemas se modeló con 3 estados. Las transiciones permitidas van del estado 1 al 2 y al 3, del estado 2 al 3 y desde cada estado hacia sí mismo (estructura similar a la de la Figura 1.22 de la página 46). En cada estado se modelan las observaciones con distribuciones continuas de probabilidad en $\mathbb{R}^{26}$. Para cada estado se simplifica la parametrización de las gaussianas con un vector de medias $\boldsymbol{\mu}_{jk} \in \mathbb{R}^{26}$ y los 26 elementos de la diagonal principal de $\mathbf{U}_{jk}$. Con esta misma estructura se incorporó un modelo para los silencios y con el segundo estado de este modelo entrenado se construyó un modelo de pausa corta para agregar al final de todas las palabras.

5.1.3. Modelos de lenguaje

El modelo de lenguaje (ML) utilizado fue una bi-gramática con probabilidades estimadas mediante el método de *back-off*, ya descrito en el Capítulo 2. La Figura 5.1 muestra un ejemplo del modelo de lenguaje con red recursiva (MLRR).

Algunos arcos y sus probabilidades asociadas se obtienen directamente desde el corpus de entrenamiento, por simples cuentas. Los arcos relacionados con el modelo de silencio son comúnmente incluidos para otorgar mayor flexibilidad al reconocedor y poder incluir situaciones naturales en el lenguaje hablado. En esta red, los arcos que unen directamente una palabra con el modelo de silencio se corresponden con aquellas palabras que han sido encontradas al principio o al final de una frase. Por último es importante mencionar a los arcos para el suavizado de la gramática. Estos arcos unen a todos los modelos a través de un nodo nulo. El nodo nulo —indicado con un círculo vacío en la Figura 5.1— no posee relación con ningún MA y se utiliza para simplificar la representación del suavizado de la gramática en la red. Estos pasajes, aunque generalmente con probabilidades más pequeñas, permiten que cualquier secuencia de palabras no presente en el corpus de entrenamiento pueda ser reconocida como parte de una frase.

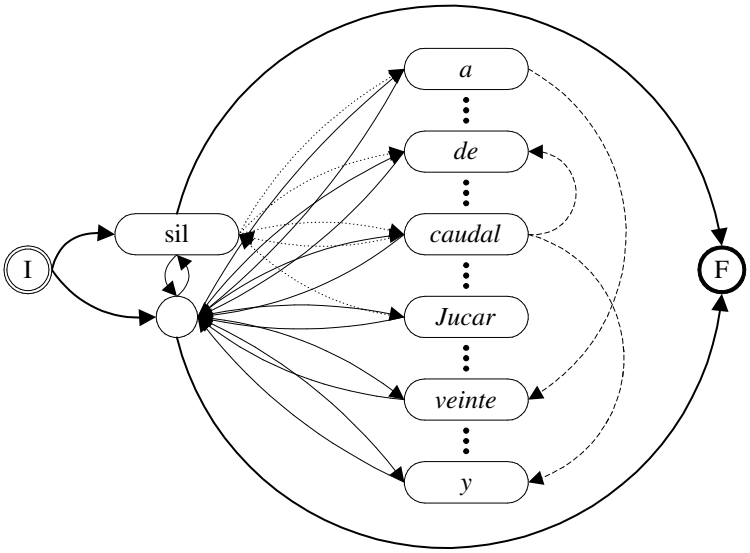

Figura 5.1. Modelo de lenguaje con red recursiva para una bi-gramática. En esta figura se pueden observar los diferentes tipos de arcos que posee el modelo. En líneas de trazo se indican los arcos que relacionan una secuencia de dos palabras que se encontró en el corpus de entrenamiento. El nodo nulo se indica con un círculo vacío y en él convergen los arcos relacionados con el suavizado de la gramática. En líneas de punto se distinguen los arcos relacionados con el modelo de silencio. Los que surgen del nodo "sil" son aquellos que van hacia palabras que se encontraron al principio de alguna frase en el corpus de entrenamiento. Los que llegan al nodo "sil" tienen que ver con palabras que estaban al final de una frase. Con líneas continuas más gruesas se indican los arcos que se relacionan con los nodos de inicio y fin de frase.

5.1.4. Entrenamiento

Tanto el entrenamiento de los MA como la estimación de las probabilidades del ML siguen los métodos descritos en el Capítulo 2. En esta sección se presentarán algunos detalles de índole práctica relacionados con el entrenamiento del sistema de referencia.

Los modelos para cada fonema y el modelo de silencio son inicializados realizando una estimación de las probabilidades sobre toda la base de datos, sin considerar las transcripciones. Luego, a partir de las transcripciones, se construye los modelos compuestos (MC) de cada una de las frases del corpus de entrenamiento. Con estos MC se realizan 3 reestimaciones de todas las probabilidades para el corpus de entrenamiento mediante el algoritmo de Baum-Welch (Sección 2.2.7).

A partir de los parámetros del estado central del modelo de silencio se construye el modelo de pausa corta, que se concatena al final de todas las palabras como una transición opcional. Después de esta modificación se realizan 2 nuevas reestimaciones con todas las frases del corpus de entrenamiento.

A continuación se enlazan los parámetros y se construyen los MOMSC. En este sistema de referencia se comparten $N_c = 200$ las gaussianas para los estados de cada modelo y las del estado 2 del modelo de silencio con las del estado único de la pausa corta. Finalmente se reestiman los modelos definitivos recorriendo 8 veces más el corpus de entrenamiento.

5.1.5. Métodos de validación

Los resultados de referencia se obtuvieron con el subconjunto frases SC1, extraídas del corpus de habla Albayzin, ya citado con anterioridad (Apéndice A.2). Si se utiliza una única partición de entrenamiento y prueba se pueden introducir sesgos en la estimación del error de reconocimiento. Estos sesgos, a favor o en contra, pueden ser ocasionados por la particular selección de las frases en cada uno de los conjuntos. Para evitar estos problemas, todas las pruebas se realizaron por validación cruzada según el método denominado "dejar k afuera promediado" (del inglés *averaged leave-k-out*) [Michie et al., 1994].

Las 600 frases se separaron al azar en 10 particiones de entrenamiento y prueba. En cada partición se utilizaron 481 frases de entrenamiento y se dejaron las restantes 119 para la prueba. A partir de las frases de los conjuntos de entrenamiento se estimaron los parámetros para 10 juegos de MA y 10 ML. Luego se probaron los 10 sistemas de reconocimiento con las

Partición	Palabras de prueba
1	1193
2	1140
3	1079
4	1068
5	1148
6	1073
7	1137
8	1115
9	1108
10	1058
Total	11119

Tabla 5.1. Cantidad de palabras por conjunto de prueba.

frases de sus respectivos conjuntos de prueba. La cantidad de palabras en cada uno de estos conjuntos se muestra en la Tabla 5.1.
Para evaluar el rendimiento de los reconocedores se utilizaron 3 medidas:

- Tasa de palabras reconocidas correctamente: en esta medida se considera la cantidad de palabras que han sido eliminadas ($E_{\mathcal{P}}$) o sustituidas ($S_{\mathcal{P}}$), en relación al total ($T_{\mathcal{P}}$) de palabras consideradas,

$$c_{\mathcal{P}} = \frac{T_{\mathcal{P}} - E_{\mathcal{P}} - S_{\mathcal{P}}}{T_{\mathcal{P}}} = 1 - \varepsilon_{\mathcal{P}}$$

  En base a esta medida se puede definir la tasa de *error* en el reconocimiento de palabras $\varepsilon_{\mathcal{P}}$, que es también conocida en inglés como *word error rate*.

- Tasa de palabras reconocidas considerando las inserciones: esta media es algo más completa ya que incluye los errores por inserción ($I_{\mathcal{P}}$),

$$c_{\mathcal{I}} = \frac{T_{\mathcal{P}} - E_{\mathcal{P}} - S_{\mathcal{P}} - I_{\mathcal{P}}}{T_{\mathcal{P}}} = 1 - \varepsilon_{\mathcal{I}}$$

  taza que en inglés es conocida como *word accuracy* y a partir de la cual se puede definir la tasa de error de reconocimiento de palabras con inserciones ($\varepsilon_{\mathcal{I}}$). Como puede apreciarse, esta tasa puede tomar valores negativos. Sin embargo, para reconocedores con bajas tasas de error se observa en general que $\varepsilon_{\mathcal{I}} \approx \varepsilon_{\mathcal{P}}$.

- Tasa de frases reconocidas correctamente: donde se consideran las frases en su totalidad, es decir, la cantidad de frases en las que no existe ningún error,

$$c_{\mathcal{F}} = \frac{T_{\mathcal{F}} - S_{\mathcal{F}}}{T_{\mathcal{F}}} = 1 - \varepsilon_{\mathcal{F}}$$

también se define la tasa de error de reconocimiento de frases $\varepsilon_{\mathcal{F}}$.

5.1.6. Resultados de referencia

En la Tabla 5.2 se muestran los resultados de reconocimiento para cada partición en el sistema de referencia. En la Tabla 5.3 se resumen los errores promedio que servirán de referencia en el resto del capítulo.

Partición	$c^r_{\mathcal{P}}$ %	$c^r_{\mathcal{I}}$ %	$c^r_{\mathcal{F}}$ %
1	91.62	90.95	54.62
2	92.11	90.79	55.46
3	91.94	90.55	57.14
4	94.01	93.73	67.80
5	94.16	93.38	63.03
6	91.71	90.21	58.47
7	91.73	91.20	62.18
8	92.02	90.76	63.87
9	94.04	93.05	66.39
10	91.30	90.08	68.07

Tabla 5.2. Resultados de reconocimiento para las 10 particiones con que se probó el sistema de referencia. En la primera columna se muestran las tasas de palabras reconocidas correctamente, en la segunda las tasas de palabras reconocidas considerando las inserciones y en la tercera las tasas de frases bien reconocidas.

	mín	máx	μ	σ
$\varepsilon^r_{\mathcal{P}}$ %	5.84	8.70	7.54	1.06
$\varepsilon^r_{\mathcal{I}}$ %	6.27	9.92	8.53	1.17
$\varepsilon^r_{\mathcal{F}}$ %	31.93	45.38	38.30	2.24

Tabla 5.3. Errores de reconocimiento para el sistema de referencia. En la primera columna se presenta el mínimo error de todas las particiones, en la segunda el máximo, luego el error promedio y finalmente la desviación estándar.

Análisis de los intervalos de confianza

El error $\varepsilon^r_{\mathcal{P}}$ puede interpretarse como una estimación de la *probabilidad* de reconocer incorrectamente una palabra p^r_ε. Esta estimación puede ser mejor o peor de acuerdo, principalmente, a la cantidad de experimentos realizados. Es por esto que resulta interesante poder calcular el intervalo dentro del que se encuentra la verdadera probabilidad de error para un determinado porcentaje de confianza.

Para este cálculo hay que considerar que se poseen tantos ejemplos como posibles oportunidades de generarse un error existan. En este caso se deben considerar las $n_{\mathcal{P}} = 11119$ palabras con que se realizaron las pruebas (Tabla 5.1).

La distribución de probabilidad de errores de reconocimiento para pruebas con más de 1000 palabras puede ser aproximada mediante distribuciones gaussianas [Torre-Vega, 1999, Apéndice B]. Asumiendo la independencia estadística de los experimentos, se pueden calcular los intervalos de confianza para el error de referencia $\varepsilon^r_{\mathcal{P}} = 0{,}0754$:

$$p^r_\varepsilon \in \begin{cases} [0{,}072; 0{,}078] & \text{con } 80{,}0\,\% \text{ de confianza} \\ [0{,}071; 0{,}079] & \text{con } 90{,}0\,\% \text{ de confianza} \\ [0{,}070; 0{,}080] & \text{con } 95{,}0\,\% \text{ de confianza} \\ [0{,}069; 0{,}081] & \text{con } 99{,}0\,\% \text{ de confianza} \\ [0{,}067; 0{,}083] & \text{con } 99{,}9\,\% \text{ de confianza} \end{cases}$$

5.1.7. Comparación de reconocedores

Para comparar los resultados de referencia con los obtenidos en los distintos experimentos se utilizaron diferentes indicadores que miden la reducción de las tasas de error. Suponiendo que ε es una de las tasas de error en el sistema en evaluación y ε^r la misma tasa en el sistema de referencia, se definen 2 medidas de reducción de la tasa error:

- Reducción absoluta de la tasa de error:

$$\Delta\varepsilon = \varepsilon^r - \varepsilon$$

- Tasa de reducción relativa de la tasa de error:

$$\delta\varepsilon = \frac{\varepsilon^r - \varepsilon}{\varepsilon^r}$$

Esta última medida brinda una idea más acabada de las mejoras obtenidas: al especificar $\delta\varepsilon$ ya no es necesario aclarar cuál fue la referencia ε^r. Sin embargo, $\delta\varepsilon$ está calculada a partir de estimaciones de la probabilidad de error y no a partir de las verdaderas probabilidades de error. En el trabajo de [Torre-Vega, 1999] se describe un método para calcular la probabilidad de que el error obtenido sea mayor al error de referencia $\Pr(\varepsilon < \varepsilon^r)$. En base a suposiciones como la independencia estadística de los experimentos y la distribución gaussiana de las probabilidades de error, se demuestra que:

$$\Pr(\varepsilon < \varepsilon^r) = \int_{-\infty}^{\Gamma} \mathcal{N}(x)dx$$

con:

$$\Gamma = \frac{\sqrt{2}(\varepsilon^r - \varepsilon)}{\sqrt{\varepsilon^r(1-\varepsilon^r)/n_{\mathcal{P}}} + \sqrt{\varepsilon(1-\varepsilon)/n_{\mathcal{P}}}}$$

Resolviendo estas ecuaciones para el sistema de referencia descrito y con un ε estimado en las mismas condiciones, se puede encontrar que:

- Para alcanzar una $\Pr(\varepsilon < \varepsilon^r) > 95\,\%$ se requiere:

$$\begin{aligned} \varepsilon &< 6{,}96\,\% \\ \Delta\varepsilon &> 0{,}58\,\% \\ \delta\varepsilon &> 7{,}75\,\% \end{aligned}$$

- Para alcanzar una $\Pr(\varepsilon < \varepsilon^r) > 99{,}99\,\%$ se requiere:

$$\begin{aligned} \varepsilon &< 6{,}22\,\% \\ \Delta\varepsilon &> 1{,}32\,\% \\ \delta\varepsilon &> 17{,}47\,\% \end{aligned}$$

- Para alcanzar una $\Pr(\varepsilon < \varepsilon^r) > 99{,}9999\,\%$ se requiere:

$$\begin{aligned} \varepsilon &< 5{,}85\,\% \\ \Delta\varepsilon &> 1{,}69\,\% \\ \delta\varepsilon &> 22{,}35\,\% \end{aligned}$$

5.2. Penalización prosódico acentual

5.2.1. Modelos de lenguaje variantes en el tiempo

En trabajos anteriores se ha incorporado información adicional a un sistema de RAH en una etapa posterior al proceso de reconocimiento. Por ejemplo, en [Nöth et al., 2000] se incorporó información prosódica modificando las probabilidades de la red de hipótesis de palabras, salida de un reconocedor basado en MOM. Otros antecedentes han sido detallados más extensamente en la Sección 1.5.2 (página 65).

No es usual la incorporación de información extra en etapas previas o durante el mismo proceso de reconocimiento. El desarrollo teórico de esta propuesta integra, a través del ML, información que cambia en el tiempo durante el proceso de reconocimiento de cada frase del corpus. La principal idea de los modelos de lenguaje variantes en el tiempo (MLVT) es modificar un ML de referencia a medida que el tiempo avanza durante el proceso de reconocimiento de una frase. Con esto en mente, supongamos que el reconocedor se encuentra en medio de una búsqueda y que una de las hipótesis acústicamente plausible está dada por:

$$\mathbf{h}_{i_1}^n = w_{i_1-1}, w_{i_1-2}, \ldots, w_{i_1-n+1}$$

con una probabilidad de transición $\hat{p}(w_{i_1}|\mathbf{h}_{i_1}^n)$ hacia la siguiente palabra w_{i_1}. En un instante de tiempo posterior, otra hipótesis acústicamente plausible podría ser:

$$\mathbf{h}_{i_2}^n = w_{i_2-1}, w_{i_2-2}, \ldots, w_{i_2-n+1}$$

con una probabilidad $\hat{p}(w_{i_2}|\mathbf{h}_{i_2}^n)$ para la transición hacia la próxima palabra w_{i_2}. Para un n fijo, como generalmente sucede en los sistemas de reconocimiento actuales, se tiene:

$$\mathbf{h}_{i_1}^n = \mathbf{h}_{i_2}^n \wedge w_{i_1} = w_{i_2} \Rightarrow \hat{p}(w_{i_1}|\mathbf{h}_{i_1}^n) = \hat{p}(w_{i_2}|\mathbf{h}_{i_2}^n)$$

es decir, para iguales historias en diferentes posiciones dentro de una frase, corresponden iguales probabilidades de transición entre palabras. Sin embargo, podrían existir otras evidencias indicando que dados dos tiempos de

análisis diferentes en la frase (i_1 e i_2) se requiera $\hat{p}(w_{i_1}|\mathbf{h}_{i_1}^n) \neq \hat{p}(w_{i_2}|\mathbf{h}_{i_2}^n)$. Por ejemplo, cuando $n = 2$, esto sería $\mathbf{h}_{i_1}^1 = w_{i_1-1}$ y $\mathbf{h}_{i_2}^1 = w_{i_2-1}$, con probabilidades de bi-gramática $\hat{p}(w_{i_1}|w_{i_1-1})$ y $\hat{p}(w_{i_2}|w_{i_2-1})$. Obviamente si $w_{i_1} = w_{i_2}$ y $w_{i_1-1} = w_{i_2-1}$, la probabilidad del ML es independiente de la posición de la palabra en la frase: $\hat{p}(w_{i_1}|w_{i_1-1}) = \hat{p}(w_{i_2}|w_{i_2-1})$.

Para los MLVT, la idea es permitir que esta probabilidad sea adaptada en diferentes momentos del proceso de reconocimiento de una frase y para las diferentes frases a reconocer. Para adaptar las probabilidades del ML durante el reconocimiento se propone la incorporación de una función de penalización:

$$\hat{p}_t(w_i|\mathbf{h}_i^n) = \varphi_i(w_i, \mathbf{h}_i^n, E_t)\hat{p}(w_i|\mathbf{h}_i^n)$$

donde E_t representa cualquier información Extra para el tiempo t de la frase que está siendo reconocida. La función φ_i genera un valor numérico en el rango real [0,1]. Esta función reduce la probabilidad del ML de referencia cuando la evidencia E no sea favorable a la transición de palabra hipotética en el tiempo t.

Un ejemplo puede ser útil para terminar de clarificar estas ideas. Consideremos dos palabras consecutivas en un ML de bi-gramática. Supongamos que estamos interesados en la probabilidad de que la próxima palabra sea *río* siendo que estamos actualmente en el final (acústicamente más probable) de la palabra *el*. En la corpus de habla Albayzin se pueden encontrar los siguientes dos casos:

1. *El río Ebro, ¿pasa por la Comunidad Autónoma de Navarra?*

2. *Mar donde desemboca el río Pisuerga.*

Supongamos que la frase 1 es la que realmente ha pronunciado el locutor. La secuencia de estructuras acentuales correcta para esta frase es: /A TA TA TA A A AAAT ATAA A ATA/. Al introducir esta información a nivel del ML evidentemente la probabilidad de transición entre *el* y *río* no deberá ser la misma si *el* es la primera palabra y *río* la segunda, que si *el* es la cuarta y *río* la quinta palabra de la frase. Esto se sigue del hecho de que hay distintas estructuras acentuales en las diferentes posiciones, es decir, la evidencia E es diferente en cada posición de la frase. Así, la función $\varphi_i(w_i, \mathbf{h}_i^n, E_t)$ deberá reducir la probabilidad de transición entre *el* y *río*

cuando se evalúe como hipótesis de búsqueda a la frase 2 y no deberá cambiarla cuando la hipótesis en la búsqueda sea la frase 1 (o cualquier otra que tenga a *el* como primer palabra y *río* como segunda).

5.2.2. Modelos de lenguaje con red expandida

Como se puede ver en la Figura 5.1, usando un MLRR para representar gramáticas no es posible cambiar las probabilidades de transición de acuerdo a la posición de las palabras dentro de la frase. En estas redes, la probabilidad de transición entre dos palabras dadas no depende de la posición de las palabras en la frase. Es por esto que se requiere una nueva estructura de red para poder implementar los MLVT.

Para incorporar la función de penalización φ_i directamente en el algoritmo de decodificación de un reconocedor basado en MOM, se propone utilizar una estructura alternativa denominada modelo de lenguaje con red expandida (MLRE). En base a un ML de bi-gramática, para permitir que $\hat{p}(w_{i_1}|w_{i_1-1}) \neq \hat{p}(w_{i_2}|w_{i_2-1})$ cuando $i_1 \neq i_2$ se puede usar un autómata no recursivo (probabilístico y de estados finitos), en lugar del caso recursivo de la Figura 5.1. Formalmente esta gramática no es una bi-gramática pero se pueden tomar algunas precauciones para que sea funcionalmente equivalente. En la Figura 5.2, se muestra una representación simplificada de un MLRE en el que solamente se permiten conexiones mediante arcos hacia delante, salvo en la última capa, donde también se pueden realizar transiciones hacia atrás. En principio, la red resultante deberá tener tantas capas como la máxima cantidad admitida de palabras por frase a reconocer. Sin embargo, las realimentaciones en la última capa permitirían reconocer frases más largas si la evidencia acústica fuera favorable.

Para construir el MLRE en primer lugar se estima el MLRR mediante el método de *back-off*. Luego se realiza la "expansión" de gramática de forma que $\hat{p}_\ell(w_i|w_j) = \hat{P}(w_i|w_j)$ para todas las capas ℓ en la Figura 5.2. Dada una frase, cada transición de una palabra a la siguiente corresponde a una capa de probabilidades en un MLRE (y a un bucle en el MLRR). Sin embargo, con la red expandida cada transición dentro de la frase puede ser modificada independientemente en relación con la posición de las palabras en la frase. Ahora, por ejemplo, la probabilidad asociada para la transición de una palabra en la primera capa (primera palabra de la frase) hacia una palabra en la segunda capa (segunda palabra en la frase) puede ser diferente a la probabilidad de transición entre las mismas palabras en las capas 3 y 4.

Hay que observar que el MLRE de la Figura 5.2 es más complejo que el MLRR de la Figura 5.1. Sin embargo, el segundo modelo no es recursivo

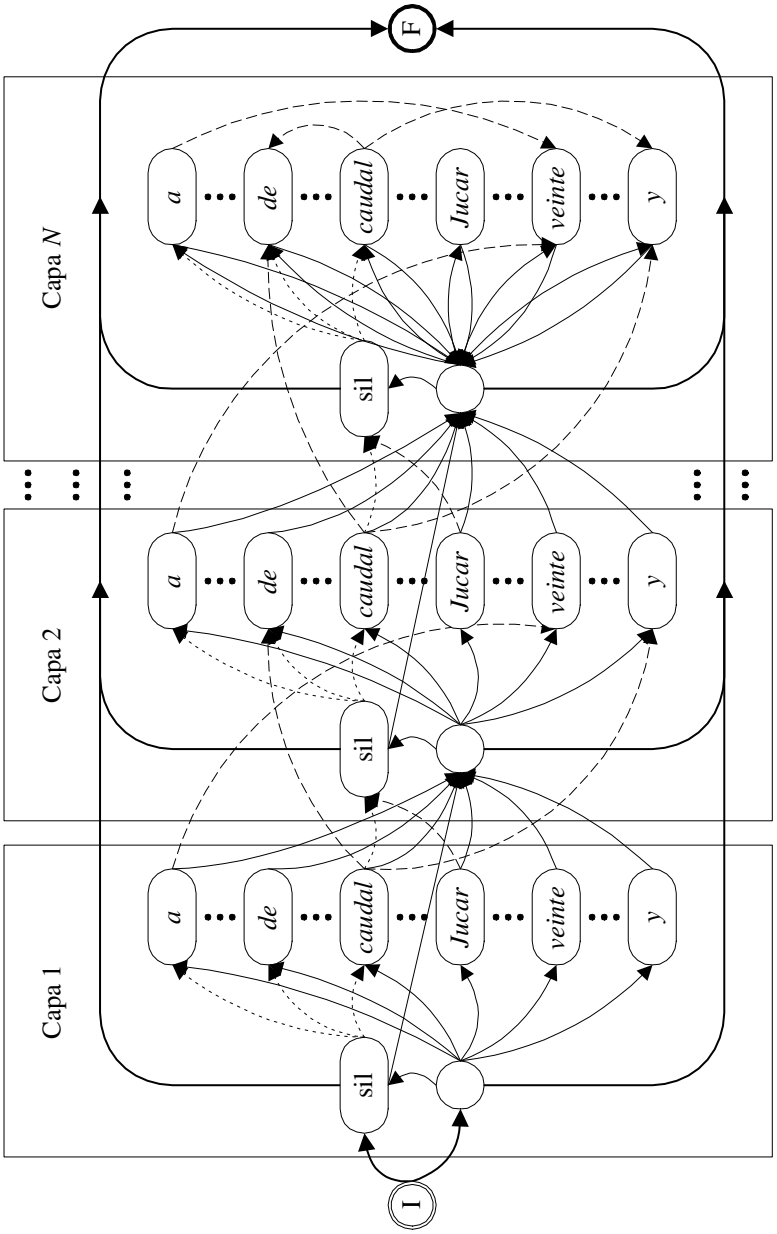

Figura 5.2. Modelo de lenguaje con red expandida para una bi-gramática. Al igual que en la Figura 5.1, se pueden observar los diferentes tipos de arcos que posee el modelo. En líneas de trazo se indican los arcos que relacionan una secuencia de dos palabras que se encontró en el corpus de entrenamiento. Con líneas continuas se indican los arcos del suavizado de gramática. En líneas de punto se distinguen los arcos relacionados con el modelo de silencio. Con líneas continuas más gruesas se indican los arcos que se relacionan con los nodos de inicio y fin de frase.

y cada capa del MLRE se corresponde con una conexión hacia atrás en el MLRR. La propiedad recursiva del MLRR estándar ha sido sustituida por la repetición de capas idénticas. De esta forma, es una tarea sencilla obtener la versión expandida de una red de bi-gramática.

Este método reduce la complejidad de implementación de un MLVT y provee la flexibilidad necesaria para realizar experimentos de laboratorio. Sin embargo, cuando el MLRE es utilizado en RAH para simular MLVT, el reconocimiento debe hacerse en dos etapas: primero el ML se modifica para la frase que se va a reconocer (se expande hacia un MLRE y se penaliza) y luego se realiza un reconocimiento estándar con la red adaptada.

5.2.3. Secuencias de estructuras acentuales y penalización

En las ecuaciones (4.16) y (4.17) se definieron los modelos de tonicidad silábica y estructuras acentuales (EA) respectivamente. Estas últimas estructuras, que eran modeladas mediante MOM tal como si se tratara de palabras en el uso más corriente en RAH, pueden verse también como parte de una secuencia en un ML. Si $\mathcal{A} = a_1, a_2, \ldots, a_P$ es el conjunto de las P posibles EA, se puede definir una función de mapeo $g : \mathcal{W} \rightarrow \mathcal{A}$ que asigna a cada palabra $w_k \in \mathcal{W}$ una EA $a_i \in \mathcal{A}$. Dentro del conjunto $\mathcal{A}$ se puede considerar una medida de distancia $\xi(a_i, a_j)$, que asignas valores en $[0, 1]$ a cada par de EA.

Durante la estimación de las secuencia de estructuras acentuales (SEA) se puede obtener:

$$\hat{\mathbf{a}}_t^q = \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_q; \quad \hat{a}_i \in \mathcal{A} \tag{5.1}$$

La EA de una palabra puede ser comparada con la EA estimada y luego aplicar a este camino una penalización proporcional a la distancia entre ambas, $\xi\left(g(w_i), \hat{a}_i\right)$. Por ejemplo, la EA estimada podría ser $\hat{a}_i = /\text{TTA}/$. Por otro lado, la hipótesis que está evaluando el reconocedor en esa posición podría ser $g(w_i) = g(/\text{estable}/) = /\text{ATA}/$. La penalización a introducir debería estar basada en la distancia $\xi\left(/\text{ATA}/, /\text{TTA}/\right)$.

Sin embargo, existen algunos problemas en la definición de la función de penalización. En primer lugar, los q elementos de la SEA estimada no tiene necesariamente que coincidir con las m palabras de la hipótesis en evaluación (ni mucho menos con las que realmente haya pronunciado el locutor). Además, debe considerarse que pueden ser incorrectas tanto la estimación de la SEA, como la frase reconocida y más aún cualquiera de las

hipótesis que evalúa el reconocedor. Para definir la función de penalización será necesario considerar éstas y otras situaciones particulares:

$$\varphi_i(w_i, \mathbf{h}_i^n, \hat{\mathbf{a}}_l^q) = \begin{cases} \gamma_e & \text{si } i > q \\ (\gamma_s - 1)\,\xi\,(g(w_i), \hat{a}_i) + 1 & \text{si } i = 1 \vee i = m \\ (\gamma_n - 1)\,\xi\,(g(w_i), \hat{a}_i) + 1 & \text{si } \mathcal{C}(w_i, \mathbf{h}_i^n) = 0 \\ (\gamma_w - 1)\,\xi\,(g(w_i), \hat{a}_i) + 1 & \text{si } \mathcal{C}(w_i, \mathbf{h}_i^n) > 0 \end{cases} \tag{5.2}$$

condiciones que se evalúan de forma excluyente, de arriba hacia abajo. Estas expresiones están basadas en una simple regla lineal de la forma $\varphi = (\gamma - 1)\xi + 1$, esto es $\xi = 1 \Rightarrow \varphi = \gamma$ y $\xi = 0 \Rightarrow \varphi = 1$. Para los siguientes experimentos se ha utilizado una medida de distancia basada simplemente en el delta de Kronecker $\xi(a_i, a_j) = 1 - \delta_{i,j}$. Las constantes γ deberán ajustarse de acuerdo al peso que se quiera dar a cada tipo de penalización.

La primera condición en (5.2) considera el caso en que la frase a ser evaluada contenga más palabras que la cantidad de EA en la SEA estimada. Esta penalización se aplicará a todas aquellas transiciones que lleven a las palabras que estén más allá de la finalización de la SEA estimadas. La segunda condición contempla a las palabras relacionadas con un modelo de silencio. Esto es necesario ya que, como se vio en el Capítulo 3, la presencia de un silencio antes o después de la palabra afecta considerablemente sus rasgos prosódicos, más aún en el principio o fin de cada frase. Esto conlleva a una estimación menos confiable de la EA en cuestión.

En tercer lugar, la ecuación (5.2) considera las historias $\mathbf{h}_i^n$ que no se encontraron en el corpus de entrenamiento durante la estimación del MLRR. Estas probabilidades son el resultado del proceso de suavizado de la gramática. La última condición contempla a las transiciones cuyas probabilidades fueron calculadas por simples cuentas en el corpus de entrenamiento [Milone y Rubio, 2003].

5.2.4. Influencia de las constantes de penalización

Para estudiar la influencia de las constantes de penalización en el proceso de reconocimiento se realizó un análisis exhaustivo donde se experimentó con diferentes valores para las 4 constantes. Se obtuvieron los errores $\varepsilon_{\mathcal{P}}$, $\varepsilon_{\mathcal{I}}$ y $\varepsilon_{\mathcal{F}}$ para cada combinación de valores a partir de 10 particiones de entrenamiento y prueba del subconjunto SC2, con 1000 frases del corpus de habla Albayzin (Apéndice A.3). En estos experimentos se penalizaron los MLRE en base a las SEA correctas, obtenidas a partir de las transcripciones

del corpus, siguiendo las reglas que se describieron en las Secciones 1.3.3 y 1.3.3

La combinación de constantes de penalización que dio los mejores resultados fue, en escala logarítmica:

$$\gamma_w = -2 \qquad \gamma_s = -4$$

$$\gamma_n = -4 \qquad \gamma_e = 0$$

Para analizar con mayor detalle la influencia de cada constante de penalización sobre los errores de reconocimiento se promediaron todos los resultados para un valor dado de cada constante y de esta forma se obtuvo el error promedio para este valor de la constante. Por ejemplo, se promediaron todos los resultados que se obtuvieron con $\gamma_s = 0$, sin importar cuanto valieran las restantes contantes, y así se obtuvo el primer punto de la curva para γ_s. Este procedimiento se repitió para todos los valores de la contante y para todas las contantes obteniendo las curvas de la Figura 5.3. Se muestran solamente los $\varepsilon_{\mathcal{P}}$ ya que los restantes errores siguen la misma forma general.

Dado que las curvas de la Figura 5.3 sugieren que para $\gamma_s, \gamma_n < -4$ se sigue reduciendo el error, se realizaron pruebas adicionales haciendo llegar los valores de estas constantes hasta -8. En el caso de γ_s no se obtuvieron mejores resultados. Sin embargo, para el caso de γ_n los errores de reconocimiento siguieron bajando lo que se explica de la siguiente forma: dado que los γ_n corresponden al suavizado de la gramática y dado que los conjuntos de prueba y entrenamiento poseen una estructura gramatical muy similar, al reducir los γ_n se da preponderancia a las probabilidades que unen palabras que se encontraban contiguas en el conjunto de entrenamiento y la gramática se ajusta mejor a las condiciones en que es probada. Sin embargo, no es beneficioso reducir excesivamente los γ_n ya que se le quita al reconocedor la capacidad de adaptarse a estructuras gramaticales no contempladas durante el entrenamiento.

Para tener una mejor idea de los beneficios que pueden obtenerse por el simple hecho de introducir las penalizaciones pero sin considerar las SEA o por el solo hecho de no considerar el suavizado de la gramática, se realizaron tres experimentos adicionales:

1. Penalizaciones fijas e independientes: se introdujeron las penalizaciones en todas los arcos del MLRE sin importar las SEA. Además, cuando se utilizaba una penalización se dejaban las restantes en cero. De esta forma se obtuvieron las siguientes mejoras relativas:

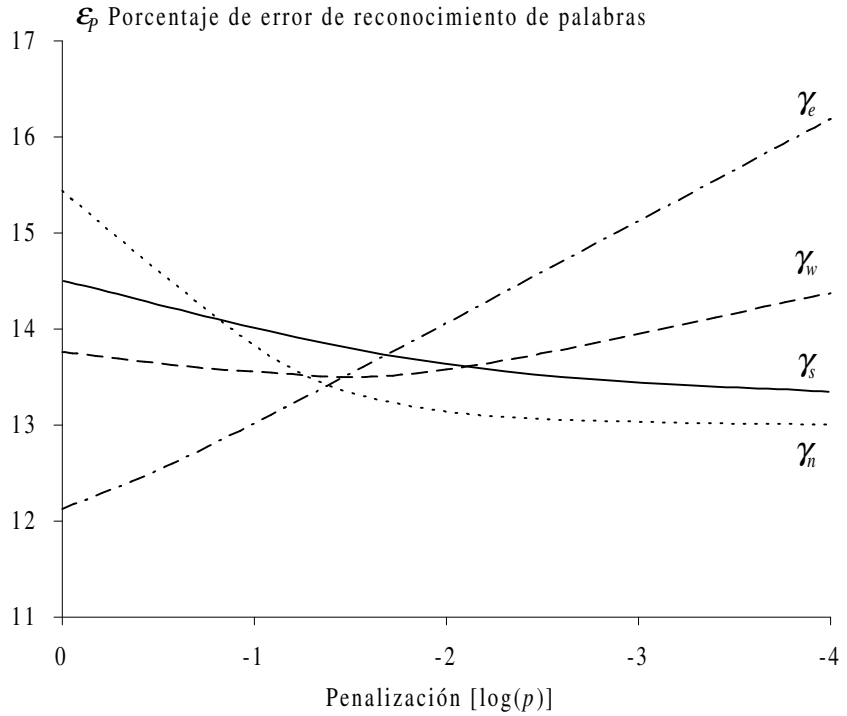

Figura 5.3. Influencia de las contantes de penalización prosódico acentual en la tasa de error en reconocimiento de palabras. En línea de trazo y punto se indica el error promedio para distintos valores de la contante relacionada con la extensión de las frases. En de trazos se indica el error promedio para variaciones de la constante que penaliza los arcos entre palabras. En línea continua se observa la influencia en el error de la constante que afecta a todos los arcos en relación con un modelo de silencio. Por último, en línea de puntos se indica el error promedio para diferentes valores de la constante relacionada con el suavizado de gramática.

- Para γ_s fija el mejor $\delta\varepsilon_{\mathcal{P}}$ fue 2.17 %
- Para γ_n fija el mejor $\delta\varepsilon_{\mathcal{P}}$ fue 16.57 %
- Para γ_w fija el mejor $\delta\varepsilon_{\mathcal{P}}$ fue -59.87 %

Como indica la Figura 5.3, no tiene sentido realizar pruebas para γ_e fija ya que no hay mejoras en ningún caso.

2. Penalizaciones al azar: se utilizaron constantes γ con valores aleatorios entre 0 y -4. Los resultados fueron claramente desfavorables, con un $\delta\varepsilon_{\mathcal{P}} = -44{,}69\,\%$.

3. Eliminación del suavizado de gramática: se quitaron completamente las transiciones por suavizado de gramática y se obtuvo una mejora $\delta\varepsilon_{\mathcal{P}} = 15{,}43\,\%$. Cabe aclarar que en este caso la información de las SEA se utilizó para elegir los mejores caminos con las tansiciones restantes.

Otro experimento que cabe mencionar aquí es el de incorporar la frecuencia fundamental (F_0) al vector de características utilizado en el sistema de referencia. Esta idea surge de un primer enfoque para agregar un rasgo prosódico al reconocedor. Debe considerarse que, si bien la energía está presente de forma explicita en el vector de características, la F_0 se elimina en la integración por bandas que se realiza para calcular los CCEM. El vector de características utilizado en estos experimentos fue:

$$\mathbf{x}_t = [\epsilon(t), \Delta\epsilon(t), F_0(t), \Delta F_0(t), \mathbf{c}_{mel}(t), \Delta\mathbf{c}_{mel}(t)]$$

Los resultados para este caso dieron un $\delta\varepsilon_{\mathcal{P}} = -7{,}16\,\%$.

5.3. Resultados

Al igual que en el sistema de referencia, todos los resultados finales se obtuvieron a partir de 10 particiones de entrenamiento y prueba del subconjunto SC1 del corpus de habla Albayzin. En cada partición de entrenamiento se incluyeron 481 frases y las restantes 119 frases se utilizaron para la prueba. Los experimentos se realizaron a partir de las SEA correctas y las SEA estimadas con los modelos TAM y TA-Q detallados en el capítulo anterior y resumidos en la Tabla 4.6 (página 176). El procedimiento general cosiste en obtener los MA y MLRR para cada partición de entrenamiento, expandir cada MLRR a una MLRE, penalizar la MLRE para cada frase de prueba a partir de la SEA y reconocer con esa MLRE penalizada.

5.3.1. Reconocimiento con estructuras acentuales correctas

Es interesante conocer como funcionaría el método de penalización prosódico acentual si se pudiese estimar perfectamente las SEA. Este experimento es fácil de realizar dado que a partir de las transcripciones de las frases del corpus es posible obtener todas las SEA siguiendo las reglas que se detallaron en las Secciones 1.3.3 y 1.3.3.

A partir de las constantes de penalización encontradas en la sección anterior se obtuvieron los resultados detallados en las Tablas 5.4 y 5.5. Si se considera, por ejemplo, el $\varepsilon_{\mathcal{P}}$ promedio en relación al $\varepsilon^r_{\mathcal{P}}$ se puede encontrar que $\Pr(\varepsilon < \varepsilon^r) > 99{,}9999999999998\,\%$ ($\Gamma = 7{,}85$).

5.3.2. Reconocimiento con estructuras acentuales estimadas

Estos resultados se obtuvieron utilizando las SEA estimadas con los modelos TAM y TA-Q. Los experimentos relacionados con los modelos TAM se descartaron rápidamente ya que en ningún caso se logró un $\delta\varepsilon_{\mathcal{P}} > 15\,\%$. En la Tabla 4.6 se mostró que las estimaciones con los modelos TA-Q alcanzaron el 54.41 % de EA bien reconocidas. Estas estimaciones se realizaron con MOM de 7 estados, mezclas de 4 gaussianas y ventanas de análisis de 25 ms de paso y 100 ms de ancho.

A partir de las constantes de penalización que mejores resultados dieron en las pruebas de la sección anterior, se obtuvieron los resultados detallados en las Tablas 5.6 y 5.7. El análisis comparativo de estos resultados con los de referencia se muestra en la Tabla 5.8. Se debe destacar que, a diferencia de las dos primeras tasas de error, para el cálculo de $\Pr(\varepsilon_{\mathcal{F}} < \varepsilon^r_{\mathcal{F}})$ se han utilizado 1190 ejemplos, es decir, 10 particiones de 119 frases cada una.

Partición	$c_{\mathcal{P}}\,\%$	$\delta\varepsilon_{\mathcal{P}}$	$c_{\mathcal{I}}\,\%$	$\delta\varepsilon_{\mathcal{I}}$	$c_{\mathcal{F}}\,\%$	$\delta\varepsilon_{\mathcal{F}}$
1	94.72	36.99	94.13	35.14	68.07	29.64
2	95.18	38.91	94.30	38.11	75.63	45.29
3	94.90	36.72	93.42	30.37	74.79	41.18
4	96.63	43.74	96.07	37.32	81.36	42.11
5	96.34	37.33	95.56	32.93	71.43	22.72
6	94.59	34.74	93.38	32.38	68.64	24.49
7	94.20	29.87	93.58	27.05	69.75	20.02
8	94.62	32.58	93.81	33.01	76.47	34.87
9	96.84	46.98	95.94	41.58	78.15	34.99
10	94.33	34.83	92.82	27.62	80.67	39.46

Tabla 5.4. Resultados de reconocimiento para cada partición utilizando las estructuras acentuales correctas. Los resultados están separados en tres grupos (reconocimiento de palabras, palabras con inserciones y frases) y en cada uno de ellos se presenta la tasa de reconocimiento acompañada por la tasa de reducción relativa de la tasa de error correspondiente.

	mín	máx	μ	σ	δ
$\varepsilon_{\mathcal{P}}\,\%$	3.16	5.80	4.76	1.02	36.87
$\varepsilon_{\mathcal{I}}\,\%$	3.93	7.18	5.70	1.20	33.18
$\varepsilon_{\mathcal{F}}\,\%$	18.64	31.93	25.50	5.00	33.42

Tabla 5.5. Errores de reconocimiento utilizando las estructuras acentuales correctas. En la primera columna se presenta el mínimo error de todas las particiones, en la segunda el máximo, luego el error promedio y la desviación estándar. En la última columna se presenta la tasa de reducción relativa de la tasa de error.

Partición	$c_{\mathcal{P}}\,\%$	$\delta\varepsilon_{\mathcal{P}}$	$c_{\mathcal{I}}\,\%$	$\delta\varepsilon_{\mathcal{I}}$	$c_{\mathcal{F}}\,\%$	$\delta\varepsilon_{\mathcal{F}}$
1	93.10	17.66	91.30	3.87	73.11	40.74
2	94.22	26.74	93.13	25.41	63.87	18.88
3	95.00	37.97	93.25	28.57	69.75	29.42
4	94.35	5.68	92.59	-18.18	68.91	03.45
5	95.69	26.20	94.76	20.85	72.88	26.64
6	96.08	52.71	94.95	48.42	68.91	25.14
7	93.66	23.34	92.26	12.05	66.95	12.61
8	94.02	25.06	93.23	26.73	65.55	4.65
9	94.26	3.69	92.83	-3.17	71.43	15.00
10	96.03	54.37	95.04	50.00	73.11	15.78

Tabla 5.6. Resultados de reconocimiento para cada partición utilizando las estructuras acentuales estimadas mediante modelos ocultos de Markov. Los resultados se han separado en tres grupos (reconocimiento de palabras, palabras con inserciones y frases) y en cada uno de ellos se presenta la tasa de reconocimiento acompañada por la tasa de reducción relativa de la tasa de error correspondiente.

	mín	máx	μ	σ	δ
$\varepsilon_{\mathcal{P}}\,\%$	3.92	6.90	5.36	1.07	28.91
$\varepsilon_{\mathcal{I}}\,\%$	4.96	8.70	6.67	1.32	21.80
$\varepsilon_{\mathcal{F}}\,\%$	26.89	36.13	30.55	3.42	20.23

Tabla 5.7. Errores de reconocimiento utilizando las estructuras acentuales estimadas mediante modelos ocultos de Markov. En la primera columna se presenta el mínimo error de todas las particiones, en la segunda el máximo, luego el error promedio y la desviación estándar. En la última columna se presenta la tasa de reducción relativa de la tasa de error.

	μ^r	μ	Γ	$\Pr(\varepsilon < \varepsilon^r)$
$\varepsilon_{\mathcal{P}}\,\%$	7.54	5.36	6.16	99.9999999636275
$\varepsilon_{\mathcal{I}}\,\%$	8.53	6.67	4.96	99.9999647534101
$\varepsilon_{\mathcal{F}}\,\%$	38.30	30.55	3.89	99.9949877889003

Tabla 5.8. Análisis comparativo de los errores de reconocimiento. En la primera columna se presentan los errores de referencia y en la segunda los errores obtenidos con el sistema con penalización prosódico acentual utilizando las secuencias de estructuras acentuales estimadas. En la tercera columna se muestran los límites de integración para los cálculos de la última columna, donde se presentan las probabilidades de que el reconocedor propuesto sea mejor que el de referencia.

5.4. Discusión

En la Figura 5.3 se puede ver la forma en que influye cada tipo de penalización en el error de reconocimiento. La penalización para los finales de frase (γ_e), que en principio parece que debería ser beneficiosa, no produce ninguna mejora en los resultados promedio. En cuanto a la penalización para las transiciones que no pertenecen al suavizado de gramática, la curva de error poseen un mínimo que permite elegir fácilmente γ_w. Algo similar sucede con la penalización γ_s, que afecta principalmente a los principios y finales de frase. Esta constante muestra un mínimo cercano a -4. No sucede lo mismo con el error en el caso de la constante γ_n que, aunque muy lentamente, sigue bajando para valores menores a -4. La influencia de esta constante en el error de reconocimiento fue medida independientemente y así se estableció claramente el máximo beneficio que podría obtenerse reduciendo simplemente las probabilidades asociadas con todos los nodos nulos. Aún más, si se eliminan por completo a los nodos nulos y se realiza la penalización con las SEA correctas sobre las restantes transiciones, se pudo ver que el máximo beneficio posible es de aproximadamente $\delta\varepsilon_{\mathcal{P}} = 15{,}43\,\%$.

En principio, se podría haber esperado que todas las penalizaciones, en su medida, beneficiaran al reconocimiento ya que todas incorporan alguna información útil. Incluso podría pensarse que, por grandes que éstas fueran, deberían seguir beneficiando al reconocimiento ya que, independientemente de su magnitud, siguen respondiendo a informaciones útiles que provienen de evidencias acústicas (prosódicas). Sin considerar las SEA estimadas para acotar la discusión, aún cuando las SEA sean las correctas desde un punto de vista ortográfico-gramatical, ya se ha estudiado en el Capítulo 3 que en el discurso *continuo* no siempre se corresponden con los rasgos prosódicos medidos. De esta forma se vuelve a introducir un elemento de duda en la búsqueda de Viterbi y sería poco adecuado cortar totalmente un camino de hipótesis con una gran penalización. De hecho, esto es lo que se refleja experimentalmente en las curvas de la Figura 5.3.

Si ocurren errores de eliminación o inserción durante el reconocimiento con un MLRE, todas las transiciones que se encuentran luego quedarán desalineadas y la penalización para los caminos que se siguen a este tipo de errores podría ser excesiva. Esta desincronización entre las hipótesis en evaluación y la SEA podría hacer que se eliminen muchos buenos caminos ya que las penalizaciones resultantes por errores de eliminación o inserción (implícitas en este fenómeno) serían tan grandes como la acumulación de todas las penalizaciones que siguen hasta terminar la frase. Para evitar este eventual fenómeno de desincronización se realizaron experimentos con

un MLRE modificado donde se incorporaron arcos hacia atrás que no eran afectados por las constantes de penalización. Sin embargo, no se han encontrado mejoras significativas, probablemente debido a que las bajas tasas de error del sistema de referencia dejan poco margen para que ocurran estos fenómenos de desincronización por eliminaciones e inserciones.

Los resultados finales presentados en las Tablas 5.6 y 5.7 muestran los beneficios de la incorporación de información prosódica y acentual en el RAH. Más aún, si fuera posible contar con una SEA totalmente correcta, desde el punto de vista ortográfico-gramatical, los resultados de las Tablas 5.4 y 5.5 proveen una buena perspectiva de las reducciones de error que se podrían alcanzar. El análisis comparativo de los resultados finales (Tabla 5.8) muestra claramente significancia estadística de las mejoras obtenidas, incluso partiendo de una estimación pobre de las SEA (54.41 %).

Capítulo 6

Conclusiones

6.1. Conclusiones particulares

6.1.1. Prosodia y acentuación en el discurso continuo

1. En comparación con las palabras aisladas, en el discurso continuo se observa una disminución importante de las coincidencias entre el máximo de los tres rasgos prosódicos (energía, frecuencia fundamental y duración del núcleo vocálico) y la posición de la sílaba tónica. En muchos casos ninguno de los tres máximos coincide con la posición de la sílaba tónica. Una conclusión importante es que existe una muy baja tasa de coincidencias entre la máximo de frecuencia fundamental y la posición de la sílaba tónica según la acentuación.

2. Los mínimos de energía y los mínimos de duración del núcleo vocálico son muy poco representativos de la posición de la sílaba tónica en la palabra. Sin embargo, existe una mayor tasa de coincidencias entre los mínimos de frecuencia fundamental y la acentuación. Si bien esta tasa no alcanza (en promedio) a la de los máximos de energía y duración del núcleo vocálico, para las primeras sílabas el mínimo de frecuencia fundamental es más representativo de la acentuación que el máximo de la energía.

3. Cuando se estudió la influencia de las pausas en las coincidencias entre los rasgos prosódicos y la acentuación se pudo encontrar que sin bien las coincidencias entre los máximos de energía y frecuencia fundamental aumentaban levemente al no considerar las palabras afectadas por una pausa, las coincidencias entre la duración del núcleo vocálico y la acentuación se redujeron en aproximadamente un 10 %.

4. Mediante la diferencia de entonación por ajuste es posible reducir el efecto de la función distintiva en la curva de entonación pero, si bien se observó alguna mejora, no hubo un impacto significativo en las tasas de coincidencia con la acentuación, tanto para los mínimos como para los máximos de frecuencia fundamental.

5. Mediante el análisis de cadencias de frecuencia fundamental se alcanzó un 49.96 % de coincidencias entre las anticadencias y la posición del acento. Esta tasa es similar a la tasa de coincidencias para los máximos de energía por lo que se puede conlcuir que las anticadencias de la frecuencia fundamental son tan representativas de la acentuación como los máximos de energía.

6. Estos estudios fueron ampliados en un análisis más detallado donde se consideró la relación entre rasgos prosódicos y acentuación para cada una de las vocales y diptongos. Los resultados reafirman las conclusiones anteriores en cuanto a la importancia de los máximos de energía y duración del núcleo vocálico y la poca relevancia de los máximos de frecuencia fundamental. En relación al análisis de cadencias se puede concluir que la vocales acentuadas /á/, /ó/, /ú/ y sobre todo la /í/ son bien caracterizadas por una anticadencia de frecuencia fundamental.

6.1.2. Estimación de estructuras acentuales

1. El método de los árboles de redes neuronales permite clasificar estructuras acentuales con un porcentaje de aciertos del 89.98 %, superando ampliamente a los otros métodos evaluados. El algoritmo de entrenamiento optimiza de forma automática la estructura topológica y la implementación de una clasificador en base a un árbol de redes neronales sencillas y de bajo costo computacional.

2. Los árboles de redes neuronales son clasificadores estáticos que requieren de una segmentación silábica previa a la extracción de características de la señal de voz. En base a una segmentación silábica correcta, este método proporciona una muy buena solución al problema de clasificación de estructuras acentuales.

3. El problema de la segmentación silábica automática es un obstáculo importante en la estimación de las estructuras acentuales. Se propusieron dos métodos basados en una medida de distancia segmental a lo largo de la frase. El primer método realiza una optimización global mediante computación evolutiva y el segundo realiza una búsqueda local con un algoritmo de detección de máximos. Ambos métodos se desarrollaron con una perspectiva amplia que permite aplicarlos a la segmentación ciega de señales en general.

4. El método de segmentación evolutiva permite segmentar corpus de habla conociendo la cantidad de segmentos por frase. Las diferentes modalidades de segmentación probadas se ajustan en gran medida a la segmentación ideal pero sería necesario reducir el costo computacional de cálculo si se pretende utilizarlo en un sistema de tiempo real.

5. El método de segmentación evolutiva fue comparado con la segmentación realizada mediante modelos ocultos de Markov y se observaron

fallas por omisión e inserción de segmentos. Las posibles causas de estas fallas fueron discutidas y sopesadas en vista de que el algoritmo no utiliza ninguna información a priori acerca de la fonética y estructura gramatical de la frase a segmentar.

6. El método de segmentación por detector de máximos posee un menor costo computacional y no requiere la especificación a priori de la cantidad de segmentos en la frase. Mediante este método se segmentaron 600 frases y se alcanzó un error promedio del 32.36 %.

7. La dudosa fiabilidad de las segmentaciones silábicas hizo que la experimentación conjunta con los algoritmos de segmentación y clasificación estática no se presentase como buena alternativa. Por esta razón se adaptó la estructura de un modelo oculto de Markov para la estimación de secuencias de estructuras acentuales.

8. Se experimentaron muy diversas alternativas para el procesamiento de la señal, el modelado acústico y el modelado del lenguaje. En el mejor de los casos se obtuvo un 56.94 % de aciertos en la estimación de estructuras acentuales. Este resultado no es satisfactorio viendo a la etapa aisladamente, pero constituye un punto de partida para los experimentos en penalización prosódica.

6.1.3. Reconocimiento del habla con penalización prosódica

1. El método de la penalización prosódico acentual permite incorporar la información contenida en las secuencias de estructuras acentuales a un sistema de reconocimiento automático del habla basado en modelos ocultos de Markov.

2. Este método se basa en los modelos de lenguaje variantes en el tiempo y en la práctica puede implementarse a través de los modelos de lenguaje con red expandida. Con estos métodos es posible adaptar el modelo de lenguaje dentro de cada cada frase del corpus de habla a medida que se van conociendo evidencias prosódicas que cambien las probabilidades de ciertas hipótesis en la búsqueda del algoritmo de Viterbi.

3. Se estudió el comportamiento de las diferentes constantes de penalización prosódico acentual observándose que la penalización por fin de frase no beneficia al reconocimiento. Los errores para diferentes constantes de penalización en las transiciones entre palabras y en las

asociadas a un modelo de silencio tienen un mínimo local que permite seleccionarlas fácilmente.

4. La penalización para el suavizado de gramática siempre beneficia al reconocedor ya que ajusta el modelo a las condiciones de entrenamiento. Mediante diversos experimentos se cuantificaron los beneficios obtenidos por la penalización del suavizado de gramática que no tuviesen que ver con la información prosódica. Se aplicaron penalizaciones fijas e independientes, penalizaciones al azar y penalizaciones excesivamente grandes que anulaban determinadas partes del modelo de lenguaje.

5. Para los resultados finales se utilizó otro subconjunto de frases del corpus de habla. Se realizaron comparaciones con un sistema de referencia probado en las mismas condiciones y se obtuvieron diversas medidas de las mejoras obtenidas y sus significancias estadísticas. Para penalizar los modelos de lenguaje variantes en el tiempo se utilizaron las secuencias de estructuras acentuales estimadas mediante el método de los modelos ocultos de Markov y las extraídas directamente de las transcripciones del corpus de habla.

6. Todas las medidas de mejora favorecieron ampliamente al sistema con información acentual. El error promedio para el reconocimiento de palabras pasó de 7.54 % a 5.36 %, lo que representa una reducción relativa del 28.91 %, con una probabilidad del 99.9999999 % de que el sistema de reconocimieto propuesto sea mejor que el de referencia. Si se pudiese contar con una secuencia de estructuras acentuales en total concordancia con la acentuación los resultados del método de penalización prosódico acentual alcanzarían una mejora relativa del 36.87 % en el reconocimiento de palabras.

6.2. Conclusiones generales

1. Existen relaciones importantes entre los rasgos prosódicos y la acentuación. En el discurso continuo resulta más complejo asociar el comportamiento de un rasgo prosódico a la posición de la sílaba tónica y esto requiere de la utilización de técnicas de clasificación más sofisticadas.

2. Los métodos propuestos para la estimación de estructuras acentuales proveen una muy buena estimación cuando se cuenta con la segmentación silábica de la señal de voz. El método para realizar la segmentación y clasificación conjunta posee un rendimiento que, si bien es pobre visto en forma aislada, es de gran utilidad integrado a un sistema de reconocimiento automático del habla.

3. La penalización prosódico acentual es un método relativamente simple y flexible que permite incorporar la información acentual a un sistema de reconocimiento automático del habla basado en modelos ocultos de Markov.

4. Mediante la incorporación explícita de la información acentual es posible mejorar significativamente el rendimiento de un sistema de reconocimiento automático del habla continua en español.

6.3. Direcciones para continuar la investigación

1. Los estudios de la relación entre rasgos prosódicos y acentuación pueden ser ampliados para incluir todos los corpus de Albayzin y otros corpus con mayor variedad de expresiones del lenguaje hablado.

2. La extensión de los estudios realizados a otras variantes del castellano, las lenguas de España e Hispanoamérica y otros idiomas (particularmente para el inglés que está tan difundido en la actualidad). La tarea de extender los estudios a otros idiomas no es nada simple e implicaría, en cada caso, una nueva investigación prácticamente desde cero.

3. En cuanto al análisis de la prosodia y la acentuación sería interesante encontrar relaciones claras entre los rasgos prosódicos y la palabra que se pronuncia, quizás pasando a un segundo plano la acentuación definida por las reglas ortográficas y definiendo una nueva forma de clasificar las prominencias acentuales del idioma.

4. En lo relativo a la estimación de estructuras acentuales es necesario aumentar significativamente los rendimiento obtenidos ya que el éxito en la reducción de las tasas de error en el reconocimiento automático del habla está asegurado por las pruebas que se realizaron con las secuencias de estructuras acentuales extraídas de las transcripciones del corpus de habla.

5. Queda pendiente la integración del sistema de segmentación evolutiva con el clasificador basado en árboles de redes neuronales. Si bien no se esperaban altas tasas de rendimiento dado que una segmentación pobre puede arruinar por completo el proceso de clasificación estática, también es cierto que en definitiva el sistema basado en modelos ocultos de Markov no superó el 60 %.

6. En general, todos los experimentos realizados no fueron orientados a minimizar el costo computacional ni a optimizar los algoritmos en algún sentido para aplicaciones de tiempo real. En base a los resultados positivos que se alcanzaron se puede hacer una revisión de aspectos relacionados con la programación y la adaptación a sistemas operativos.

7. También con relación a la implementación práctica es interesante estudiar la integración del método de estimación de estructuras acentuales

en el mismo algoritmo de la búsqueda de Viterbi en lugar de usar los modelos de lenguaje.

8. En base a los experimentos de reconocimiento automático del habla realizados hasta ahora, sería necesario ampliar la variabilidad del material de habla utilizado. Se propone realizar pruebas utilizando corpus de habla que posean:

   - mayor vocabulario,
   - mayor cantidad de frases y locutores,
   - más perplejidad en las estructuras gramaticales,
   - más naturalidad en la pronunciación (habla espontánea),
   - contaminación con ruidos de diferente tipo y magnitud.

Apéndice A

Corpus de habla “Albayzin”

A.1. Generalidades

El corpus de habla Albayzin ha sido desarrollado con el objetivo de contribuir al desarrollo y la evaluación de sistemas de reconocimiento y procesamiento del habla. El diseño fue realizado a principios de la década del 90 [Casacuberta et al., 1991, Casacuberta et al., 1992] aunque la producción completa se finalizó en 1998. El proyecto “Albayzin” fue llevado adelante por 5 Universidades de España:

- Universidad de Granada (UGR) Dpto. ETC
- Universidad Politécnica de Valencia (UPV) Dpto. SIS
- Universidad Politécnica de Madrid (UPM) Dpto. IE y Dpto. SSR
- Universidad Autónoma de Barcelona (UAB) Dpto. FE
- Universidad Politécnica de Catalunya (UPC) Dpto. TSC

El corpus se compone de 15600 elocuciones pronunciadas por 152 hombres y 152 mujeres de entre 18 y 55 años de edad. Los hablantes pertenecen a la variedad central del castellano, en su mayor parte de las comunidades de Castilla-La Mancha, Castilla-León, Cantabria y Madrid. El material que contiene el corpus es leído aunque para el diseño se ha utilizado como punto de partida un estudio del habla espontánea. En promedio las frases poseen 4 s. de duración y fueron muestreadas a 16 KHz con una resolución de 16 bits. Se pudo medir una relación señal a ruido promedio de 48 dB.

Las frases de la base de datos se encuentran distribuidas en 3 corpus bien diferenciados:

1. Corpus fonético: es un conjunto genérico de 6800 elocuciones equilibradas fonéticamente, sin restricciones sintáctico-semánticas, que brinda un marco de referencia de la lengua castellana [Moreno et al., 1993]. Para el diseño de este corpus se han considerado tanto la proporción como la cobertura de las elocuciones de cada alófono en cada contexto. El corpus ha sido dividido en dos subconjuntos, uno de aprendizaje y otro de prueba. El subconjunto de aprendizaje consiste en la elocución de 200 frases diferentes por 4 locutores y 160 frases por otros 25 locutores (4800 elocuciones en total). El subconjunto de prueba consiste en 40 frases diferentes pronunciadas por 50 locutores.

2. Corpus geográfico: es un conjunto de 6800 elocuciones de frases dependientes de la aplicación, con restricciones semánticas y sintácticas relacionadas con la consulta de una base de datos de geografía española [Diaz et al., 1993]. Las construcciones sintácticas reflejan la forma natural del habla en el lengua castellana. Para extraerlas se analizaron 14918 frases obtenidas mediante entrevistas a 408 personas que intentaban obtener información sobre geografía española. Todas las frases se clasificaron según criterios lingüísticos, semánticos y de complejidad estructural. El subconjunto de entrenamiento consta de 50 frases diferentes pronunciadas por 88 locutores y el subconjunto de prueba consta de otras 50 frases diferentes pronunciadas por 48 locutores.

3. Corpus "Lombard": se compone de 2000 elocuciones de los corpus anteriores, producidas en condiciones adversas. El efecto Lombard consiste en un conjunto de modificaciones de la voz que se producen cuando el locutor se encuentra sometido a un nivel alto de ruido. Este corpus consta de las elocuciones de 40 locutores que pronuncian 50 frases diferentes cada uno.

A.2. Subconjunto 1 (SC1)

A.2.1. Características generales

El subconjunto SC1 contiene 600 elocuciones y está diseñado con las pautas generales del corpus geográfico [Diaz et al., 1998]. En la siguiente tabla se resumen las características más importantes de este subconjunto.

Total de elocuciones	600
Total de frases con texto diferente	200
Frases interrogativas	258
Duración promedio de las frases	3.55 s.
Duración total	2442 s.
Total de palabras	5678
Total de palabras diferentes	202
Perplejidad de la gramática	5.9
Hablantes femeninos	6
Hablantes masculinos	6

A.2.2. Frases

Cada locutor se identifica por las dos primeras letras del nombre de archivo: **aa**, **ac**, **al**, **an**, **aq**, **ar**, **ma**, **mg**, **mj**, **mk**, **mm** y **mo**. Las que comienzan con **a** corresponden a elocuciones de mujeres y las que comienzan por **m** a los hablantes masculinos. Los últimos tres números del archivo identifican la frase pronunciada y se detallan a continuación.

001	¿A qué mar va a parar el río español de mayor longitud?
002	¿Cómo se llama el mar que baña Valencia?
003	¿Cuál es el caudal de todos los ríos de la Comunidad Valenciana?
004	¿Cuál es el caudal del Ebro?
005	¿Cuál es el caudal del río más largo que pasa por Andalucía?
006	¿Cuál es el caudal máximo de los ríos españoles?
007	¿Cuál es el caudal y longitud del Tajo?
008	¿Cuál es el mar en el que desembocan mayor número de ríos con una longitud mayor de 200 kilómetros?
009	¿Cuál es el mar que rodea las Canarias?
010	¿Cuál es el nombre del río más largo de la Península?
011	¿Cuál es el río de mayor longitud que desemboca en el mar Cantábrico?
012	¿Cuál es el río más caudaloso que pasa por Extremadura?
013	¿Cuál es el río más largo que atraviesa por lo menos 2 comunidades?
014	¿Cuál es la comunidad autónoma de mayor extensión por la que pasa el río Ebro?
015	¿Cuál es la extensión de la comunidad autónoma en la que nace el río Ebro?
016	¿Cuál es la longitud de todos los ríos?
017	¿Cuáles son las comunidades autónomas con una extensión superior a 20.000 kilómetros cuadrados?
018	¿Cuáles son las comunidades autónomas por las que pasan más ríos?
019	¿Cuáles son las comunidades que atraviesa el Tajo?
020	¿Cuáles son las comunidades que lindan con el mar?
021	¿Cuáles son los ríos catalanes más largos que 100 kilómetros?
022	¿Cuáles son los ríos cuya longitud es superior a 100 kilómetros?
023	¿Cuáles son los ríos que desembocan en el Cantábrico?
024	¿Cuáles son los ríos que pasan por Extremadura y otras 2 comunidades autónomas?
025	¿Cuáles son los ríos que pasan por la comunidad de Valencia?
026	¿Cuántas comunidades están bañadas por 2 mares?
027	¿Cuánto mide el Tajo?
028	¿Cuántos metros cúbicos por segundo lleva el Turia?
029	¿Cuántos mares reciben agua de un río?
030	¿Cuántos ríos con caudal mayor de 800 metros cúbicos por segundo pasan por la Comunidad Valenciana?
031	¿Cuántos ríos de Castilla y León tienen más de 100 kilómetros?
032	¿Cuántos ríos pasan por Aragón y Cataluña?
033	¿Cuántos ríos son más largos de 200 kilómetros?
034	¿Dónde desemboca el Guadiana?
035	¿Dónde nace el río Duero?
036	¿Dónde nace el río Ebro?
037	¿En qué comunidad autónoma está el río más caudaloso?
038	¿En qué comunidad autónoma hay más ríos?
039	¿En qué comunidad autónoma pasan nacen y desembocan más ríos?
040	¿En qué comunidad desemboca el río Ebro?
041	¿En qué comunidad nace y pasa el Pisuerga?
042	¿En qué comunidad nacen más ríos?
043	¿En qué mar desemboca el río más caudaloso de la comunidad andaluza?
044	¿En qué mar desembocan mayor número de ríos?
045	¿Es el Ebro más caudaloso que el Tajo?
046	¿Hay algún río cuyo caudal sea mayor que 100 metros cúbicos por segundo?
047	¿Me podría decir cuál es la comunidad donde está el nacimiento del Guadiana?
048	¿Pasa algún río por más de 4 comunidades?
049	¿Pasa el río Duero por la Comunidad de Madrid?
050	¿Por cuántas comunidades pasa el Ebro?

051	¿Por dónde pasa el río Duero?
052	¿Por dónde pasa el río con más caudal?
053	¿Por qué comunidad pasan más ríos?
054	¿Por qué mar está bañada Asturias?
055	¿Qué caudal tiene el Ebro?
056	¿Qué caudal tiene el Miño?
057	¿Qué comunidad autónoma es menos extensa?
058	¿Qué comunidad bañada por el Mediterráneo es la más extensa?
059	¿Qué comunidades no son bañadas por algún mar?
060	¿Qué comunidades son bañadas por el Tajo?
061	¿Qué comunidades tienen una extensión mayor de 1.000 kilómetros cuadrados?
062	¿Qué extensión tiene el País Vasco?
063	¿Qué longitud tiene el río más largo?
064	¿Qué mar baña Asturias?
065	¿Qué mar baña las costas de la Comunidad de Madrid?
066	¿Qué mar baña las costas del País Vasco?
067	¿Qué mar está junto a la Comunidad Valenciana?
068	¿Qué río cruza menos comunidades?
069	¿Qué río desemboca en el mar Mediterráneo y pasa por Murcia?
070	¿Qué río es más largo el Tajo o el Ebro?
071	¿Qué río tiene más caudal el Tajo o el Ebro?
072	¿Qué ríos desembocan en el mar Menor?
073	¿Qué ríos extremeños tienen una longitud superior a los 200 kilómetros?
074	¿Qué ríos hay en Asturias?
075	¿Qué ríos nacen en Cantabria?
076	¿Qué ríos pasan por Asturias y no nacen allí?
077	¿Qué ríos poseen un caudal superior a 800 metros cúbicos por segundo?
078	¿Qué ríos tienen más caudal que el río Duero?
079	¿Qué ríos tienen una longitud comprendida entre 500 y 1.000 kilómetros?
080	¿Seguro que el Segura pasa por la Comunidad de Valencia?
081	¿Tiene alguna comunidad más extensión que la comunidad andaluza?
082	¿Tienen la misma longitud y el mismo caudal el río Guadiana y el río Guadalquivir?
083	Caudal de los ríos con más de 100 kilómetros de longitud.
084	Caudal de los ríos que pasan por Castilla y León.
085	Caudal del río que pasa por la comunidad de Valencia.
086	Comunidad autónoma más grande.
087	Comunidades autónomas más grandes que Cataluña.
088	Comunidades con más de 5 ríos.
089	Comunidades por las que pasa el río Ebro.
090	Comunidades que baña el mar Mediterráneo.
091	Dígame el nombre del río más largo.
092	De los ríos del estado ¿cuántos desembocan en el Mediterráneo?
093	Deseo saber el caudal del río Miño.
094	Di el caudal del río menos caudaloso.
095	Di el río más caudaloso que desemboca en el Cantábrico.
096	Dime comunidades cuya superficie sea mayor a 1.000 kilómetros cuadrados.
097	Dime cuál es la comunidad autónoma de menor extensión.
098	Dime cuáles son las comunidades autónomas.
099	Dime cuántos ríos de la Comunidad Valenciana tienen más de 200 kilómetros de longitud.
100	Dime dónde desemboca el río Júcar.

101	Dime dónde muere el río Ebro.
102	Dime dónde nace el río Júcar.
103	Dime el caudal de los ríos de Cataluña.
104	Dime el caudal de todos los ríos que desembocan en el mar Mediterráneo.
105	Dime el caudal del río Cuervo.
106	Dime el caudal del río más pequeño que pasa por La Rioja.
107	Dime el caudal máximo de los ríos.
108	Dime el mar donde desemboca el río Turia.
109	Dime el mar en que desemboca el Miño.
110	Dime el número de ríos que desembocan en el Mediterráneo y que sean entre 1.000 y 200 kilómetros de largo.
111	Dime el nombre de las 3 comunidades autónomas más grandes.
112	Dime el nombre de las comunidades que linden con 2 mares.
113	Dime el nombre de los mares que bañan la comunidad de Andalucía.
114	Dime el nombre de los ríos que desembocan en el océano Atlántico.
115	Dime el nombre de los ríos que pasan por la Comunidad de Madrid.
116	Dime el nombre de los ríos que tienen menos de 100 kilómetros.
117	Dime el nombre de todas las comunidades que tienen mar.
118	Dime el río de mayor caudal que pase por la comunidad de Valencia.
119	Dime el río de menor longitud de Cataluña.
120	Dime en qué comunidad autónoma nace el Tajo.
121	Dime en qué comunidad nace el río Turia.
122	Dime la comunidad en la que desemboca el río Turia.
123	Dime la extensión de la comunidad asturiana.
124	Dime la extensión de las comunidades por donde pasa el Ebro.
125	Dime la longitud de los ríos que pasan por la Comunidad de Madrid.
126	Dime la longitud del río Guadalquivir.
127	Dime la longitud del río más largo.
128	Dime las comunidades autónomas con extensión superior a 1.000 kilómetros cuadrados.
129	Dime las comunidades autónomas.
130	Dime las comunidades que lindan con más de un mar.
131	Dime lo grande que es el Ebro.
132	Dime los mares que bañan Andalucía.
133	Dime los mares.
134	Dime los ríos con una longitud superior a 500 kilómetros.
135	Dime los ríos de la comunidad autónoma gallega.
136	Dime los ríos que desembocan en Andalucía.
137	Dime los ríos que desembocan en el Atlántico.
138	Dime los ríos que nacen en la Comunidad Foral de Navarra.
139	Dime los ríos que nacen y desembocan en la misma comunidad.
140	Dime los ríos que pasan por la Comunidad de Madrid.
141	Dime los ríos que tengan una longitud mayor que 500 kilómetros.
142	Dime qué longitud tiene el río Júcar.
143	Dime qué río tiene el caudal más grande.
144	Dime si por la comunidad de Valencia pasa o no más de un río.
145	Dime todos los mares que bañan Andalucía.
146	Dime todos los ríos que desembocan en el mar Cantábrico.
147	El río Ebro ¿pasa por la comunidad autónoma de Navarra?
148	El río Miño ¿por cuántas comunidades autónomas pasa?
149	Entre el río Ebro y el Júcar ¿cuál de ellos es más corto?
150	Enumera las comunidades autónomas por donde pasa el Ebro.

151	Enumera los ríos que tienen una longitud mayor de 100 kilómetros.
152	Enumerar los ríos que atraviesan la comunidad autónoma de Asturias.
153	Extensión de la comunidad autónoma por la cual pasa el río cuyo nombre es Guadalquivir.
154	Extensión del País Vasco.
155	La extensión de las comunidades autónomas que dan al mar Atlántico.
156	Lista de las comunidades por las que pase algún río de longitud mayor de 1.000 kilómetros.
157	Listado de todos los ríos con una longitud menor que la del Júcar.
158	Longitud de los ríos que desembocan en el mar Cantábrico.
159	Longitud del río Ebro.
160	Longitud del río que pasa por la Comunidad Valenciana.
161	Lugar donde desemboca el Júcar.
162	Mar en el que desembocan más ríos.
163	Mares en los que desembocan 5 o más ríos de longitud superior a 100 kilómetros.
164	Mares que bañan la comunidad gallega.
165	Nómbrame los ríos que pasan exactamente por 3 comunidades autónomas.
166	Número de mares del Estado Español.
167	Número de ríos que nacen y desembocan en la Comunidad Valenciana.
168	Nombra los ríos que pasan por las comunidades autónomas que no dan al mar.
169	Nombre de la comunidad autónoma en la que desemboquen mayor número de ríos.
170	Nombre de las 3 comunidades de menor extensión.
171	Nombre de las comunidades con extensión mayor que la Comunidad Valenciana.
172	Nombre de los mares que están en la Comunidad Valenciana.
173	Nombre de los ríos cuya longitud no supere los 1.000 kilómetros y no sea menor de 100 kilómetros.
174	Nombre de los ríos cuyo caudal es superior a 800 metros cúbicos por segundo.
175	Nombre de los ríos que desembocan en cada mar.
176	Nombre de los ríos que nacen en La Rioja y pasan por aquellas comunidades por las que sólo pasa ese río.
177	Nombre de los ríos que pasen por Castilla y León desembocan en el Atlántico y su caudal sea menor que el del río Tajo.
178	Nombre de todos los mares que bañan Andalucía.
179	Nombre del mar en el que desemboca un río que nace en Aragón.
180	Nombres de comunidades autónomas cuya extensión se encuentra entre 1.000 y 2.000 kilómetros cuadrados.
181	Obtener las comunidades autónomas por donde pasa el Ebro.
182	Quiero saber los nombres de los ríos más largos de 200 kilómetros.
183	Quisiera conocer cuántos ríos tienen un caudal de más de 200 metros cúbicos por segundo y son de menos de 1.000 kilómetros de largo.
184	Quisiera saber en qué mar desemboca el Segura.
185	Quisiera saber qué comunidades autónomas no tienen salida al mar.
186	Río más corto que desemboca en el Cantábrico.
187	Río más largo que nazca en Extremadura.
188	Ríos con caudal superior al del río Guadalquivir.
189	Ríos cuya longitud sea mayor de 1.000 kilómetros.
190	Ríos de Cantabria de más de 100 kilómetros de longitud.
191	Ríos de la comunidad autónoma gallega.
192	Ríos que atraviesen más de 3 autonomías.
193	Ríos que desembocan en el Cantábrico con una longitud mayor a 100 kilómetros.
194	Ríos que desemboquen en el Cantábrico.
195	Ríos que mueren en el Cantábrico.
196	Ríos que nacen en la Comunidad de Madrid.
197	Ríos que nacen en una comunidad bañada por el mar y desembocan en otra comunidad.
198	Ríos que pasan por la comunidad autónoma de Valencia.
199	Ríos que tengan un caudal superior a 800 metros cúbicos por segundo.
200	Todos los ríos.

A.2.3. Acentuación

En esta sección se detalla la clasificación de las palabras del diccionario según su función gramatical y acentuación. Siguiendo a [Quilis, 1993], las palabras pueden clasificarse en acentuadas e inacentuadas. A continuación se indican las palabras acentuadas con una **A** y las inacentuadas con una **I**. Hay situaciones particulares en donde la acentuación es dependiente del contexto en el que se encuentra la palabra y estos casos se indicarán con una **D**. Para distinguir las diferentes funciones que cumple una palabra en la frase se ha utilizado la siguiente notación.

Abreviatura	Función
S	sustantivo
V	verbo
A	adjetivo
Ai	adjetivo indefinido
Ap	adjetivo posesivo
B	artículo
D	adverbio
Dm	adverbio terminado en *mente* (doble acentuación)
P	pronombre
Pp	pronombre posesivo
Q	preposición
C	conjunción
I	formas interrogativas *qué*, *cuál*, etc.
nI	formas no interrogativas *que*, *cual*, etc.
N	numerales

A continuación se detallan las funciones que cumple cada palabra del diccionario *en las frases del subconjunto SC1* y la tipología acentual correspondiente.

Palabra	Ac.	Fn.	Palabra	Ac.	Fn.	Palabra	Ac.	Fn.
a	I	Q	cúbicos	A	A	largo	A	AS
agua	A	S	Cuervo	A	S	largos	A	AS
al	I	QB	cuya	A	Pp	las	I	B
algún	A	Ai	cuyo	A	Pp	León	A	S
alguna	A	Ai	dan	A	V	lindan	A	V
allí	A	A	de	I	Q	linden	A	V
Andalucía	A	S	decir	A	V	lista	A	S
andaluza	A	A	del	I	QB	listado	A	S
aquellas	A	A	desemboca	A	V	llama	A	V
Aragón	A	S	desembocan	A	V	lleva	A	V
asturiana	A	S	desemboquen	A	V	lo	I	BP
Asturias	A	S	deseo	A	V	longitud	A	S
Atlántico	A	AS	di	A	V	los	I	BP
atraviesa	A	V	dígame	A	V	lugar	A	S
atraviesan	A	V	dime	A	V	Madrid	A	S
atraviesen	A	V	donde	I	nI	mar	A	S
autónoma	A	A	dónde	A	I	mares	A	S
autónomas	A	A	dos	D	N	más	A	AD
autonomías	A	S	doscientos	D	N	máximo	A	S
baña	A	V	Duero	A	S	mayor	A	A
bañada	A	V	Ebro	A	S	me	I	P
bañadas	A	V	el	I	B	Mediterráneo	A	A
bañan	A	V	ellos	I	P	menor	A	A
cada	A	A	en	I	Q	menos	A	A
Canarias	A	S	encuentra	A	V	metros	A	S
Cantabria	A	S	entre	I	Q	mide	A	V
Cantábrico	A	A	enumera	A	V	mil	D	N
Castilla	A	S	enumerar	A	V	Miño	A	S
catalanes	A	A	es	A	V	misma	A	A
Cataluña	A	S	ese	A	S	mismo	A	A
caudal	A	S	español	A	A	muere	A	V
caudaloso	A	A	españoles	A	A	mueren	A	V
cien	D	N	está	A	V	Murcia	A	S
cinco	D	N	estado	A	S	nace	A	V
cómo	A	I	están	A	V	nacen	A	V
comprendida	A	V	exactamente	A	Dm	nacimiento	A	S
comunidad	A	S	extensa	A	A	Navarra	A	S
comunidades	A	S	extensión	A	S	nazca	A	V
con	I	Q	Extremadura	A	S	no	A	D
conocer	A	V	extremeños	A	A	nombra	A	V
corto	A	A	Foral	A	A	nómbrame	A	V
costas	A	S	gallega	A	A	nombre	A	S
cruza	A	V	grande	A	A	nombres	A	S
cuadrados	A	S	grandes	A	A	número	A	S
cual	I	nI	Guadalquivir	A	S	o	I	C
cuál	A	I	Guadiana	A	S	obtener	A	V
cuáles	A	I	hay	A	V	océano	A	S
cuántas	A	I	Júcar	A	S	ochocientos	D	N
cuánto	A	I	junto	A	D	otra	A	A
cuántos	A	I	kilómetros	A	S	otras	A	A
cuatro	D	N	la	I	B	País	A	S

Palabra	Ac.	Fn.	Palabra	Ac.	Fn.	Palabra	Ac.	Fn.
parar	A	V	río	A	S	superficie	A	S
pasa	A	V	Rioja	A	S	superior	A	A
pasan	A	V	ríos	A	S	Tajo	A	S
pase	A	VS	rodea	A	V	tengan	A	V
pasen	A	V	saber	A	V	tiene	A	V
península	A	S	salida	A	S	tienen	A	V
pequeño	A	A	se	I	P	todas	A	A
Pisuerga	A	S	sea	A	V	todos	A	A
podría	A	V	sean	A	V	tres	D	N
por	I	Q	segundo	A	S	Turia	A	S
poseen	A	V	Segura	A	S	un	A	A
que	I	nI	seguro	A	A	una	A	A
qué	A	I	si	A	D	va	A	V
quiero	A	V	sólo	A	D	Valencia	A	S
quinientos	D	N	son	A	V	valenciana	A	A
quisiera	A	V	su	I	Ap	Vasco	A	A
reciben	A	V	supere	A	V	veinte	D	N
						y	I	C

A.3. Subconjunto 2 (SC2)

El subconjunto SC2 está formado por 1000 elocuciones y posee las mismas características generales que el SC1. Las frases son diferentes a las del SC1 y también los locutores participantes. Este subconjunto se encuentra dividido en otros dos subconjuntos de aprendizaje y prueba.

El subconjunto de aprendizaje consta de 600 elocuciones y 5678 palabras en total. De estas 600 frases hay 128 que tienen caracter interrogativo. Los 12 locutores pronunciaron 50 frases cada uno a partir de 300 textos diferentes. El subconjunto de prueba consta de 400 elocuciones (86 interrogativas) y 3770 palabras en total. En este subconjunto participaron 8 locutores pronunciado 50 frases cada uno a partir de un conjunto de 200 frases diferentes.

Los archivos involucrados quedan identificados por las letras iniciales: **eu**, **ev**, **ew**, **ik**, **il**, **lj**, **ru**, **rv**, **rw**, **vk**, **vl**, **zj** en el subconjunto de aprendizaje y **bx**, **by**, **gu**, **ju**, **nx**, **ny**, **tu**, **xu** en el subconjunto de prueba.

Apéndice B

Glosario

B.1. Notación

a, i	itálica minúscula: variables escalares.
$\mathbf{x}, \mathbf{v}$	negrita minúscula: vectores columna.
$\mathbf{A}, \mathbf{B}$	negrita mayúscula: matrices, secuencias de secuencias.
$\mathbf{h}^2, \mathbf{q}^T$	negrita minúscula: secuencias, el superíndice indica la cantidad de elementos.
$\mathbf{X}^T$	negrita mayúscula: secuencia de secuencias, el superíndice indica la cantidad de secuencias.
$\mathcal{T}_F$	caligráfica mayúscula: operadores, funcionales.
$\mathbb{R}, \mathbb{V}$	doble borde mayúscula: espacios vectoriales.
$\mathcal{A}, \mathcal{Q}$	caligráfica mayúscula: conjuntos.
N_v, N_ω	itálica mayúscula: cantidad de elementos, dimensiones. El subíndice indica la variable de que se trata.
$\approx$	aproximadamente igual a.
$\propto$	proporcional a.
$\triangleq$	igual por definición.
$\doteq$	debe ser igual a.
$\arg\max_x f(x)$	valor de x que maximiza $f(x)$.
$\lvert\mathcal{Q}\rvert$	cardinalidad, cantidad de elementos del conjunto $\mathcal{Q}$.
$\mathbf{x}^{\mathrm{T}}, \mathbf{A}^{\mathrm{T}}$	transpuesta de un vector o matriz.
∇_μ	operador gradiente en las coordenadas μ.
$\Pr(\cdot)$	probabilidad.
$p(\cdot)$	función de densidad de probabilidad.
$\Pr(x, y)$	probabilidad conjunta.
$\Pr(x\lvert y)$	probabilidad condicional.
$\mathcal{N}(x)$	distribución gaussiana unidimensional, con media 0 y desviación estándar 1.
$\mathcal{N}(\cdot, \boldsymbol{\mu}, \mathbf{U})$	distribución gaussiana multidimensional, con media $\boldsymbol{\mu}$ y matriz de covarianza $\mathbf{U}$.
$\sqcup(a, b)$	generador de números al azar con distribución uniforme en el rango a y b.
$\mathbf{x}_t$	evidencia acústica en el tramo de tiempo t
Θ, $\tilde{\Theta}$	modelo oculto de Markov, estimación inicial y siguiente de los parámetros en el proceso de optimización
${}^W\Theta_{w_m}$	modelo oculto de Markov de la palabra w_m
$i_{(m)}$	i-ésimo estado del modelo de palabra w_m
$1_{(m)}$, $\lvert\mathcal{Q}\rvert_{(m)}$	primer y último estado del modelo de palabra w_m
$\mathbf{q}^T_{(m)}$	secuencia de T estados en ${}^W\Theta_{w_m}$

[r̄], [∂]	representación fonética: fonos, alófonos.
/a/, /A/	representación fonológica: fonemas, tonicidades silábicas, estructuras acentuales, secuencias de estructuras acentuales.
{hora}	representación morfológica: morfemas.
palabra	representación ortográfica: en general palabras o frases.

B.2. Acrónimos

RAH	reconocimiento automático del habla
MM	modelos de Markov
MOM	modelos ocultos de Markov, en inglés HMM de *hidden Markov models*
MOMC	modelos ocultos de Markov continuos
MOMSC	modelos ocultos de Markov semicontinuos o de parámetros enlazados
MA	modelo acústico
ML	modelo de lenguaje
MC	modelo compuesto (MA+ML)
MLVT	modelo de lenguaje variante en el tiempo
MLRE	modelo de lenguaje con red expandida
MLRR	modelo de lenguaje con red recursiva
CE	coeficientes espectrales
CPL	coeficientes de predicción lineal
CC	coeficientes cesptrales
CCEM	coeficientes cesptrales en escala de mel
TC	transformada coseno
TDF	transformada discreta de Fourier
TDFI	transformada discreta de Fourier inversa
EA	estructura acentual
SEA	secuencia de estructuras acentuales
TAM	modelos con sílabas tónicas, átonas y monosílabos sin especificar su tonicidad.
TA-v	modelos de tonicidad silábica para cada vocal y diptongo
TA*-v	modelos de tonicidad silábica para cada vocal y diptongo sin formar estructuras acentuales, organizados en frases como secuencias de sílabas
TA-Q	modelos con sílabas tónicas y átonas donde existen palabras clasificadas como inacentuadas siguiendo a [Quilis, 1993]
AD	árbol de decisión
ARN	árbol de redes neuronales
CVA	cuantización vectorial con aprendizaje, en inglés LVQ de *learn vector quantization*
CVA1-O	algoritmo 1 optimizado para cuantización vectorial con aprendizaje
MAO	mapa autoorganizativo
RNA	red neuronal artificial

B.3. Terminología

voz: realización física o emisión sonora del habla. Por ejemplo, se dice que un cantante o locutor tiene "buena voz" sin importar cual es el mensaje que transmite. En el contexto del procesamiento de señales suele utilizarse también **señal de voz**.

elocución: acto que realiza el locutor. Tiene un alcance similar al término **voz** y también puede asemejarse a **emisión de voz**.

habla: en un sentido más amplio, incluye a la voz y todos los niveles de organización estructural, desde el físico hasta el semántico y pragmático.

---

fonema: modelo de un sonido elemental del habla. Hace referencia al modelo para el estudio fonológico y no a sus posibles pronunciaciones en diferentes contextos (Sección 1.3.2, página 27).

alófonos: diferentes realizaciones de un mismo fonema. También se utiliza el término **fono** como sinónimo de alófono (página 27).

vocoide: alófono del fonema de una vocal. Se utiliza como sinónimo de **sonido vocálico** (página 29).

gramática: cuando está relacionado con la lingüística es la agrupación de palabras en las clases fundamentales sustantivo, adjetivo, verbo y adverbio y en un sentido más amplio incluyendo tanto al conocimiento lexicográfico como al sintáctico (Sección 1.3.4, página 39). Cuando el término está en el contexto de los modelos para reconocimiento automático del habla es la estructura matemática que se utiliza en teoría de lenguajes formales (Sección 1.4.5, página 54). Esta última es la acepción más utilizada en los capítulos 2 al 6.

---

prosodia: desde un punto de vista físico es el efecto resultante de las diferentes combinaciones de energía, frecuencia fundamental y duración de suprasegmentos, aplicadas al lenguaje hablado. Desde una perspectiva lingüística se define mejor como un conjunto de reglas generales que rigen la superposición de rasgos como la cantidad, la duración y la entonación en el lenguaje hablado (Sección 1.3.3, página 31).

---

prosodema: conjunto de elementos relacionados con la expresión y representados principalmente por el acento, la cantidad, la duración y la entonación. El término **suprasegmento** se utiliza como sinónimo de prosodema y alude claramente al hecho de que estos rasgos se superponen a los propios de cada segmento (página 31).

rasgos prosódicos: manifestaciones físicas de la prosodia, esto es, energía, frecuencia fundamental y duración de un tramo de la señal de voz (página 32).

---

frecuencia fundamental: (o simplemente F_0) rasgo prosódico en un sentido físico, esto es, el valor de frecuencia a la que vibran las cuerdas vocales cuando se pronuncia algún fonema sonoro (Sección 1.2.4 página 19).

entonación: en un sentido amplio es un conjunto de fenómenos lingüísticos relacionados directamente con la frecuencia fundamental de las emisiones de voz (Sección 1.3.3, página 36). En un sentido más restringido es la curva de frecuencia fundamental en función del tiempo a lo largo de una frase completa o **curva melódica** (página 38).

tonema: entonación analizada al nivel de una sílaba o algún suprasegmento entre los fonemas y las sílabas (página 36). De forma similar se utiliza el término **tono** cuando la entonación es analizada a nivel de una palabra.

cadencias de entonación: en una clasificación lingüística se utiliza para describir la entonación a partir de un diccionario de estructuras tonemáticas que se clasifican como cadencias, mesetas y anticadencias, según la curva melódica posea un descenso, se mantenga estable o ascienda, respectivamente (página 37). Cuando se realiza un análisis matemático de la entonación a nivel de tonemas el término **cadencia** alude a una pendiente negativa y **anticadencia** a una pendiente positiva en la curva de frecuencia fundamental. Cuando la curva posee una pendiente próxima a cero se habla de **meseta** de frecuencia fundamental (Sección 3.3.5, página 123).

---

acentuación: es la representación del acento en el lenguaje escrito y queda establecida por las reglas ortográficas (Sección 1.3.3, página 35). No se hace referencia solamente a la tilde sino también a las situación en

---

que, a pesar de no utilizarse esta grafía, las reglas ortográficas definen inequívocamente la sílaba acentuada. Se denomina **acentuación prosódica** a la manifestación del acento en los rasgos prosódicos de una emisión de voz (Sección 3.1, página 108).

acento: es uno de los prosodemas más importantes del habla y en general el término se utiliza en un sentido amplio, incluyendo a la acentuación (páginas 32 y 34).

tónica: y su antónimo **átona** se utilizan para distinguir la acentuación a nivel de sílabas, es decir, la **tonicidad silábica**. Se representan con una /T/ las sílabas tónicas y con una /A/ las sílabas átonas (página 34).

acentuada: y su antónimo **inacentuada** se utilizan principalmente para distinguir la acentuación a nivel de palabras (página 32) y vocales. Las palabras acentuadas pueden ser **oxítonas**, **paroxítona**, **proparoxítonas** o **superproparoxítonas** según la posición de la tónica en relación a la última sílaba de la palabra, /-T/, /-TA/, /-TAA/ o /-TAAA/, respectivamente (página 35). El término también se aplica a las vocales.

estructura acentual: concatenación de las tonicidades silábicas de una palabra (página 34).

secuencia de estructuras acentuales: secuencia que forma la transcripción de una frase en estructuras acentuales (página 34).

Bibliografía

[Aguilar et al., 1997] Aguilar, L., Giménez, J. A., Machuca, M., Marín, R., y Riera, M. "Catalan vowel duration". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 2, páginas 771–774.

[Akaike, 1974] Akaike, H. "A new look at the statistical model identification". *IEEE Trans. on Automatic Control*, volumen 19, número 6, páginas 716–723.

[Almiñana, 1991] Almiñana, J. M. G. *Modelización de Patrones Melódicos del Español para la Síntesis y el Reconocimiento del Habla*. Servei de Publicacions de la Universitat Autònoma de Barcelona, Facultad de Filosofia i Lletres, Departament de Filologia Espanyola, Barcelona.

[Arslan y Hansen, 1996] Arslan, L. M. y Hansen, J. H. L. "Language accent classification in american english". *Speech Communication*, volumen 18, páginas 353–367.

[Bäck et al., 1997] Bäck, T., Hammel, U., y Schewfel, H.-F. "Evolutionary computation: Comments on history and current state". *IEEE Trans. on Evolutionary Computation*, volumen 1, número 1, páginas 3–17.

[Bartkova y Jouvet, 1999] Bartkova, K. y Jouvet, D. "Selective prosodic post-processing for improving recognition of french telephone numbers". En *Proceedings of the 7th European Conference on Speech Communication and Technology*, volumen 1, páginas 267–270.

[Batliner et al., 1997] Batliner, A., Kießling, A., Kompe, R., Niemann, H., y Nöth, E. "Tempo and its change in spontaneous speech". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 2, páginas 763–766.

[Bhandarkar y Zhang, 1999] Bhandarkar, S. M. y Zhang, H. "Image segmentation using evolutionary computation". *IEEE Trans. on Evolutionary Computation*, volumen 3, número 1.

[Bishop, 1995] Bishop, C. M. *Neural Networks for Pattern Recognition*. Oxford University Press.

[Bonafonte et al., 1997] Bonafonte, A., Esquerra, I., Febrer, A., y Vallverdu, F. "A bilingual text-to-speech system in spanish and catalan". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 5, páginas 2455–2458.

[Bosch y Gallés, 1997] Bosch, L. y Gallés, N. "The role of prosody in infants' native-language discrimination abilities: the case of two phonologically close languages". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 1, páginas 231–234.

[Bourlard et al., 1996] Bourlard, H., Hermansky, H., y Morgan, N. "Towards increasing speech recognition error rates". *Speech Communication*, volumen 18, número 3, páginas 205–231.

[Breiman et al., 1984] Breiman, L., Friedman, J. H., Olshen, R. A., y Stone, C. J. *Classification and Regression Trees.* Wadsworth Int.

[Brindöpke et al., 1999] Brindöpke, C., Fink, G. A., y Kummert, F. "A comparative study of HMM-based approaches for the automatic recognition of perceptually relevant aspects of spontaneous german speech melody". En *Proceedings of 7th European Conference on Speech Communication and Technology*, volumen 2, páginas 699–702.

[Brindöpke et al., 1998] Brindöpke, C., Fink, G. A., Kummert, F., y Sagerer, G. "A HMM-based recognition system for perceptive relevant pitch movements of spontaneous german speech". En *Proceedings of the 5th International Conference on Spoken Language Processing.* Prosody and Emotion 6.

[Brugnara et al., 1993] Brugnara, F., Falavigna, D., y Omologo, M. "Automatic segmentation and labeling of speech based on hidden Markov models". *Speech Communication*, volumen 12, número 4, páginas 357–370.

[Buckow et al., 1998] Buckow, J., Batliner, A., Huber, R., Nöth, E., Warnke, V., y Niemann, H. "Dovetailing of acoustic and prosody in spontaneous speech recognition". En *Proceedings of 5th International Conference on Spoken Language Processing.* Prosody and Emotion 2.

[Busdhtein, 1996] Busdhtein, D. "Robust parametric modeling of durations in hidden Markov models". *IEEE Trans. on Speech and Audio Processing*, volumen 4, número 3.

[Cahn, 1998] Cahn, J. E. "A computational memory and processing model for prosody". En *Proceedings of the 5th International Conference on Spoken Language Processing.* Prosody and Emotion 2.

[Campione y Véronis, 1998] Campione, E. y Véronis, J. "A statistical study of pitch target points in five languages". En *Proceedings of the 5th International Conference on Spoken Language Processing.* Prosody and Emotion 5.

[Casacuberta et al., 1992] Casacuberta, F., García, R., Llisterri, J., Nadeu, C., Pardo, J. M., y Rubio, A. "Desarrollo de corpus para investigacion en tecnologias del habla". *Boletín de la Sociedad Española de Procesamiento del Lenguaje Natural*, volumen 1, número 12, páginas 35–42.

[Casacuberta et al., 1991] Casacuberta, F., García, R., Llisterri, J., Nadeu, C., Prado, J. M., y Rubio, A. "Development of a spanish corpora for the speech research". En *Proceedings of the Workshop on International Cooperation and Standardisation of Speech Databases and Speech I/O Assessment Methods*, Chiavari, Italy. CEC DGXIII, ESCA and ESPRIT PROJECT 2589.

[Caspers, 1997] Caspers, J. "Testing the meaning of four dutch pitch accent types". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 2, páginas 863–866.

[Chellapilla, 1998] Chellapilla, K. "Combining mutation operators in evolutionary programming". *IEEE Trans. on Evolutionary Computation*, volumen 2, número 3.

[Chen et al., 1998] Chen, S.-H., Hwang, S.-H., y Wang, Y.-R. "An RNN-based prosodic information synthesizer for mandarin text-to-speech". *IEEE Trans. on Speech and Audio Processing*, volumen 6, número 3.

[Chiang et al., 1996] Chiang, T.-H., Lin, Y.-C., y Su, K.-Y. "On jointly learning the parameters in a character synchronous integrated speech and language model". *IEEE Trans. on Speech and Audio Processing*, volumen 4, número 3.

[Chih-Heng et al., 1996] Chih-Heng, L., Chien-Hsing, W., Pei-Yih, T., y Hsin-Min, W. "Frameworks for recognition of mandarin syllables with tones using sub-syllabic units". *Speech Communication*, volumen 18, páginas 175–190.

[Chung y Seneff, 1998] Chung, G. y Seneff, S. "Improvements in speech understanding accuracy through the integration of hierarchical linguistic, prosodic, and phonological contraints in the jupiter domain". En *Proceedings of the 5th International Conference on Spoken Language Processing*. Spoken Language Understanding Systems 1.

[Cingolani y Houssay, 1988a] Cingolani, H. E. y Houssay, A. B. *Fisiología Humana*, volumen 1. El Ateneo, Buenos Aires, 6° edición.

[Cingolani y Houssay, 1988b] Cingolani, H. E. y Houssay, A. B. *Fisiología Humana*, volumen 2. El Ateneo, Buenos Aires, 6° edición.

[Davis y Mermelstein, 1980] Davis, S. B. y Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in contunuously spoken sentences". *IEEE Trans. on Acoust. Speech, Signal Processing*, volumen 28, número 4, páginas 357–366.

[Deller et al., 1993] Deller, J. R., Proakis, J. G., y Hansen, J. H. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing, NewYork.

[Diaz et al., 1998] Diaz, J. E., Peinado, A. M., Rubio, A. J., Segarra, E., Prieto, N., y Casacuberta, F. "Albayzin: A task-oriented spanish speech corpus". En *Proceedings of the 1st International Conference in Language Resources and Evaluation*, volumen 1, páginas 497–501, Granada.

[Diaz et al., 1993] Diaz, J. E., Rubio, A. J., Peinado, A. M., Segarra, E., Prieto, N., y Casacuberta, F. "Development of a task-oriented spanish speech corpora". En *Proceedings of the 2th European Conference of Speech Communication and Technology*, Berlin.

[Ducrot y Todorov, 1984] Ducrot, O. y Todorov, T. *Diccionario enciclopédico de las ciencias del lenguaje*. Siglo Veintiuno, Mexico, 10° edición.

[Duda et al., 1999] Duda, R. O., Hart, P. E., y Stork, D. G. *Pattern Classification*. John Wiley and Sons, 2° edición.

[Ferguson, 1980] Ferguson, J. *Hidden Markov Models for Speech*. IDA, Princeton, NJ.

[Gallwitz et al., 1998] Gallwitz, F., Batliner, A., Buckow, J., Huber, R., Niemann, H., y Nöth, E. "Integrated recognition of words and phrase boundaries". En *Proceedings of 5th International Conference on Spoken Language Processsing*, páginas 328–331, Sydney.

[Goldberg, 1997] Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley.

[Gray, 1984] Gray, R. "Vector quantization". *IEEE Acoustics Speech and Signal Processing Magazine*, volumen 4, páginas 4–29.

[Hemert, 1991] Hemert, J. V. "Automatic segmentation of speech". *IEEE Trans. on Signal Processing*, volumen 39, número 4, páginas 1008–1012.

[Hess, 1991] Hess, W. J. "Pitch and voicing determination". En Furui, S. y Sondhi, M. M., editores, *Advances in Speech Signal Processing*, páginas 3–48. Marcel-Dekker, New York.

[Hirose y Iwano, 1998] Hirose, K. y Iwano, K. "Accent type recognition and syntactic boundary detection of japanese using statistical modeling of moraic transitions of fundamental frequency contours". En *Proceedings of the IEEE 23rd International Conference on Acoustics, Speech and Signal Processing*, volumen 1, páginas 25–28, Seattle.

[Hirose y Iwano, 2000] Hirose, K. y Iwano, K. "Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition". En *Proceedings of the IEEE 25rd International Conference on Acoustics, Speech and Signal Processing*, volumen 3, páginas 1763–1766.

[Hoskins, 1997] Hoskins, S. "The prosody of broad and narrow focus in english: Two experiments". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 2, páginas 791–794.

[Huang et al., 1990] Huang, X. D., Ariki, Y., y Jack, M. A. *Hidden Markov Models for Speech Recognition*. Edinburgh University Press.

[Humphries y Woodland, 1998] Humphries, J. J. y Woodland, P. C. "The use of accent-specific pronunciation dictionaries in acoustic model training". En *Proceedings of the IEEE 23rd International Conference on Acoustics, Speech and Signal Processing*, volumen 1, páginas 317–320.

[Iparraguirre y Torres, 1996] Iparraguirre, P. y Torres, M. I. "Acoustic parameters for place of articulation identification and classification of spanish unvoiced stops". *Speech Communication*, volumen 18, páginas 369–379.

[Jelinek, 1999] Jelinek, F. *Statistical Methods for Speech Recognition*. MIT Press, Cambrige, Masachussets.

[Jeong y Jeong, 1996] Jeong, C. y Jeong, H. "Automatic phone segmentation and labelling of continuous speech". *Speech Communication*, volumen 20, páginas 291–311.

[Junqua y Haton, 1996] Junqua, J. C. y Haton, J. P. *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic Publishers.

[Kay y Marple, 1981] Kay, S. M. y Marple, S. L. "Spectrum analisis". En *Proceedings of the IEEE*, volumen 69, páginas 1380–1419.

[Kohonen, 1990] Kohonen, T. "The self-organizing map". *Proceedings of the IEEE*, volumen 78, número 9, páginas 1464–1480.

[Kohonen, 1995] Kohonen, T. *The Self-Organizing Map*. Springer-Verlag.

[Kohonen et al., 1984] Kohonen, T., Makisara, K., y Saramaki, T. "Phonotopics maps - insightful representation of phonological features for speech recognition". En *Proceedings of the IEEE 7th International Conference on Pattern Recognition*, páginas 182–185, Montreal, Canada.

[Koza, 1992] Koza, J. R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press.

[Kuc, 1988] Kuc, R. *Introduction to digital signal processing*. McGraw-Hill Book Company.

[Kuijk y Boves, 1999] Kuijk, V. y Boves, L. "Acoustic characteristics of lexical stress in continuous telephone speech". *Speech Communication*, volumen 27, páginas 95–111.

[Laan, 1997] Laan, G. "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style". *Speech Communication*, volumen 22, páginas 43–65.

[Latarjet y Liard, 1989] Latarjet, M. y Liard, A. R. *Anatomía Humana*, volumen 1. Editorial Médica Panamericana, 2° edición.

[Lee y Hirose, 1999] Lee, S.-W. y Hirose, K. "Dynamic beam-search strategy using prosodic-syntactic information". En *Workshop on Automatic Speech Recognition and Understanding*, páginas 189–192.

[Lee y Ching, 1999] Lee, T. y Ching, P. C. "Cantonese syllable recognition using neural networks". *IEEE Trans. on Speech and Audio Processing*, volumen 7, número 4, páginas 466–472.

[Li y Gibson, 1996] Li, T.-H. y Gibson, J. D. "Speech analysis and segmentation by parametric filtering". *IEEE Trans. on Speech and Audio Processing*, volumen 4, número 3.

[Lieske et al., 1997] Lieske, C., Bos, J., Emele, M., Gambäck, B., y Rupp, C. J. "Giving prosody a meaning". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 3, páginas 1431–1434.

[Liporace, 1982] Liporace, L. A. "Maximum likelihood estimation for multivariate stochastic observations of Markov chains". *IEEE Trans. Information Theory*, volumen 28, número 5.

[Lippmann, 1997] Lippmann, R. P. "Speech recognition by machines and humans". *Speech Communication*, volumen 22, número 1, páginas 1–15.

[Llorach, 1999] Llorach, E. A. *Gramática de la Lengua Española*. Real Academia Española. Colección Nebrija y Bello. Editorial Espasa Calpe, Madrid.

[López et al., 1998] López, E., Caminero, J., Cortázar, I., y Hernández, L. "Improvement on connected numbers recognition using prosodic information". En *Proceedings of the 5th International Conference on Spoken Language Processing*. Prosody and Emotion 2.

[López et al., 1997] López, E., Rodríguez, J. M., Hernández, L., y Villar, J. M. "Automatic corpus-based training of rules for prosodic generation in text-to-speech". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 5, páginas 2515–2518.

[Lublinskaja y Sappok, 1996] Lublinskaja, V. y Sappok, C. "Speaker attribution of successive utterances: The role of discontinuities in voice characteristics and prosody". *Speech Communication*, volumen 19, páginas 145–159.

[Makhoul, 1975] Makhoul, J. "Linear prediction: A tuturial review". En *Proceedings of the IEEE*, volumen 63, páginas 561–580.

[Manrique, 1980] Manrique, A. M. B. *Manual de Fonética Acústica*. Hachette, Buenos Aires.

[Marini, 1989] Marini, J. "Recent advances in speech processing". En *Proceedings of the IEEE International Conference on Acoustic, Speech & Signal Processing*, volumen 1, páginas 429–440.

[Merelo et al., 2000] Merelo, J. J., Carpio, J., Castillo, P., Rivas, V. M., Romero, G., y Schoenauer, M. "Evolving objects". En *Third International Workshop on Frontiers in Evolutionary Algorithms*, Atlantic City.

[Michalewicz, 1992] Michalewicz, Z. *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag.

[Michie et al., 1994] Michie, D., Spiegelhalter, D., y Taylor, C. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, University College, London.

[Milone et al., 2002] Milone, D. H., Merelo, J. J., y Rufiner, H. L. "Evolutionary algorithm for speech segmentation". En *Proceedings of the 2002 IEEE World Congress on Computational Intelligence*, páginas 741–744, Honolulu, HI. Paper No. 7270.

[Milone y Rubio, 2003] Milone, D. H. y Rubio, A. J. "Prosodic and accentual information for automatic speech recognition". *IEEE Trans. on Speech and Audio Processing*. (Por aparecer).

[Milone et al., 1998a] Milone, D. H., Sáez, J. C., Simón, G., y Rufiner, H. L. "Árboles de redes neuronales autoorganizativas". *Revista Mexicana de Ingeniería Biomédica*, volumen 19, número 4, páginas 13–26.

[Milone et al., 1998b] Milone, D. H., Sáez, J. C., Simón, G., y Rufiner, H. L. "Self–organizing neural tree networks". En *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, volumen 3, páginas 1348–1351, Hong Kong.

[Molloy y Isard, 1998] Molloy, L. y Isard, S. "Suprasegmental duration modeling with elastic contraints in automatic speech recognition". En *Proceedings of the 5th International Conference on Spoken Language Processing*. Hidden Markov Model Techniques 3.

[Moreno et al., 1993] Moreno, A., Poch, D., Bonafonte, A., E.Lleida, J.Llisterri, J.B.Marino, y Nadeu, C. "Albayzin speech data base: design of the phonetic corpus". En *Proceedings of the 2th European Conference of Speech Communication and Technology*, páginas 175–178, Berlin.

[Ney y Ortmanns, 1999] Ney, H. y Ortmanns, S. "Dynamic programming search for continuous speech recognition". *IEEE Signal Processing Magazine*, volumen 16, número 5, páginas 64–83.

[Noll, 1967] Noll, A. M. "Cepstrum pitch determination". *Journal of the Acoustic Society of America*, volumen 41, páginas 293–309.

[Nöth et al., 2000] Nöth, E., Batliner, A., Kießling, A., Kompe, R., y Niemann, H. "Verbmobil: The use of prosody in the linguistic components of a speech understanding system". *IEEE Trans. on Speech and Audio Processing*, volumen 8, número 5, páginas 519–532.

[Olaszy y Németh, 1997] Olaszy, G. y Németh, G. "Prosody generation for german CTS/TTS systems (from theoretical intonation patterns to practical realisation)". *Speech Communication*, volumen 21, páginas 37–60.

[Oppenheim y Schafer, 1989] Oppenheim, A. V. y Schafer, R. W. *Discrete-Time Signal Processing*. Prentice-Hall, Inc., Englewood Cliffs, NJ.

[Pallier et al., 1997] Pallier, C., Cutler, A., y Sebastián-Gallés, N. "Prosodic structure and phonetic processing: A cross-linguistic study". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 4, páginas 2131–2134.

[Pauws et al., 1996] Pauws, S., Kamp, Y., y Willens, L. "A hierarchical method of automatic segmentation for synthesis applications". *Speech Communication*, volumen 19, páginas 207–220.

[Pierrehumberg, 1980] Pierrehumberg, J. B. *The phonology and phonetics of English intonation*. Ph.D. thesis, MIT, Cambridge, Massachusetts.

[Pols et al., 1996] Pols, L. C. W., Wang, X., y Bosch, L. F. M. "Modeling of phone duration (using the TIMIT database) and its pottential benefit for ASR". *Speech Communication*, volumen 19, páginas 161–176.

[Portele y Heuft, 1997] Portele, T. y Heuft, B. "Towards a prominence-based synthesis system". *Speech Communication*, volumen 21, páginas 61–72.

[Potamianos y Jelinek, 1998] Potamianos, G. y Jelinek, F. "A study of n-gram and decision tree letter language modeling methods". *Speech Communication*, volumen 24, páginas 171–192.

[Potisuk et al., 1999] Potisuk, S., Harper, M. P., y Gandour, J. "Classification of thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method". *IEEE Trans. on Speech and Audio Processing*, volumen 7, número 1.

[Press et al., 1997] Press, W., Teukolsky, S., Vetterling, W., y Flannery, B. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 2° edición.

[Quilis, 1993] Quilis, A. *Tratado de Fonología y Fonética Españolas.* Biblioteca Románica Hispánica. Editorial Gredos, Madrid.

[Quinlan, 1993] Quinlan, J. R. *C4.5: Programs for Machine Learning.* The Morgan Kaufmann Series in Machine Learning.

[Rabiner y Gold, 1975] Rabiner, L. R. y Gold, B. *Theory and Application of Digital Signal Processing.* Prentice Hall.

[Rabiner y Juang, 1986] Rabiner, L. R. y Juang, B. H. "An introduction to hidden Markov models". *IEEE Acoustics Speech and Signal Processing Magazine*, volumen 3, número 1, páginas 4–16.

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

[Rajendran y Yegnanarayana, 1996] Rajendran, S. y Yegnanarayana, B. "Word boundary hypothesization for continuous speech in Hindi based on F0 patterns". *Speech Communication*, volumen 18, páginas 21–46.

[Reddy, 1966] Reddy, D. R. "An approach to computer speech recognition by direct analysis of the speech wave". Reporte técnico CS59, Computer Science Department, Stanford University.

[Ross y Ostendorf, 1999] Ross, N. K. y Ostendorf, M. "A dynamical system model for generating fundamental frequency for speech synthesis". *IEEE Trans. on Speech and Audio Processing*, volumen 7, número 3.

[Rossi, 1997] Rossi, M. "Is syntactic structure prosodically retrievable?". En *Proceedings of the 5th European Conference on Speech Communication and Technology.* Keynote Speech.

[Rouvière y Delmas, 1988a] Rouvière, H. y Delmas, A. *Anatomía Humana. Descriptiva, Topográfica y Funcional. Cabeza y Cuello.*, volumen 1. Masson, Barcelona, 9° edición.

[Rouvière y Delmas, 1988b] Rouvière, H. y Delmas, A. *Anatomía Humana. Descriptiva, Topográfica y Funcional. Tronco.*, volumen 2. Masson, Barcelona, 9° edición.

[Salomon, 1998] Salomon, R. "Evolutionary algorithms and gradient search: Similarities and differences". *IEEE Trans. on Evolutionary Computation*, volumen 2, número 2, páginas 45–55.

[Sestito y Dillon, 1994] Sestito y Dillon. *Automated Knowledge Acquisition.* Prentice Hall.

[Shimamura y Kobayashi, 2001] Shimamura, T. y Kobayashi, H. "Weighted autocorrelation for pitch extraction of noisy speech". *IEEE Trans. on Speech and Audio Processing*, volumen 9, número 7, páginas 727–730.

[Sönmez et al., 1997] Sönmez, M. K., Heck, L., Weintraub, M., y Shriberg, E. "A lognormal tied mixture model of pitch for prosody based speaker recognition". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 3, páginas 1391–1394.

[Sosa, 1999] Sosa, J. M. *La Entonación en el Español. Su estructura Fónica, Variabilidad y Dialectología.* Editorial Cátedra, Madrid.

[Stevens, 1998] Stevens, K.Ñ. *Acoustic Phonetics.* MIT Press.

[Stolcke et al., 1999] Stolcke, A., Shriberg, E., Hakkani-Tür, D., y Tür, G. "Modeling the prosody of hidden events for improved word recognition". En *Proceedings of the 7th European Conference on Speech Communication and Technology*, volumen 1, páginas 311–314.

[Strangert, 1997] Strangert, E. "Relating prosody to syntax: Boundary signalling in swedish". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 1, páginas 239–242.

[Svendsen y Soong, 1987] Svendsen, T. y Soong, F. K. "On the automatic segmentation of speech signals". En *Proceedings of the IEEE International Conference on Acoustic and Signal Processing*, volumen 1, páginas 77–80, Dallas, Texas.

[Swerts y Ostendorf, 1997] Swerts, M. y Ostendorf, M. "Prosodic and lexical indications of discourse structure in human-machine interactions". *Speech Communication*, volumen 22, número 25-41.

[Torre-Vega, 1999] Torre-Vega, A. *Técnicas de Mejora de la Representación en los Sistemas de Reconocimiento Automático de Voz.* Sc.D. thesis, Universidad de Granada, Granada, España.

[Van Santen, 1997] Van Santen, J. P. H. "Prosodic modeling in text-to-speech synthesis". En *Proceedings of the 5th European Conference on Speech Communication and Technology*. Keynote Speech.

[Vereecken et al., 1997] Vereecken, H., Vorstermans, A., Martens, J. P., y Van Coile, B. "Improving the phonetic annotation by means of prosodic phrasing". En *Proceedings of the 5th European Conference on Speech Communication and Technology*, volumen 1, páginas 179–182.

[Vorstermans et al., 1996] Vorstermans, A., Martens, J.-P., y Van Coile, B. "Automatic segmentation and labelling of multi-lingual speech data". *Speech Communication*, volumen 19, páginas 271–293.

[Véronis et al., 1998] Véronis, J., Di Cristo, P., Courtois, F., y Chaumette, C. "A stochastic model of intonation for text-to-speech synthesis". *Speech Communication*, volumen 26, páginas 233–244.

[Waibel et al., 1989] Waibel, A. H., Hanazawa, T., Hiton, G., Shikano, K., y Lang, K. "Phoneme recognition using time-delay neural networks". *IEEE Trans. on Acoustic Speech and Signal Processing*, volumen 37, número 3, páginas 328–339.

[Wang y Seneff, 1998] Wang, C. y Seneff, S. "A study of tones and tempo in continuous mandarin digit strings and their application in telephone quality speech recognition". En *Proceedings of the 5th International Conference on Spoken Language Processing*. Prosody and Emotion 2.

[Warnke et al., 1999] Warnke, V., Gallwitz, F., Batliner, A., Buckow, J., Huber, R., Nöth, E., y Höthker, A. "Integrating multiple knowledge sources for word hypotheses graph interpretation". En *Proceedings of 7th European Conference on Speech Communication and Technology*, volumen 1, páginas 235–238.

[Wu et al., 1998] Wu, S.-L., Kingsbury, B., Morgan, N., y Greenberg, S. "Incorporating information from syllable-length time scales into automatic speech recognition". En *Proceedings of the IEEE 23rd International Conference on Acoustics, Speech and Signal Processing*, volumen 2, páginas 721–724, Seattle.

[Yaeger-Dror, 1996] Yaeger-Dror, M. "Register as a variable in prosodic analysis: The case of the English negative". *Speech Communication*, volumen 19, número 39-60.

[Ying, 1998] Ying, G. S. *Automatic measurement and representation of prosodic features*. Ph.D. thesis, Purdue University, Purdue.

[Young et al., 2000] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., y Woodland, P. *HMM Toolkit*. Cambridge University, http://htk.eng.cam.ac.uk.

Esta Tesis fue escrita en LaTeX, compilada con MiKTeX y editada en TeXnicCenter.