Biblioteca Digital
F C E N - U B A

BIBLIOTECA CENTRAL LUIS F LELOIR
BIBLIOTECA CENTRAL LELOIR
FACULTAD DE CIENCIAS EXACTAS Y NATURALES UBA

Tesis Doctoral

Incorporación de información suprasegmental en el proceso de reconocimiento automático del habla

Evin, Diego Alexis

2011

Este documento forma parte de la colección de tesis doctorales y de maestría de la Biblioteca Central Dr. Luis Federico Leloir, disponible en digital.bl.fcen.uba.ar. Su utilización debe ser acompañada por la cita bibliográfica con reconocimiento de la fuente.

This document is part of the doctoral theses collection of the Central Library Dr. Luis Federico Leloir, available in digital.bl.fcen.uba.ar. It should be used accompanied by the corresponding citation acknowledging the source.

Cita tipo APA:
Evin, Diego Alexis. (2011). Incorporación de información suprasegmental en el proceso de reconocimiento automático del habla. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.

Cita tipo Chicago:
Evin, Diego Alexis. "Incorporación de información suprasegmental en el proceso de reconocimiento automático del habla". Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2011.

EXACTAS UBA
Facultad de Ciencias Exactas y Naturales

UBA
Universidad de Buenos Aires

Dirección: Biblioteca Central Dr. Luis F. Leloir, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires. Intendente Güiraldes 2160 - C1428EGA - Tel. (++54 +11) 4789-9293

Contacto: digital@bl.fcen.uba.ar

UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Incorporación de Información Suprasegmental en el Proceso de Reconocimiento Automático del Habla

Tesis presentada para optar al título de Doctor
de la Universidad de Buenos Aires
en el área Ciencias de la Computación

Diego Alexis Evin

Director: Jorge Alberto Gurlekian

Consejera de Estudios: Ana Ruedin

Lugar de trabajo: Laboratorio de Investigaciones Sensoriales - UBA - CONICET

Buenos Aires, 2011

Diego Alexis Evin: *Incorporación de Información Suprasegmental en el Proceso de Reconocimiento Automático del Habla,*
Tesis Doctoral
Director: Jorge Alberto Gurlekian

Coordinadora de Estudios: Ana Ruedin
© Abril de 2011.

WEBSITE:
http://www.lis.secyt.gov.ar/

E-MAIL:
diegoevin@gmail.com

Incorporación de Información Suprasegmental en el Proceso de Reconocimiento Automático del Habla

RESUMEN

Desarrollar sistemas informáticos capaces de interactuar con sus usuarios de la forma más natural y eficiente posible es uno de los requisitos esenciales para lograr la integración del mundo tecnológico en la sociedad.

En ese marco el habla se presenta como una de las formas de comunicación más eficientes y naturales que posee el ser humano. Es por ello que desde el origen mismo de la investigación en ciencias de la computación, el desarrollo de interfaces hombre-máquina a través de la voz ha despertado un gran interés.

Uno de los elementos que componen dicha interfaz oral es el **Reconocimiento Automático del Habla** (RAH), área de la Inteligencia Artificial que busca desarrollar sistemas computacionales capaces de transformar un fragmento de habla en su trascripción textual.

El RAH es un problema de gran complejidad, lo que se puede atribuir principalmente a dos factores: en primer lugar a la variabilidad de la señal de habla, que responde a múltiples factores como características particulares del locutor y medio acústico donde se registra, la velocidad y estilos de elocución; y en segundo lugar a la necesidad de encontrar palabras individuales en un continuo acústico, es decir realizar al mismo tiempo las tareas de segmentación y clasificación.

Si bien se pueden encontrar en los últimos años avances significativos en el desempeño de los sistemas de RAH, aún hay mucho por mejorar en relación a la capacidad de reconocimiento que presentan los oyentes humanos para las mismas tareas y bajo las mismas condiciones. Varias hipótesis intentan explicar esta diferencia de desempeño: información insuficiente o representada de manera inadecuada en los sistemas automáticos, problemas en el modelado del sistema de reconocimiento, insuficientes cantidades de ejemplos empleados para lograr tasas de reconocimiento similares, etc.

Con respecto al primero de estos puntos, los sistemas de RAH no utilizan toda la información acústica disponible en la señal de habla. Dichos sistemas interpretan el habla como secuencias de unidades cuyas duraciones se encuentran a nivel segmental (fonético). Por lo tanto procesan la información acústica en la escala segmental para obtener las hipótesis de secuencias de unidades emitidas. Sin embargo estudios tanto psicoacústicos como psicolingüísticos resaltan el rol crucial que posee la información de una escala temporal mayor: la información suprasegmental, en la percepción humana. Se entiende por información suprasegmental toda aquella que está dada en segmentos de duración superior al fonético, y cuyas propiedades están determinadas principalmente por la prosodia de una frase.

Además se argumenta que en la tarea de reconocimiento e interpretación del habla los seres humanos emplean e integran varios niveles de conocimiento lingüístico, muchos de los cuales aún no han sido incorporados o aprovechados eficientemente en el RAH.

A partir de esas evidencias resulta interesante investigar cuál es el aporte que puede brindar la información suprasegmental o prosódica para mejorar el desempeño de los sistemas de RAH estándar.

En esta Tesis se investiga el empleo de información suprasegmental como factor de mejora en el desempeño, así como alternativas para su integración en sistemas de RAH estándar.

En el **Capítulo** 1 se exponen argumentos que muestran la necesidad de mejorar los sistemas de RAH actuales a la luz del desempeño mostrado en esta tarea por los seres humanos. Se presentan las bases de los mecanismos de producción, percepción y reconocimiento humano, así como un resumen de las principales aproximaciones al reconocimiento automático. Posteriormente se introducen los aspectos generales de la información suprasegmental y su rol en el mecanismo de comunicación oral. Seguidamente se hace una revisión de los antecedentes en el empleo de información suprasegmental dentro del proceso de RAH. Finalmente se delinean los objetivos de esta tesis.

Debido a que en esta tesis se analiza la utilización de información suprasegmental en distintos módulos de los sistemas de RAH convencionales, en el **Capítulo** 2 se presenta la arquitectura y componentes principales de los reconocedores del habla actuales. Se detalla la arquitectura y forma de funcionamiento de estos sistemas, los fundamentos teóricos de los modelos de Markov, así como la forma de medir sus desempeños.

El **Capítulo** 3 profundiza los aspectos de la información suprasegmental introducidos en el **Capítulo** 1. Hace principal hincapié en las características prosódicas del español de Argentina, y presenta las técnicas computacionales empleadas en la tesis para la extracción automática de sus atributos a partir de la señal de habla.

El **Capítulo** 4 contiene una serie de estudios en que se busca vincular patrones de los atributos suprasegmentales con información lingüística útil para el proceso de RAH. En el primero de estos estudios se analiza la posibilidad de establecer agrupamientos de frases entonativas a partir de semejanzas en sus atributos suprasegmentales. El segundo experimento indaga la posibilidad de obtener información del número de palabras de contenido presentes en una frase, a partir de la morfología de sus curvas de F0. Finalmente el tercer estudio explora la viabilidad de establecer la tipología acentual de las palabras finales de frase utilizando rasgos suprasegmentales.

En el **Capítulo** 5 se presenta una metodología para utilizar información suprasegmental a nivel de los modelos acústicos de un sistema de RAH. Específicamente se realiza una distinción entre modelos acústicos correspondientes a sonidos vocálicos acentuados y no acentuados. La metodología propuesta se evalúa y contrasta con distintas versiones de sistemas de RAH convencionales, empleando un corpus de habla continua.

En el **Capítulo** 6 se expone una alternativa para emplear información suprasegmental durante la selección de hipótesis de reconocimiento. Esta alternativa contempla la definición de un índice de semejanza entonativa entre la curva de F0 correspondiente a la frase a reconocer, y las posibles curvas de F0 correspondientes a las hipótesis de reconocimiento y obtenidas mediante un proceso de predicción. Se propone y desarrolla un modelo para su implementación y se realizan comparaciones de desempeño con respecto a un sistema de RAH de referencia.

Finalmente en el **Capítulo** 7 se presentan las conclusiones y aportes de la tesis, juntamente con posibles líneas de investigación futura.

Palabras Clave:

Prosodia • Entonación • Acentuación • Modelos Ocultos de Markov • Reconocimiento Automático del Habla.

Incorporation of Suprasegmental Information into Automatic Speech Recognition Process

ABSTRACT

The development of computational systems capable of interacting with users in the most natural and efficient way is one of the essential requirements for the integration of the technological world in society.

In this context speech is presented as one of the most efficient form of communication mechanisms available for human beings. That is why from the very beginning of research in computer science, the development of human-machine interfaces through voice have gain great interest.

One of the elements that compose such interfaces is the **Automatic Speech Recognition** (ASR). ASR is a field of Artificial Intelligence which searches for the development of computational systems that transform speech segments into text transcriptions.

ASR is a very complex problem, which can be attributed mainly to two factors: first, to the huge variability of the speech signal, depending on multiple factors such as the speaker, the acoustic environment, linguistic context, speech rate, emotional states, locution styles, and many others; and secondly to the need of finding isolated words in an acoustic continuum, that is to say solving segmentation and classification problems simultaneously.

Even thought we can find significant advances in the performance of ASR systems in recent years, there is still much space for improvement to match human recognition ability for the same tasks under the same conditions.

Several hypotheses attempt to explain these differences on performance: insufficient information, inadequate way to represent it, problems in modelling, insufficient quantities of used examples to achieve similar recognition rates, etc.

Regarding the first point, ASR systems do not use all available acoustic information in speech signal. These systems interpret the speech as sequences of units whose durations spans in a segmental (phonetic) level. Therefore they process the acoustic information at a segmental scale to obtain the hypotheses of sequences of uttered units.

Nevertheless psychoacoustic and psycholinguistic research emphasize the essential role of information at a higher temporal level for the human speech perception: the suprasegmental information. Any information whose duration spans over several phonetic units can be thought as suprasegmental, and its properties are determined principally by the prosody of an utterance.

Furthermore, it is argued that during the task of speech recognition and interpretation, various linguistic knowledge are integrated and used. It has been also argued that no much of linguistic knowledge have yet been incorporated or utilized efficiently in the ASR

From these evidences it seems relevant to investigate whether the suprasegmental or prosodic information could contribute to improve the performance of standard ASR systems.

In this thesis the use of the suprasegmental information is investigated as a factor for improving performance, as well as an alternative for the integration of this information into the architecture of standard ASR systems.

Chapter 1 arguments are presented that show the need to improve current ASR systems in the light of the performance showed by human speech recognition.
The basis and mechanisms of production, perception and human speech recognition, as well as main approaches for ASR are revised. Subsequently the general aspects of the suprasegmental information and its roll in the mechanism of oral communication are introduced. After that, a review of the employment of information into the process of ASR is devised. Finally, we discuss the objectives of this thesis.

Because this thesis examines the use of suprasegmental information in different modules of the conventional ASR systems, **Chapter** 2 presents the architecture and main components of current speech recognizers. The functionality, theoric foundations of hidden Markov models, and the performance evaluation methodology are detailed.

Chapter 3 discusses in more detail the aspects of the suprasegmental information introduced in **Chapter** 1. Special emphasis is given to the prosodic characteristics of Argentinian Spanish. The computational techniques employed in this dissertation for the automatic extraction of these attributes from the speech signal are presented.

Chapter 4 contains a series of studies that seeks to link attributes of suprasegmental patterns with linguistic information, useful in the ASR process. In the first experiment, the possibility of establishing groupings of intonative phrases based on similarities between their suprasegmental attributes is analyzed
The second experiment explores the possibility of obtaining information about the number of content words contained in a phrase by analyzing the shape the of F0 curves.
Finally the third study explores the feasibility of establishing the accentual typology of final words in sentences using suprasegmental features.

Chapter 5 introduces a methodology for using suprasegmental information at the level of acoustic models in a conventional ASR system. Specifically a distinction is made between acoustic models for accented and unaccented vowel sounds. The proposed methodology is evaluated and compared with different versions of conventional of ASR systems, using a corpus of continuous speech.

Chapter 6 exposes an alternative for using suprasegmental information in the posprocessing of ASR.
This alternative defines an index of intonative similarity, measured between the F0 contour estimated from the utterance to recognize, and the one corresponding to the alternative recognition hypotheses.

This idea is implemented and compared to a reference standard ASR system.

Finallly **Chapter** 7 concludes the Thesis. An overview of the main findings and contributions of this thesis is presented together with future works in this line of research.

Key Words:

Prosody • Intonation • Stress Patterns • Hidden Markov Models • Automatic Speech Recognition.

AGRADECIMIENTOS

Llegar al final de esta tesis no hubiese sido posible sin el apoyo y aliento de una gran cantidad de personas.

En primer lugar, quiero dedicar la tesis a Yami por su amor y paciencia.

Además deseo agradecer a mis padres por brindarme las posibilidades de estudiar, así como por su apoyo incondicional. A mis hermanas por su afecto y confianza.

Respecto al aspecto técnico, quisiera decir gracias a Jorge quien me dirigió en esta tesis, su guía y soporte fueron fundamentales. A los miembros del *Laboratorio de Investigaciones Sensoriales*: Miguelina, Humberto, Natalia, Eduardo, Amalia, Pedro, Alejandro, Reina, Agustín, Christian y Miguel por su fraternidad, y conformar un grupo en el que trabajar resultó placentero.

Agradezco a Anita quien fue mi consejera de estudios, por su generosidad y calidez.

Además quisiera reconocer a Diego Milone, del grupo de *Señales, Sistemas e Inteligencia Computacional*, de la Universidad Nacional del Litoral, por brindarme su valiosa referencia, criterio y ayuda.

También al grupo de *Inteligencia Artificial Aplicada* de la Univeridad Nacional de Entre Ríos: Bartolomé, Alejandro, Rubén, Guille, por haberme permitido iniciar el apasionante camino de la investigación.

Finalmente pero sobre todo a Dios, por darme salud y sustento en este proceso.

ÍNDICE GENERAL

experiencias, intercambiar ideas, actuar coordinadamente, y transmitir conocimiento a través de las generaciones. Si bien el acto de comunicación empleando el habla se lleva a cabo de manera natural, y sin esfuerzos aparentes, este proceso es altamente complejo, como eficiente. Esa eficiencia se expresa en su robustez ante diferencias de voces, hábitos o estilos de habla, dialectos y acentos de locutores, ruidos, distorsiones e interferencias.

Se puede ver la comunicación humana como una cadena de eventos que vinculan el cerebro del locutor y del oyente, y que se conoce como cadena del habla [34] (figura 1).

cadena del habla

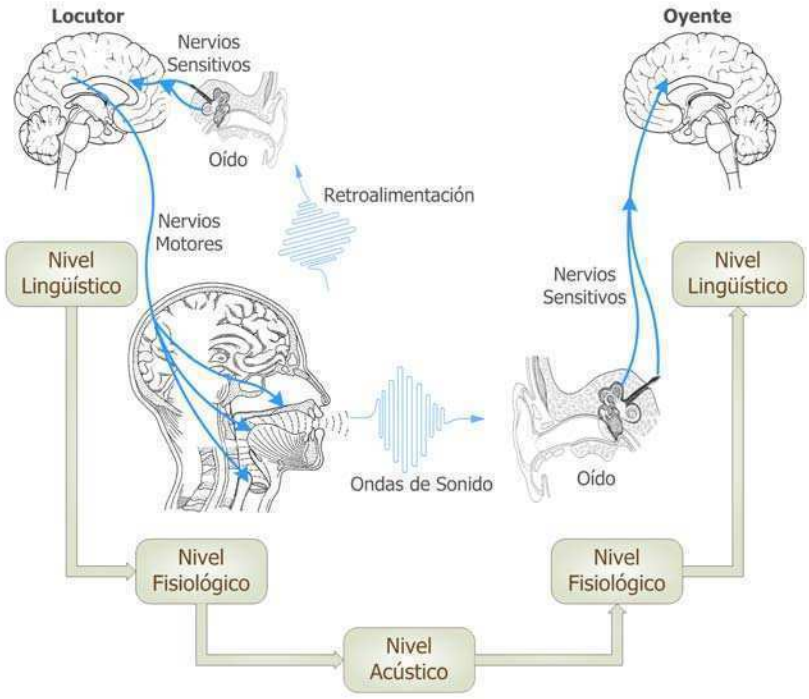

Figura 1: Representación de la cadena del habla.

Considere por ejemplo la interacción entre dos personas, donde el locutor quiere transmitir información al oyente empleando el habla. Lo primero que debe hacer es ordenar sus pensamientos, establecer qué es lo que quiere decir y transformar esa idea en un arreglo de unidades lingüísticas a través de la selección de secuencias de palabras y frases.

Tal ordenamiento debe seguir restricciones gramaticales del lenguaje usado. Todo este proceso de imaginación, selección y ordenamiento de unidades de acuerdo a normas del lenguaje se lleva a cabo en el cerebro del locutor casi instantáneamente. A este nivel se lo pude llamar **nivel lingüístico** de la cadena de habla.

También a nivel del cerebro del locutor se codifica el mensaje planificado en instrucciones motoras. A través de los impulsos eléctricos correspondientes a esas instrucciones motoras, los nervios activan y coordinan la actividad muscular de los órganos vocales: los pulmones, las cuerdas vocales, la lengua, las mandíbulas, los labios.

Este es el **nivel fisiológico**, ya que involucra a los eventos de actividad neuronal y muscular del locutor.

Finalmente, el efecto de la activación del aparato fonador provoca patrones de variación en la presión del aire, denominadas ondas de sonido, que se propagan hasta los oídos del oyente, y del propio locutor. Al hablar el locutor oye lo que articula, y emplea esa información para retroalimentar su mecanismo de fonación, comparando la calidad de los sonidos que intentó producir con los que realmente produjo y efectuando ajustes necesarios para minimizar las diferencias.

Este proceso de generación y transmisión de las ondas acústicas se corresponden con el **nivel físico o acústico**.

Del lado del oyente, los patrones de presión de aire que fueron provocados por la fonación del locutor, provocan movimientos en la membrana timpánica, que se transfieren a través de un sistema mecánico ubicado en el oído medio hasta el oído interno. Ya en el oído interno esos patrones de movimientos se transducen a impulsos eléctricos, que viajan a lo largo de los nervios acústicos hasta el cerebro. La llegada de los impulsos nerviosos modifica la actividad nerviosa que se estaba produciendo en el cerebro del oyente, y a través de un proceso que aún no se conoce completamente, se reconoce el mensaje del locutor. Es decir que del lado del oyente, el proceso se revierte: los eventos comienzan en el nivel físico, donde las ondas acústicas activan el mecanismo auditivo, continúan en el nivel fisiológico, donde se procesa la información auditiva, y termina en el nivel lingüístico en el momento que el oyente reconoce las palabras y frases pronunciadas por el locutor.

Cabe aclarar que el modelo presentado es claramente una simplificación de lo que en realidad ocurre. Por ejemplo, no aparecen los estímulos visuales que también se emplean simultáneamente con los auditivos para reconocer el habla.

1.2.1 Organización Lingüística del Habla

El mensaje antes de ser transmitido al oyente es estructurado lingüísticamente por el locutor. A través de este proceso el locutor selecciona las palabras y oraciones adecuadas para expresar un mensaje.

Se puede considerar que el lenguaje está conformado por una serie de unidades. Estas unidades son símbolos que permiten representar objetos, conceptos o ideas. El lenguaje es un sistema conformado por esos símbolos y reglas para combinarlos en secuencias que permitan expresar pensamientos, intenciones o experiencias, y que aprendemos a identificar y utilizar desde niños.

Se pueden distinguir las siguientes unidades para los distintos niveles lingüísticos:

- ***Fonemas.*** Se utiliza el término fonema para hacer referencia a un sonido contrastivo en una lengua determinada. Por lo tanto se pude definir a un fonema como la unidad mínima de diferenciación del lenguaje. Cada lengua tiene un número reducido de tales sonidos contrastivos. En el Español existen cinco sonidos

vocálicos y menos de veinte sonidos consonánticos (dependiendo del dialecto).

Aunque en sí mismo un fonema no tiene significado alguno, al reemplazar un fonema por otro en una palabra podemos obtener una palabra diferente.

El conjunto de todos los sonidos (fonos) que son aceptados como variantes de un mismo fonema se conoce como alófonos. Por ejemplo, si se analiza la forma en que se pronuncia la palabra *"dedo"* de manera aislada, se puede encontrar que sus dos consonantes, a pesar de corresponder al mismo fonema: *"d"*, se pronuncian de manera diferente. Mientras la primera se realiza apoyando el ápice de la lengua contra los dientes superiores, de manera tal que se impide el paso del aire (consonante oclusiva), la segunda se pronuncia sin llegar a provocar una oclusión total (articulación aproximante). Estos dos sonidos son variantes alofónicas del fonema *"d"*.

Al estudiar la estructura fónica de una lengua se debe considerar tanto los fonemas como sus principales alófonos. Debido a que la pronunciación de todos los sonidos varía dependiendo de aspectos como la influencia de sonidos cercanos, la rapidez de elocución, o el estilo de habla, se puede encontrar un considerable margen de detalle con el que se puede caracterizar a los sonidos. Por ello los lingüistas suelen hacer una distinción entre representación fonética estrecha, en las que se incluye un gran número de detalles de pronunciación, y representación fonética amplia en donde se incluyen los detalles no contrastivos considerados más importantes.

Los hablantes de un idioma tienden a oír solamente las diferencias entre sonidos que son relevantes para distinguir una palabra de otra para esa lengua. Es decir que muchas veces no se distinguen entre sí las variantes alofónicas. Por otro lado, diferencias alofónicas para una lengua pueden ser fonémicas en otra, por ejemplo *"pelo"* y *"pero"* pueden sonar indistintos para el Japonés o Coreano, mientras que para el Español presentan diferencias fonéticas.

Al clasificar los sonidos del habla suele hacerse una distinción entre sonidos consonánticos y vocales. En la articulación de las consonantes se produce una obstaculización u obstrucción al paso de aire procedente de los pulmones. Durante la producción de vocales, en cambio, el aire pasa por la cavidad bucal sin tales obstáculos.

Los sonidos de las vocales se pueden clasificar utilizando tres parámetros, dos que tienen que ver con la posición de la lengua: su altura y su desplazamiento hacia la parte anterior o posterior de la boca, y el tercero vinculado con la posición de los labios.

Teniendo en cuenta la altura del dorso de la lengua se distinguen vocales altas, (/i/ y /u/), vocales medias (/e/ y /o/); y vocales bajas, con descenso del dorso, que en Español es únicamente la /a/.

Según el desplazamiento de la lengua hacia adelante o hacia el velo, tenemos vocales anteriores (/i/ y /e/), una vocal central (/a/), y vocales posteriores, con retracción del dorso (/o/ y /u/).

Finalmente si se considera la disposición de los labios, tenemos dos vocales redondeadas (/o/ y /u/) y tres no redondeadas (/i/, /e/ y /a/).

Los sonidos consonánticos por su parte se suelen clasifican según tres parámetros principales: punto de articulación, modo de articulación y actividad de las cuerdas vocales.

punto de articulación

El **punto de articulación** especifica cuáles son los órganos articulatorios que provocan la oclusión o impedimento a la circulación del aire expelido. En este caso al articulador en movimiento se denomina activo y al que permanece inmóvil o presenta menor movimiento articulador pasivo. Por ejemplo, la *"p"* es un consonante bilabial, la *"t"* es ápicodental, y la *"k"* es velar.

modo de articulación

El **modo de articulación** describe la naturaleza del obstáculo que se encuentra a la salida del aire. De acuerdo a esta clasificación se puede separar a las consonantes en oclusivas, fricativas, africadas, aproximantes, nasales, laterales y vibrantes.

De acuerdo a la **actividad de las cuerdas vocales** se puede diferenciar a los **sonidos sordos**, producidos sin la vibración de las cuerdas vocales, y los **sonidos sonoros**, articulados con vibración de las cuerdas vocales.

- ***Morfemas.*** Son las unidades mínimas de significación del lenguaje. Hay dos tipos de morfema: el lexema, que en un contexto más coloquial se conoce como *"raíz de una palabra"* y es su principal aporte de significado, y el gramema, que es un morfema generalmente pospuesto al lexema para indicar accidentes gramaticales, y pueden ser de género o número. También hay gramemas independientes como las preposiciones.

- ***Lexía.***

  Son unidades léxicas compuestas por morfemas relacionados mediante un alto índice de inseparabilidad, o un agrupamiento estable de semas, que constituyen una unidad funcional.

  Se pueden encontrar: Lexía simple, que consiste en un par de morfemas y se conoce habitualmente como palabra; lexía compuesta, que es lo conocido como palabra compuesta; lexía compleja, compuesta también por varias palabras pero con separación gráfica (ej. *"a duras penas"*, *"Semana Santa"*); y lexía textual, conformada por varias palabras y una mayor elaboración formal, interviene la función poética, como en los refranes.

- ***Sintagma.***

  Constituye la unidad de función. Cada sintagma posee una función sintáctica específica dentro de la oración. Puede ser nominal, verbal, preposicional, adverbial, o adjetival, dependiendo si el

núcleo es un sustantivo, verbo, preposición, adverbio o adjetivo, respectivamente

- ***Oración.***

  Se considera la unidad mínima de comunicación. Corresponde a lo conocido tradicionalmente como oración simple.

- ***Enunciado.***

  Unidad de manifestación. Corresponde a lo conocido habitualmente como oración compuesta o frase.

- ***Secuencia Textual.***

  Es la unidad de intención, dada por el conjunto de actos comunicativos de un hablante en una situación dada.

A su vez el estudio de las unidades lingüísticas, dio lugar a distintas disciplinas:

- ***Fonética:*** estudia la forma en que se producen (articulación), e interpretan los sonidos del lenguaje. Además del estudio del aprendizaje del lenguaje por los niños y personas no nativas, analiza las modificaciones que sufren los sonidos del habla dependiendo de los estilos, contextos fonéticos, regiones geográficas, etc. Además de analizar la actividad de los órganos articulatorios durante la producción de cada sonido (fonética articulatoria), también se puede estudiar los sonidos del habla investigando la estructura de las ondas sonoras producidas (fonética acústica).

- ***Fonología:*** trata los sistemas de sonido relevantes semánticamente del lenguaje. Mientras que la fonética se centra en una base meramente acústica, la fonología tiende a considerar la imagen mental de lo percibido. Los fonemas son objeto de estudio de la fonología, mientras que los fonos son objeto de estudio de la fonética.

- ***Morfología:*** atañe las relaciones entre las formas y los significados de las palabras. Abarca el estudio de la generación y procesos morfológicos como inflexión y derivación de las palabras. Está relacionada tanto con la fonología, al estudiar las formas de los sonidos de las palabras, como con la sintaxis, al estudiar la composición de palabras y su función sintáctica, y con la semántica, por su estudio del significado del lexicón.

- ***Sintaxis:*** se centra en el modelado de la combinación de palabras y morfemas que permiten conformar significados más complejos. Es la parte de la gramática que trabaja en el sistema de reglas que describen cómo se pueden derivar estructuras bien formadas de un lenguaje a partir de elementos básicos.

- ***Semántica:*** analiza y describe el significado literal de las expresiones lingüísticas.

- ***Pragmática:*** estudia las relaciones entre las expresiones del lenguaje natural y sus usos en situaciones específicas, es decir, cómo se ejercita el lenguaje humano como práctica social en la comunidad.

Esta tesis trata especialmente con las unidades lingüísticas de nivel suprasegmental, es decir aquellas de longitud mayor al de los fonemas, y vinculada al área lingüística conocida como prosodia, que involucra tanto a la fonología como a la sintaxis. No obstante, como se verá en el curso de la misma los niveles presentados interactúan entre sí y no es posible tratarlos como entidades completamente independientes.

1.2.2 Producción del Habla

El habla surge de una compleja acción coordinada entre un sistema que almacena y genera presión de aire, un mecanismo que regula la vibración de las cuerdas vocales, y otro que controla un conjunto de resonadores.

Estructuralmente se puede dividir al sistema de producción del habla en tres partes: el sistema que se encuentra por debajo de la laringe, la propia laringe y sus estructuras circundantes, y las estructuras y conductos supra-laríngeos.

Mientras que desde un punto de vista funcional se puede separar los órganos del habla de acuerdo a dos funciones: fonación y articulación.

Los órganos fonatorios comprenden a los pulmones y la laringe, y son los responsables de generar un flujo de aire y las oscilaciones de las cuerdas vocales. Estos dos órganos también son los encargados de producir los patrones prosódicos del habla, ajustando la altura tonal, sonoridad y la calidad vocal.

Durante la inspiración, los músculos respiratorios permiten el ingreso de aire hacia los pulmones. A medida que se va alcanzando la capacidad máxima de aire en los pulmones, la presión aumenta. Ese volumen de aire a presión constituye la fuente de energía del sistema fonador.

El aire proveniente de los pulmones es expelido a una tasa relativamente constante gracias a la acción coordinada de los músculos inspiratorios y expiratorios. Ese flujo de aire encuentra una constricción a la altura de la laringe, en una región denominada glotis.

Las variaciones en la presión de aire que permiten generar los sonidos sonoros son producidas por el movimiento de las cuerdas vocales ubicadas en la región glótica. Es decir que durante la producción de sonidos sonoros las oscilaciones de las cuerdas vocales convierten el aire expiratorio en trenes intermitentes de pulsos de aire, a una frecuencia denominada frecuencia fundamental o F0.

Este movimiento surge por el paso de aire expulsado desde los pulmones, y por un mecanismo activo controlado por un conjunto de músculos laríngeos.

Si las cuerdas vocales se tensan ante el flujo de aire, vibran por efecto Bernoulli, produciendo sonidos sonoros, periódicos o cuasioperiódicos. En esa situación el flujo de aire abre y cierra la glotis y transfiere a

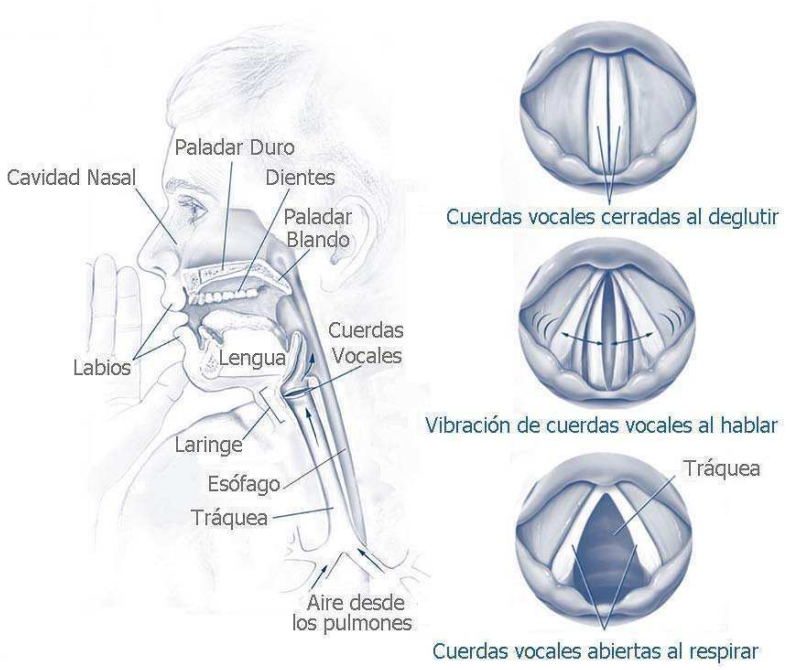

Figura 2: Estructuras y mecanismos involucradas en la producción del habla. Izquierda: componentes fonatorios y articulatorios. Derecha: configuración glótica al deglutir (arriba), pronunciar sonidos sonoros (centro), y al respirar (abajo)

las cuerdas vocales un tipo de movimiento progresivo. Para los tonos bajos, la glotis permanece más tiempo cerrada que abierta (relación 5:1 con 100 Hz). Para los tonos altos (400 Hz), esta relación disminuye a 1,4:1. Al cantar en falsete o susurrar, la glotis permanece más tiempo abierta.

Si por el contrario, las cuerdas vocales se relajan durante el paso del aire, éste continúa hacia el tracto vocal. Si encuentra alguna constricción (parcial o completa), el aire se vuelva turbulento, y genera sonidos sordos y aperiódicos.

Las señales motoras provienen de la corteza motosensorial y alcanzan la zona del núcleo del nervio vago. Éste no sólo realiza la inervación motora de la glotis, sino también la sensitiva: las fibras sensibles de la mucosa de la glotis y las sensoriales de los husos musculares informan a nivel central de la situación y la tensión de las cuerdas vocales. Estos reflejos y la estrecha relación de la vía auditiva con los centros bulbares y corticales de la motricidad del lenguaje son importantes para el ajuste fino de la voz.

Los órganos articulatorios están compuestos por todas las estructuras supra-glóticas: las cavidades de los tractos vocal, nasal y para-nasal, los labios, el velo, los dientes y la lengua. El tracto vocal se puede ver como un conjunto de filtros acoplados, cuya función es dar forma al espectro de la señal proveniente de la laringe (modulación en frecuencia), de acuerdo a las resonancias características de la faringe, la cavi-

dad bucal, la cavidad nasal, y de las posiciones y movimientos de los articuladores: lengua, velo, labios, mandíbula. Además los músculos constrictores de la faringe y laringe también participan en la articulación y la determinación de la calidad vocal. Debe tenerse en cuenta además que los sistemas fonatorio y articulatorio se influencian entre sí mutuamente.

En la figura 3 se pueden ver los modelos mecánico y eléctrico del sistema de fonación humano. En el modelo mecánico el émbolo representa la acción de los músculos respiratorios al comprimir el aire de los pulmones, representado como un fuelle. Los tubos simbolizan a los bronquios, la tráquea y los resonadores. Finalmente se emplean válvulas para simbolizar la acción de las cuerdas vocales y del velo. La configuración de los resonadores se modifica de manera dinámica al hablar, por ejemplo la cavidad bucal se altera a través de la apertura de la boca, los movimientos de la lengua y los labios.

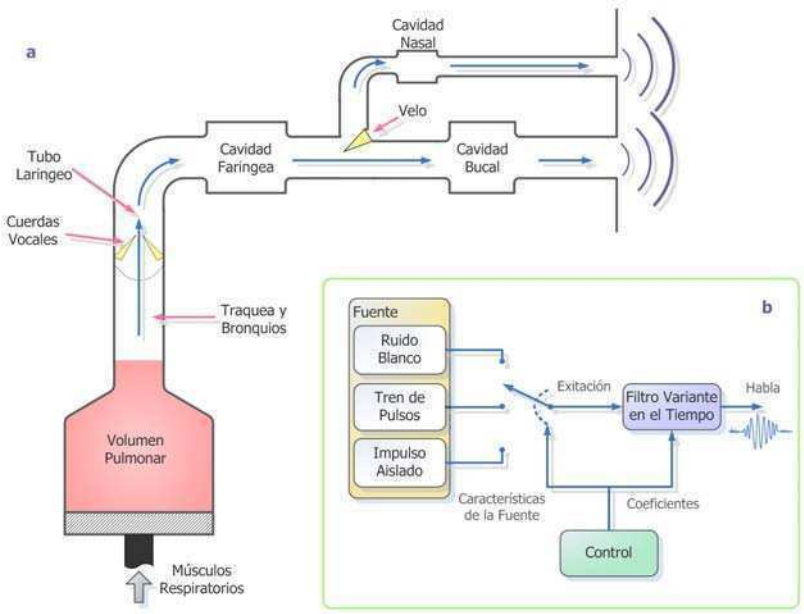

Figura 3: Representación del sistema fonatorio como modelos mecánicos y eléctricos: **a.** modelo mecánico; **b.** modelo eléctrico correspondiente al mecanismo de producción del habla.

Por otra parte el análogo eléctrico se conoce como modelo de fuente-filtro [36], y fue uno de los primeros modelos eléctricos implementados para la síntesis de habla. En este modelo los tubos acústicos con área variante en función del tiempo se caracteriza mediante un filtro eléctrico, cuyos coeficientes varían en función del tiempo. Es decir, se ve a los resonadores mecánicos como filtros que dan forma al espectro de la señal de excitación. La entrada al filtro es una mezcla de una señal cuasi-periódica y una fuente de ruido. Cuando el filtro se excita mediante la señal de entrada, la salida es un voltaje que se puede asimilar a la onda de presión generada al fonar.

1.2.3 Percepción del Habla

En la figura 4 se muestra un esquema del funcionamiento del oído. Las ondas sonoras llegan al órgano auditivo principalmente a través del pabellón y el conducto auditivo externo, terminando en el tímpano (oído externo); aunque el sonido también produce vibraciones en el cráneo que se transmiten directamente a la cóclea a través de la conducción ósea. Las ondas de presión sonoras hacen que vibre la membrana timpánica, y que se transmitan esas vibraciones por los huecesillos del oído medio a la membrana oval, justamente donde comienza el oído interno.

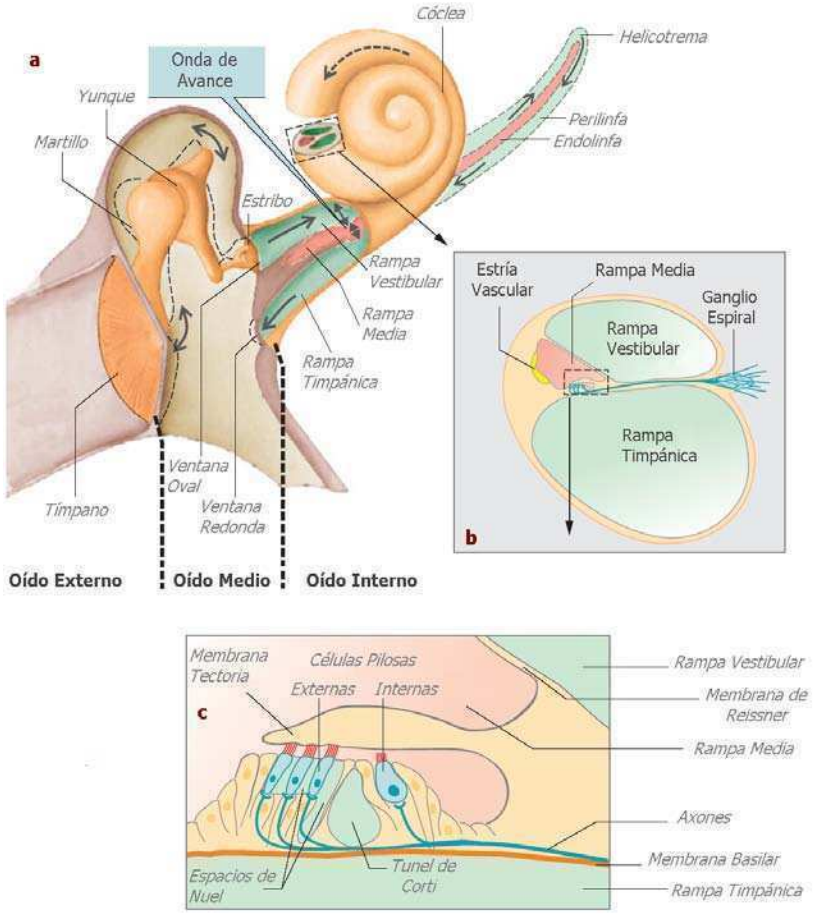

Figura 4: Estructura del oído humano: **a.** mecanismo de recepción de estímulos sonoros, **b.** corte transversal de la cóclea, **c.** órgano de Corti. Adaptado de [159]

En el oído medio el mecanismo compuesto por tres pequeños huecesillos: el martillo, el yunque y el estribo realizan la transmisión mecánica del sonido y actúan como un adaptador de impedancias. La impedancia es la resistencia que presenta un medio al paso de las ondas sonoras. Cuando el sonido se propaga de un medio de baja impedancia como el aire a uno de alta impedancia como un medio líquido,

gran parte de la energía acústica se pierde por reflexión en la interfase aire-líquido. Para evitar esa pérdida de energía acústica el mecanismo conformado por los huecesillos y el tímpano intensifican alrededor de 20 veces la presión de las ondas que llegan a la membrana timpánica a partir de la relación de superficies tímpano/ventana oval de 17:1; y el efecto de palanca de los huecesillos, factor 1:3.

Esa conversión de impedancias es eficaz para frecuencias inferiores a 2400 Hz.

En este mismo sistema de transmisión en el oído medio, la acción de dos músculos filtran la señal de entrada, reduciendo la transmisión de sonidos de baja frecuencia, y son los responsables del mantenimiento reflejo de la intensidad de los sonidos, de la protección ante sonidos fuertes y la reducción de ruidos desagradables.

El oído interno está formado por el órgano del equilibrio y por la cóclea, un conducto de $3-4$ cm de longitud con forma de caracol. En la cóclea hay un tubo, llamado rampa media que contiene un líquido denominado endolinfa, y que transcurre paralela a otras dos cavidades: la rampa vestibular y la rampa timpánica, hasta llegar al vértice de la cóclea. Estos dos conductos, al igual que la rampa media, están llenos de líquido, en este caso de perilinfa. La rampa vestibular comienza en la ventana oval y la rampa timpánica termina en la membrana de la ventana redonda.

Las estructuras sensoriales del órgano auditivo están constituidas por 10000 a 12000 células ciliadas externas y 3500 células ciliadas internas ubicadas en la membrana basilar. Las células pilosas externas están dispuestas en tres hileras, con sus cuerpos celulares rodeados de perilinfa del espacio de Nuel y presentan alrededor de un centenar de cilios (microvellosidades) relacionados en su base a través de las células de sostén con la membrana tectoria. Estas células están innervadas por la rama eferente del ganglio espiral. Las células pilosas internas están rodeadas por células de sostén y sus cilios, a diferencia de los de las células ciliadas externas, se sitúan libres en la endolinfa, en una sola hilera, y tienen contacto sináptico con las fibras del ganglio espiral. Los axones eferentes del núcleo olivar superior lateral se unen a las terminaciones eferentes.

La transmisión del sonido en el oído interno se produce de la siguiente manera: los movimientos del estribo provocan vibraciones en la membrana de la ventana oval, que son transmitidas a través de la perilinfa hasta la ventana redonda (figura 4 **a**). Las paredes de la rampa coclear, es decir (membranas de Reissner y basilar) ceden ante las ondas de presión y vibran alternando de manera sucesiva contra la rampa vestibular y la rampa timpánica, como se muestra en las figuras 5 **a** y 5 **b**.

Las propiedades mecánicas de la membrana basilar van cambiando a lo largo de su eje longitudinal, siendo más rígida y delgada en su base (cerca de la ventana oval). Ello hace que esa región resuene con las altas frecuencias. En cambio, las vibraciones de baja frecuencia, generan una mayor amplitud de desplazamiento de la membrana basilar cerca del ápex coclear. Esta representación de las frecuencias de los

sonidos en diferentes ubicaciones de la membrana basilar se conoce como organización tonotópica.

Las vibraciones de la rampa coclear provocan pequeños desplazamientos relativos entre la membrana tectoria y basilar. Estos movimientos generan una curvatura de las cilias de las células ciliadas externas, que conducen a cambios en la conductividad de la membrana de estas células (transducción mecanosensible). Si el desplazamiento es en un sentido dado se produce la despolarización celular, que causa un acortamiento de las células ciliadas externas en sincronía con el estímulo. Cuando el movimiento de corte es en la dirección opuesta se produce una hiperpolarización celular y la extensión de las células ciliadas externas figura 5 **c**.

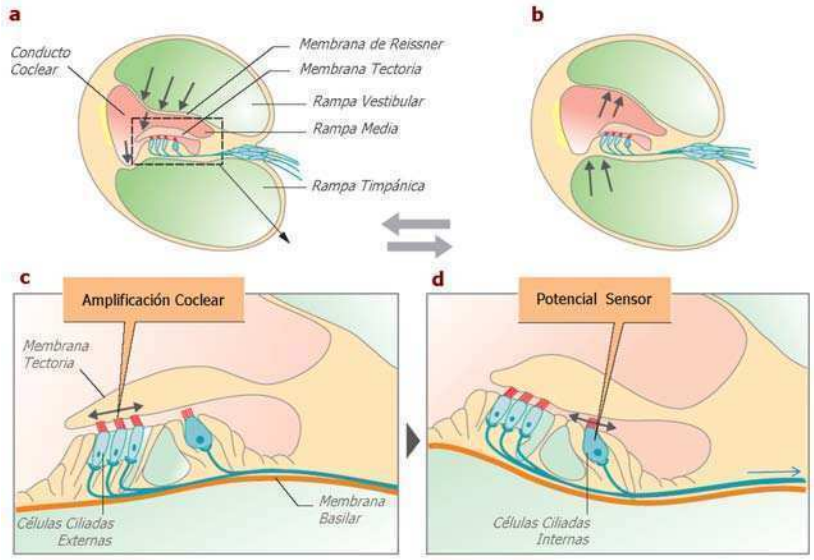

Figura 5: Estimulación de células ciliadas por deformaciones mecánicas: **a.** y **b.** movimiento alternado de las paredes de la rampa coclear generado por las ondas de presión sonoras; **c.** mecanismo de amplificación coclear; **d.** potencial sensor. Adaptado de [159]

Esa electromotilidad de las células pilosas externas contribuyen a la amplificación coclear (se amplifica 100 veces el estímulo), y se lleva a cabo antes que el estímulo alcance los verdaderos sensores acústicos (las células cicliadas internas).

La electromotilidad genera ondas en la endolinfa que se encuentra debajo de la membrana tectoria las cuales ejercen fuerzas de corte en las cilias de las células pilosas internas, causando la apertura de los canales de transducción y la despolarización de las células (potencial sensor), figura 5 **d**. Esto lleva a la liberación de neurotransmisores por parte de las células ciliadas internas y la conducción del impulso nervioso hacia el sistema nervioso central.

El sonido descompuesto a nivel de la cóclea se transmite por fibras separadas de las vías auditivas hasta el nivel central. La vía auditiva es una cadena interconectada de núcleos en el tronco cerebral que van desde la cóclea hasta el córtex auditivo en los lóbulos temporales del

cerebro (figura 6). Esta vía se interrumpe haciendo sinapsis con distintos núcleos, varias veces a lo largo de su trayecto, y en cada uno de ellos se analizan aspectos diferentes del sonido. Al parecer también se producen inhibiciones laterales en cada nivel.

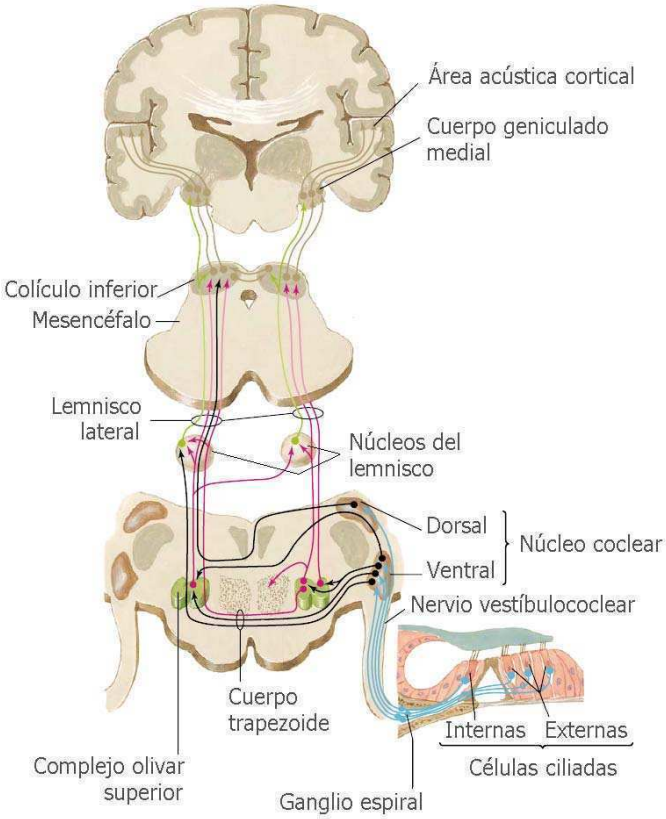

Figura 6: Diagrama esquemático de la vía auditiva. Adaptado de [70]

El cuerpo (soma) de las primeras neuronas de esta vía se ubican en el ganglio espiral de la cóclea. Estas primeras neuronas son células monopolares y existen unas 30000 en cada ganglio espiral. Sus prolongaciones periféricas hacen sinapsis con las células ciliadas (en un 90 % con células ciliadas internas). Una neurona está conectada a una sola célula ciliada interna, pero cada célula ciliada interna está conectada con aproximadamente 10 de esas neuronas. Este patrón de innervación tiene importancia funcional, indica por un lado que casi toda la información auditiva proviene de las células ciliadas internas, y además que varias fibras pueden codificar de manera independiente (y algo diferente) la información procedente de una sola célula ciliada interna. La inervación de las células ciliadas externas es diferente, ya que cada neurona ganglionar puede innervar varias células ciliadas externas.

Al igual de lo que sucede en la membrana basilar, cada fibra responde mejor a la estimulación de una frecuencia determinada (su frecuencia característica). Si bien estímulos de frecuencias diferentes a la característica también pueden llegar a excitar estas fibras, ello ocurre solamente a intensidades mayores. Las células ciliadas también reciben

fibras eferentes, cuyo número es mucho mayor para las células ciliadas externas.

Las prolongaciones centrales de las neuronas del ganglio espiral, que forman parte del VIII nervio craneal, penetran en la cavidad craneal atravesando el conducto auditivo interno para posteriormente introducirse en el tronco del encéfalo, bifurcarse y hacer sinapsis con las segundas neuronas situadas en los núcleos cocleares dorsal y ventral. Los cilindroejes que salen de la segunda neuronas forman tres tractos:

- Estría o tracto auditivo dorsal: las fibras provienen del núcleo coclear dorsal y terminan en las neuronas de los núcleos contralaterales del lemnisco lateral.
- Estría auditiva intermedia: las fibras provienen de ambos núcleos cocleares y terminan en el núcleo olivar superior de ambos lados.
- Cuerpo trapezoide: se origina en el núcleo coclear ventral y termina en los núcleos del cuerpo trapezoide y en el núcleo olivar superior.

Los tres tractos contienen fibras que siguen un camino ascendente sin hacer sinapsis en los núcleos intercalados. Dichas fibras ascendentes tanto cruzadas como directas, junto con las fibras que nacen en los núcleos de la vía (núcleo del cuerpo trapezoide, olivar superior y lemnisco lateral), alcanzan el colículo inferior, formando un tracto mielinizado y organizado tonotópicamente, llamado lemnisco lateral. Adosado a éste se encuentran los núcleos del lemnisco lateral. Este tracto también presenta fibras descendentes.

El colículo inferior tiene en su centro un núcleo grande donde terminan tonotópicamente las fibras del lemnisco lateral. Además posee una corteza dorsal formada por cuatro capas neuronales que reciben aferencias descendentes de la corteza cerebral auditiva y, que a su vez se proyectan sobre la corteza del colículo inferior contralateral. Los cilindroejes originados en el núcleo del colículo inferior terminan en el núcleo geniculado medial. Los cilindroejes originados en este núcleo avanzan por el segmento sublenticular, formando las radiaciones acústicas, para terminar finalmente en la corteza auditiva.

La corteza auditiva se encuentra en el lóbulo temporal, en el labio inferior de la cisura lateral y está compuesta por el área auditiva primaria y debajo de ésta, el área auditiva secundaria. La corteza auditiva de cada hemisferio recibe información bilateral, de modo que incluso la destrucción completa de la corteza auditiva de un hemisferio apenas produce efectos. Sin embargo, la percepción de la música está lateralizada generalmente hacia el hemisferio derecho. El área auditiva primaria tiene una representación tonotópica. Además la corteza está dividida en bandas de dos tipos que se disponen alternativamente. Una de estas bandas se conoce como columnas de supresión y sus neuronas responden cuando las señales provienen de un oído. La otra banda se conoce como columnas de sumación y sus neuronas responden cuando reciben señales provenientes de los dos oídos.

Esas fibras nerviosas de la vía auditiva codifican distintos tipos de información del sonido como su frecuencia, intensidad, dirección y distancia desde la fuente.

Existe un mecanismo muy fino de procesamiento de las señales acústicas que permite discriminar pequeñas variaciones en las frecuencias de los sonidos. De hecho el umbral diferencial de frecuencias relativo es de aproximadamente 0,003, lo que implica que podemos distinguir un tono de 1003 Hz de otro de 1000 Hz. A esta capacidad de diferenciación contribuyen la representación tonotópica de la cóclea y la amplificación que realizan las células pilosas externas, así como el contraste neuronal a lo largo de la vía auditiva.

Para las intensidades, el oído presenta una menor resolución que para las frecuencias. El umbral diferencial relativo es de 0,1, es decir, un sonido se percibirá más alto o más bajo, cuando su intensidad varíe más de un factor relativo de 1,1. Las diferencias de intensidades se codifican empleando tanto la frecuencia de descargas de las fibras como el número de fibras que reclutadas. Un sonido de mayor intensidad produce potenciales de acción más frecuentes en la fibra nerviosa eferente y provocan que se recluten más fibras nerviosas vecinas.

La localización del sonido es uno de los aspectos más estudiados. Se cree que para esta tarea el sistema nervioso central emplea dos estrategias de análisis diferentes, en función del sonido recibido. Para frecuencias entre 200 y 2000 Hz, utiliza fundamentalmente el retraso relativo en la llegada de los sonidos entre ambos sonidos (retraso interauricular), análisis efectuado por las neuronas del núcleo olivar superior.

Para frecuencias entre 2 y 20 KHz emplea la diferencia de intensidad interauricular, lo que es analizado por algunas neuronas del núcleo olivar y de los núcleos del cuerpo trapezoide. Estos efectos también se suman. Por otro lado el oído externo ayuda a diferenciar si el sonido viene de adelante o atrás, abajo o arriba. La audición bilateral posibilita también que, en situaciones de mucho ruido ambiental (en una fiesta, por ejemplo), se perciba mejor el sonido de una voz, en comparación con la audición monoaural. Estos son aspectos que contribuyen a robustecer el reconocimiento del habla. El colículo inferior es un centro integrador de la información sobre la localización del sonido y existen en él una representación topográfica del espacio auditivo.

La distancia de una fuente de sonido se reconoce porque en la transmisión del sonido las frecuencias altas se reducen más que las bajas. Cuanto más largo es el camino que recorre el sonido, menos frecuencias altas llegan (p. ej., los truenos en una tormenta lejana o cercana).

El procesamiento de la secuencia temporal es muy importante para recibir información contenida en algunos sonidos como la música y las palabras. En este proceso están involucrados varios núcleos. Algunas neuronas de los núcleos del lemnisco lateral señalan el inicio del sonido, cualquiera sea su intensidad o frecuencia; otras neuronas de estos núcleos calculan otros aspectos temporales de los sonidos como las duraciones. El colículo inferior también interviene en el procesamiento de sonidos con patrones temporales complejos.

1.2.4 Psicofísica de la Audición

La psicofísica es el estudio de las relaciones entre las propiedades físicas de los estímulos y las respuestas sensoriales que ante ellos presenta el sistema nervioso humano. En esta definición queda claro que existe una distinción entre el estímulo físico y la respuesta psicológica ante él. Antes de hablar sobre los aspectos psicológicos de la percepción es necesario introducir algunos conceptos físicos básicos. La física del sonido se estudia en la rama conocida como acústica.

Las ondas de sonido se propagan por los gases (como el aire), los líquidos (como en la perilinfa), y los sólidos (como en el cráneo). Durante el habla, las fluctuaciones de presión que se generan en la fuente de sonido se propagan en el aire a una velocidad característica (c), que depende del medio. Por ejemplo en el aire a 0 ºC la velocidad de propagación de las ondas acústicas es de 332 m/s. Si se hace una representación gráfica de la variación de presión de una onda sonora en función del tiempo, se obtiene una onda como la mostrada en la figura 7 **a**.

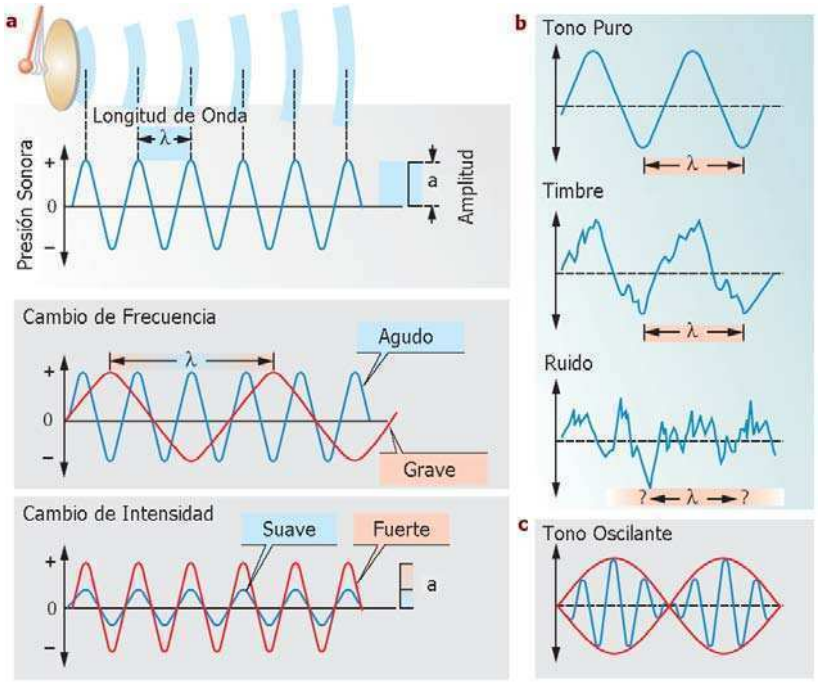

Figura 7: Ondas acústicas. **a:** longitud de onda, amplitud y frecuencia. **b:** Representación de un tono puro (*arriba*), un tono compuesto (*centro*) y ruido (*abajo*). **c:** Ilusión auditiva de una oscilación de baja frecuencia tras la combinación de dos tonos de similar amplitud y frecuencias ligeramente diferentes. Adaptado de [159]

longitud de onda
amplitud
frecuencia

La menor distancia entre dos puntos contiguos que presenten la misma presión se designa como longitud de onda (λ), y la máxima diferencia de presión respecto de la de reposo se denomina amplitud (*a*). Según aumente o disminuya λ, se oirá un tono más grave o más agudo. Dicha altura tonal viene determinada generalmente por su fre-

cuencia (f), que indica la tasa con la que oscila la presión del sonido en un punto del campo de audición. La unidad de frecuencia es el *Herz* ($Hz = s^{-1}$). Una persona joven y sana es capaz de oír sonidos cuyas frecuencias se extiendan en un rango entre 20 Hz hasta $16000 - 20000$ Hz.

Por otra parte, una disminución (o un aumento) de a, provoca un tono más fuerte o más suave.

Frecuencia, longitud de onda y velocidad del sonido se relacionan entre sí mediante la siguiente ecuación:

$$\lambda\,(m) = \frac{c\,(m/s)}{f\,(Hz)} \tag{1.1}$$

Un tono en sentido estricto corresponde a una vibración pura de forma senoidal. El "tono" o timbre de fuentes sonoras como el habla, los instrumentos musicales o el canto se compone de un conjunto de tonos puros de diferentes frecuencias y amplitudes, lo que conforma la onda que vemos en la figura 7 **b**. El tono puro más bajo contenido en el timbre determina la altura del "tonal" del mismo.

timbre

En la figura 7 **c** se puede ver el efecto auditivo que genera la superposición de dos tonos puros de amplitud similar pero frecuencias ligeramente diferentes, por ejemplo $f_1 = 1000Hz$ y $f_2 = 1{,}003Hz$. La pequeña diferencia de frecuencias provoca que estén en fase y fuera de fase de manera cíclica con una frecuencia igual a la diferencia de frecuencias $f_2 - f_1$. Por lo tanto estos tonos se oirán como un tono de 1000 Hz cuya amplitud aumenta y disminuye de manera cíclica a una tasa de 3 veces por segundo.

Potencia: mide la velocidad con la que se realiza un trabajo, es decir, cantidad de trabajo por unidad de tiempo y su unidad es el vatio (W), que equivale al trabajo de 1 Joule desarrollado en un segundo.

Intensidad (I): es una medida física que cuantifica la cantidad de potencia aplicada por unidad de área. Su unidad es W/m^2.

Presión: es una medida de fuerza por unidad de área, y su unidad en el sistema MKS es de Newtons por metro cuadrado (N/m^2), o Pascal (Pa). El rango de magnitudes de sonidos que podemos oír es enorme. El mayor nivel de presión sonora que podemos tolerar es alrededor de 10 millones de veces más grande el menor sonido perceptible. Como resulta inconveniente trabajar con ese rango de magnitudes en escala lineal, se lo transforma en una escala logarítmica cuya unidad es el decibel (dB). Para definirlo se utiliza como valor de referencia la intensidad audible ($I_0 = 10^{-12}\ W/m^2$) o mínima presión de sonido que se puede percibir ($p_0 = 2 \cdot 10^{-5}$ Pa). La presión expresada en dB se indica como dB SPL, e indica decibeles de nivel de presión sonora. De manera similar, la intensidad expresada en dB se nombra como dB IL (decibeles de nivel de intensidad).

dB de presión auditiva sonora

La fórmula general para expresar una presión p_x Pa en decibeles es:

$$dB = 10 \cdot \log\left(\frac{p_x}{p_0}\right) \tag{1.2}$$

La fórmula correspondiente para expresar una intensidad I_x expresada en W/m^2, en decibeles es:

$$dB = 10 \cdot \log\left(\frac{I_x}{I_0}\right) \tag{1.3}$$

psicoacústica

La rama de la psicofísica dedicada a estudiar la audición se denomina psicoacústica . Si esa correspondencia fuera uno a uno, entonces se podría cuantificar lo oído directamente de los atributos físicos del sonido. Esto implicaría que todo sonido que existiera sería audible, que cualquier cambio en un sonido podría ser discriminable y que cualquier cambio en la magnitud del estímulo resultaría en un cambio perceptual de la misma magnitud. Como esto claramente no es así, surge la necesidad de establecer esa correspondencia y es la razón de ser de la psicoacústica.

A continuación se presentarán algunas propiedades psicoacústicas que resultarán útiles para comprender el desarrollo de la tesis.

- **Bandas Críticas:** este concepto surge de experimentos de enmascaramiento simultáneo, en los que la presencia de un sonido (por ejemplo de ruido blanco) obscurece o incluso impide la percepción de otro sonido (por ejemplo de un tono puro). Se encontró que solamente una banda acotada de frecuencias en torno al tono puro (banda crítica) contribuye a su enmascaramiento. Una banda crítica define un rango de frecuencias (ancho de banda) en los experimentos piscoacústicos en el que la respuesta perceptual cambia abruptamente.

  Este efecto de bandas críticas está vinculado con el análisis no lineal de las frecuencias llevadas a cabo por la membrana basilar. Ese análisis se puede pensar como la acción de un banco de filtro pasa-banda, en el que la respuesta en frecuencia de cada filtro es mayor al aumentar la frecuencia central del filtro.

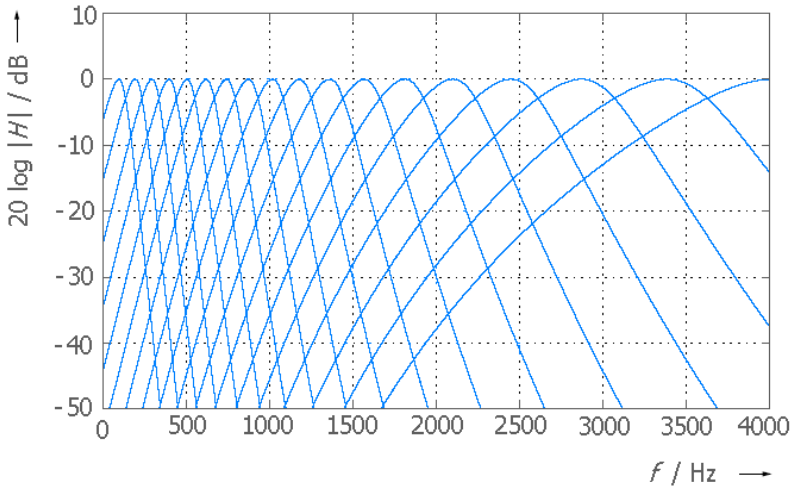

Figura 8: Modelo de banco de filtros de bandas críticas hasta 4 Khz

En la figura 8 se puede ver una representación de dicho banco de filtro. En la realidad los filtros pasa-banda no son ideales, sino que sus respuestas en frecuencia se solapan ya que dos puntos sobre la membrana basilar no pueden vibrar de manera completamente independiente entre sí.

Aún así el concepto de análisis mediante un banco de filtros coclear es muy empleado en el campo de procesamiento del habla, donde se estimaron y definieron bandas críticas que muestran que el ancho de banda efectivo es constante para los filtros cuyas frecuencias centrales se encuentran por debajo de los 500 Hz, y con un ancho de banda relativo del 20 % para aquellos con frecuencias centrales superiores a los 500 Hz.

La siguiente ecuación refleja el comportamiento observado a través de mediciones empíricas sobre el rango auditivo:

$$\Delta fc = 25 + 75 \cdot \left[1 + 1{,}4 \cdot \left(\frac{fc}{1000}\right)^2\right]^{0{,}69} \tag{1.4}$$

donde Δfc es la banda crítica asociada con la frecuencia central fc.

Se pueden encontrar aproximadamente 25 bandas críticas que van de 0 a 20 kHz.

- **Sonoridad:** es una cualidad perceptual vinculada con el nivel de presión de un sonido. Se cuantifica relacionando el nivel de presión sonora de un tono puro (en decibeles relativos al nivel de referencia estándar), con la sonoridad percibida del mismo tono, cuya unidad es el fono. Esa relación se muestra en la figura 9

  Las curvas de sonoridad presentadas indican que la percepción de ese atributo es dependiente de la frecuencia. Específicamente la curva señalada como "umbral de audibilidad" muestra el nivel de presión sonora necesario para que un tono con determinada frecuencia sea apenas audible para una persona normoyente.

  Para que los tonos de bajas frecuencias puedan ser percibidos deben ser significativamente más intensos que los de frecuencia media.

  Las líneas sólidas se denominan contornos de iso-sonoridad, y se miden comparando los sonidos de distintas frecuencias con un tono puro de 1000 Hz y nivel de presión sonora conocido. Por ejemplo el punto sobre la curva de 50 fonos correspondiente a una frecuencia de 100 Hz se obtiene ajustando la potencia del tono de 100 Hz hasta que suene tan fuerte como un tono de 1000 Hz con un nivel de presión de 50 dB.

  Para el caso anterior se observa que un tono de 100 Hz debe tener un nivel de presión de 60 dB para ser percibido con la misma sonoridad que un tono de 1000 Hz de 50 dB. Por convención se

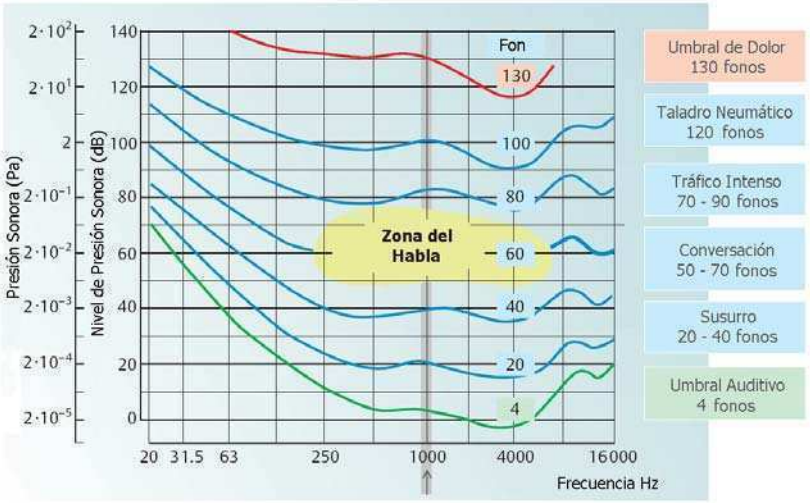

Figura 9: Relación psicoacústicas entre presión sonora, nivel de presión sonora y sonoridad. El umbral acústico depende de la frecuencia (curva inferior). Las curvas intermedias muestran contornos de equi-sonoridad. La curva superior muestra el nivel de presión sonora que exige demasiado al oído y provoca una sensación dolorosa. Adaptado de [159]

dice que tanto el tono de 1000 Hz a 50 dB y el tono de 100 Hz a 60 dB tienen un nivel de sonoridad de 50 fonos.

Las curvas de equi-sonoridad muestran que el sistema auditivo es más sensible a las frecuencias con un rango desde 100 Hz hasta 6 kHz aproximadamente, con un máximo de sensibilidad entre 3 y 4 kHz aproximadamente. Este es precisamente el rango de frecuencias que muestra la mayor parte de sonidos del habla.

- **Altura Tonal o Pitch:** es el correlato psicofísico de la frecuencia. La relación entre el pitch percibido y la frecuencia es aproximadamente lineal entre 20 Hz y 1000 Hz. Por lo tanto la magnitud del cambio percibido en el pitch entre al variar la frecuencia de un tono 200 a 300 Hz, es similar a un cambio de frecuencias de 800 a 900 Hz. Desde 1000 Hz hasta los 20000 Hz la relación entre frecuencia y pitch es logarítmica. En este rango, por ejemplo la variación de frecuencia de 1000 a 1100 Hz (10 %) es percibida con una magnitud equivalente a la que se produce entre 12000 y 13200 Hz (10 %).

  Hay varias escalas propuestas para representar el pitch en función de la frecuencia de los estímulos. Entre ellas las más utilizadas son las escalas Mel, Bark y ERB.

  Como se puede ver en la figura 10, las bajas frecuencias son mejor resueltas que las altas frecuencias. También se observa que las escalas musicales son de tipo logarítmica: aumentando una octava duplica la frecuencia. Por lo que las escalas musicales exhiben el mismo comportamiento que el pitch a partir de los 1000 Hz. Las escalas Mel y Bark se pueden interpretar en término de las

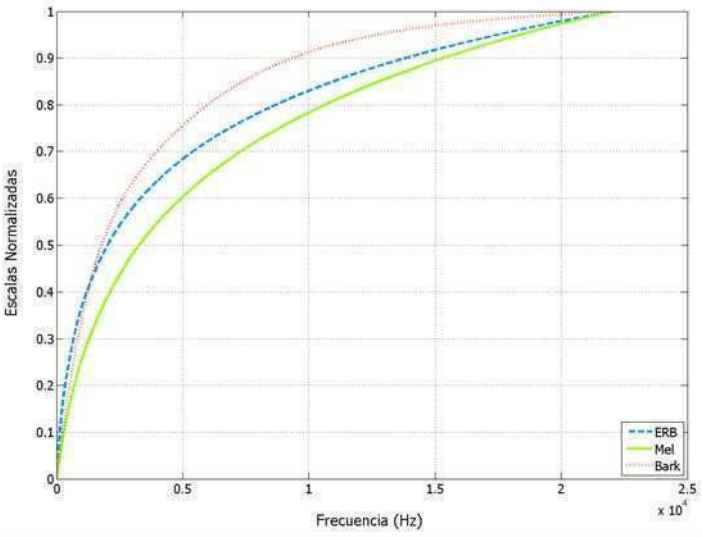

Figura 10: Escalas perceptuales de la altura tonal

distancias en la membrana basilar. La percepción de diferencias de pitch es función de la distancia de resonancia de los sonidos en la membrana basilar.

El pitch es un factor particularmente importante para la percepción de atributos como melodía u armonía en la música, y representa una de las mayores fuentes de información para la prosodia del habla. Al igual que la sonoridad y el timbre, esta medida es subjetiva y no se puede expresar en términos físicos. Para el caso de un tono puro, su principal correlato objetivo es el atributo físico de la frecuencia, pero la intensidad, duración y envolvente temporal también afectan al pitch de un tono.

Un tono complejo compuesto por sinusoides con diferentes frecuencias se puede percibir como un solo tono (como en el caso de un clarinete), como un grupo de tonos (por ejemplo un acorde ejecutado por un conjunto de instrumentos), o también se pueden percibir como si cada un esos componentes sinusoides tuvieran su propio pitch. Incluso se puede evocar sensaciones de este atributo mediante sonidos discretos y no bien definidos.

1.2.5 Teorías de Reconocimiento del Habla

Como se mencionó, los mecanismos por los cuales el oyente obtiene una secuencia de palabras de reconocimiento partiendo de la señal acústica no se conocen en la actualidad. Sin embargo se pueden distinguir algunas teorías que intentan explicar cómo se lleva a cabo el reconocimiento humano del habla (RHH). Entre ellas se pueden citar la **teoría episódica o sub-simbólica** y la **teoría simbólica**. Ambas asumen que hay un proceso de competencia léxica entre todas las hipótesis de palabras que se activan simultáneamente en el tiempo. En

otras palabras, que la elección de una hipótesis léxica específica, para un instante de tiempo dado, está basada en su activación inicial, y en la inhibición causada por las demás hipótesis que fueron activadas.

La **teoría episódica** de RHH asume que cada unidad léxica está asociada con un gran número de representaciones acústicas almacenadas en la memoria [92, 93, 55]. Durante el reconocimiento, la señal de entrada se compara con las representaciones acústicas almacenadas, y mediante un proceso de competencia se decide qué representación léxica es la elegida.

Por su parte la **teoría simbólica** de RHH sostiene que hay una etapa intermedia en la que se mapea la señal acústica de entrada en una representación pre-léxica (por ejemplo mediante fonemas), y que las unidades léxicas se almacenan en la memoria en términos de dicha representación simbólica [114, 130, 54, 109]. Es decir que el proceso de reconocimiento del habla en la teoría simbólica consta de dos niveles: el nivel prelexical y el nivel lexical, donde tiene lugar la competencia léxica. Los distintos modelos de RHH que comulgan con la teoría simbólica suelen diferir acerca de la naturaleza de la representación pre-léxica. Este tema también es importante para el reconocimiento automático del habla, dado que detrás de la representación simbólica intermedia del habla se encuentra el supuesto de las teorías respecto a la información que es relevante en la señal para distinguir palabras. Si esa información no es capturada durante la representación preléxica, se asume como irrelevante para el RHH.

A continuación se hará una breve reseña de los modelos más significativos para el RHH [71]:

- **Modelo Cohorte** [111], hace hincapié en la naturaleza temporal del reconocimiento auditivo de palabras y propone que la percepción del comienzo de una palabra activa todas las palabras conocidas que empiezan igual. Esta *"cohorte"* de candidatos léxicos activados inicialmente se va reduciendo de manera paulatina, mediante la eliminación de los candidatos incompatibles con las nuevas evidencias que se van recabando. Debido a que la representación sonora del fragmento inicial de una palabra puede ser diferente a todas las demás, este modelo predice que es posible el acceso léxico sin necesidad de oír todos los sonidos de la palabra en cuestión. Numerosos estudios confirmaron que esto se verifica en la práctica.

  Entre las mayores contribuciones del modelo Cohorte se encuentran la distinción de tres procesos específicos dentro del RHH: un proceso de activación, un proceso de evaluación y un tercero de selección. La interrelación de esos procesos resulta en un mecanismo de resolución de ambigüedades, mediante el cual el reconocimiento de una palabra procede de manera incremental. Sin embargo, su arquitectura está basada en una identificación adecuada del instante de inicio correspondiente a cada palabra, lo que es incompatible con el hecho que en el habla continua los oyentes generalmente no disponen de una identificación de esos límites entre palabras.

En reformulaciones posteriores de este modelo, su autor propuso algunas modificaciones en su arquitectura: reemplazo de las representaciones pre-léxicas de fonemas por atributos, y un mecanismo que estima el grado de soporte a los candidatos léxicos. Las palabras candidatas se activan en paralelo, cada una según el grado con el cual son soportadas por la información en la señal de habla. Así en cualquier momento dentro del procesamiento de esa señal, se puede considerar que existe un candidato de reconocimiento más fuerte que las demás.

- **Modelo TRACE** [114], fue el primer modelo computacional del RHH, es un modelo conexionista, por lo que consta de múltiples unidades de procesamiento conectadas entre sí. Esas unidades están ordenadas en tres niveles: el nivel de atributos, el fonémico y el de palabras. Los nodos de cada nivel se corresponden con hipótesis perceptuales, y la activación de cada nodo indica el grado de soporte que brindan las evidencias a esas hipótesis. Las unidades mutuamente consistentes de diferentes niveles presentan conexiones excitatorias, mientras que las unidades del mismo nivel están unidas por conexiones inhibitorias. Además todas las conexiones entre niveles son bidireccionales, por lo que la información puede fluir en ambas direcciones (puede darse un procesamiento *bottom-up* como *top-down*).

  Durante el reconocimiento, para cada intervalo de tiempo, se presentan a los nodos de entrada los atributos de la señal de habla, lo que activa algunos nodos de atributos, que a su vez activan a los nodos fonémicos y éstos a los de palabra. Es decir, la activación en la red se dispersa ya que los nodos activos excitan a sus vecinos, y como por otro lado, debido a que los nodos de un mismo nivel de procesamiento son hipótesis alternativas, éstos compiten entre sí, resultando eventualmente una sola unidad de salida activa.

  Este modelo mejoró a la versión original del modelo Cohorte al menos en dos aspectos: en primer lugar a diferencia del modelo Cohorte, TRACE no da prioridad absoluta a la información vinculada con el inicio de las palabras, sino que lo hace de manera gradual. TRACE tiene en cuenta los efectos de la posición en el reconocimiento de palabras dando más importancia para a la parte inicial de la palabra que a la final, pero de una manera tal que permite que las palabras candidatas pueden activarse en cualquier lugar de la señal de habla, y además permite reconocer palabras mal pronunciadas. En segundo lugar, en TRACE todas las palabras candidatas compiten entre sí, y el grado de activación de un candidato está influenciado por el grado de activación de los candidatos competidores. Ello provoca un incremento de las diferencias de activación iniciales de los candidatos que surge como función de la bondad de su ajuste inicial a los atributos de la señal de habla. La competencia léxica asegura que se encuentre la mejor opción, y si bien los límites entre palabras no se identifican explícitamente, dado que las palabras competidoras activadas no tienen por qué estar en correspondencia mutua, los

límites entre palabras emergen como resultado del proceso de competencia.

- **Modelo Shortlist** [130], es un modelo basado en redes conexionistas al igual que el modelo TRACE. También al igual que éste se propone que la señal percibida activa una serie de candidatos léxicos potenciales que compiten unos con otros de manera activa en un proceso de activación en el que, cuanto más activo sea un candidato determinado, más inhibirá la activación de sus competidores. Sin embargo los modelos TRACE y Shortlist se diferencian, básicamente, en el hecho de admitir o no la existencia de un flujo de información unidireccional o bidireccional entre los niveles del procesamiento.

  En el modelo Shortlist, la información puede fluir desde el nivel preléxico de procesamiento de la señal hasta el léxico, pero no al revés. Además este modelo presenta una arquitectura de dos estadíos, en las que los candidatos léxicos iniciales se generan exclusivamente a partir de información ascendente, y el proceso de competencia tiene lugar solamente entre los miembros de esta "lista de candidatos escogidos". Por su parte en el modelo TRACE, esa competencia se puede dar, en principio, entre todos los elementos del lexicón, lo cual hace que sea menos manejable computacionalmente, mientras que la estructura del modelo Shortlist tiene la ventaja práctica de permitir simulaciones con un vocabulario realista de varias decenas de miles de palabras.

Como se mencionó, el modelo TRACE contempla un procesamiento de información bottom-up y top-down, mientras que los modelos Cohorte y Shortlist solo consideran un flujo de información de abajo hacia arriba (del nivel preléxico al léxico). Sin embargo, la forma en que se procesa el lenguaje en los humanos sigue siendo un campo de debate en la psicolingüística, en el que confrontan dos hipótesis: bottom-up o top-down.

El procesamiento bottom-up implica que el cerebro del oyente distingue en primer lugar los fonemas, los procesa y determina a partir de allí qué palabras se pronunciaron. Por su parte, en el procesamiento top-down sucede lo opuesto: el oyente parte de lo que espera oír y basándose en el contexto llena los "huecos" necesarios para completar la decodificación del mensaje. El contexto, la semántica y pragmática ejercen su influencia sobre la percepción aún antes que los propios fonemas.
Se pueden encontrar evidencias a favor de ambas teorías.

Finalmente se puede encontrar una tercera posición, que es la adoptada por el modelo TRACE, donde se admite un procesamiento concomitante bottom-up y top-down, asumiendo que los oyentes son capaces de recibir las pistas acústicas disponibles y reconocer las palabras en la medida que esas palabras sean coherentes con el contexto [31].

Pruebas experimentales han confirmado que el reconocimiento auditivo de palabras es extremadamente rápido y muy eficiente [110].

Una amplia gama de descubrimientos experimentales también avalan la existencia de candidatos léxicos que se activan de manera simultánea, y compiten activamente entre sí de modo que tal activación puede producir inhibiciones entre ellas [116].

Sin embargo, hay muchas cuestiones sobre el RHH que no se pudieron verificar en la práctica, y permanecen abiertos.

Entre ellas: la naturaleza de la unidad principal de representación preléxica primaria, la contribución relativa de información fonética coherente y no coherente en la activación de palabras; el grado de explicitud fonológica de las representaciones léxicas, el procesamiento de cambios fonológicos inducidos por el contexto, el papel que desempeña la estructura prosódica en el reconocimiento de palabras y el papel de la estructura morfológica interna en el reconocimiento.

Estos conocimientos del campo de psicolingüística son indispensables para la optimización de los sistemas de reconocimiento automáticos.

1.3 RECONOCIMIENTO AUTOMÁTICO DEL HABLA

Se puede definir al reconocimiento automático del habla como el proceso por el cual se convierten señales de habla en secuencias de palabras, empleando algoritmos implementados como programas computacionales.

Este campo de investigación tiene como último fin construir sistemas que puedan reconocer eficientemente el habla, y comprender su significado para cualquier área de discurso, locutor y medio acústico.

Como área de estudio multidisciplinaria, el RAH comprende disciplinas como Física Acústica, Procesamiento de Señales, Reconocimiento de Patrones, Teoría de la Información y las Comunicaciones, Lingüística, Fisiología, Ciencias de la Computación y Psicología, entre otras.

Como ya se ha mencionado, la tarea de reconocimiento del habla mediante computadoras presenta una gran complejidad. Uno de los factores claves que así lo determinan es la variabilidad de la señal de habla [15].

El habla es una señal no estacionaria, dinámica por naturaleza. Cada sonido del habla surge de la combinación de acciones musculares y mecanismos de control, que aún en la articulación cuidadosa y de manera intencional difícilmente genere segmentos idénticos. Es decir, que la primer fuente de variabilidad de la señal de habla se debe a las diferencias de pronunciación. Se puede clasificar a dichas diferencias de pronunciación en :

- ***Diferencias de pronunciación inter-locutores:*** las señales de habla no llevan solamente información lingüística (es decir el contenido del mensaje transmitido), sino mucha información sobre el locutor: características del tracto vocal, género, edad, condiciones socio-culturales, origen geográfico, malformaciones o peculiaridades anatómicas, calidad vocal, etc. Cada uno de estos

aspectos también contribuyen en gran medida a diferencias de características de la señal de voz entre diversos locutores.

- ***Diferencias de pronunciación intra-locutores:*** Hay múltiples factores que influencian el grado de variación en la pronunciación de un mismo locutor. Ellos incluyen:
  - *Estilo de Habla o variación estilística:* este tipo de variación depende si el tipo de habla es espontáneo (como en el habla conversacional), está guionado, o es habla leída [191]. En estilos de habla más informales pueden surgir variaciones por las diferencias de pronunciación de las palabras, provocando muchas veces articulaciones reducidas, algunas inconcientes y vinculadas con el lenguaje, otras dialectales, conocidas como acentos, o de grupos sociales: las variaciones sociolécticas.
  - *Velocidad de Elocución:* se ha demostrado que la velocidad de habla tiene un efecto significativo sobre la pronunción resultante [60]. La velocidad de elocución impacta sobre las características tanto temporales como espectrales de la señal.
  - *Coarticulación:* la superposición de las articulaciones adyacentes afecta la forma en que las palabras se pronuncian [99]. Dado que el habla surge a partir de gestos articulatorios continuos, la posición anterior del sistema de producción de los sonidos, así como la posición siguiente que deberá adoptar para generar habla fluida, determinan que la morfología de la señal resultante varíe dependiendo del contexto en el que se produce. Como puede haber distintos grados de coarticulación, las características de la señal resultante también varía de forma gradual.
  - *Rasgos Suprasegmentales:* la pronunciación de las palabras se ve afectada también dependiendo de la mayor o menor intensidad de las palabras o frases, de la entonación, la frecuencia de ocurrencia de una palabra, su posición en la oración, la posición de una consonante o vocal en la sílaba y el contexto de ésta en la frase [99, 59].
  - *Estado de salud del locutor:* diversos factores como si el locutor está engripado, está cansado, alcoholizado o drogado también influencian la forma de pronunciación de las palabras.
  - *Estado emocional del locutor:* existen características en las señales de habla que se alteran con el estado de ánimo de los locutores: si está contento, aburrido, triste, o excitado por ejemplo. También hay alteraciones que dependen de la actitud del locutor frente al tema sobre el que está hablando [137].
  - *Condiciones externas:* por ejemplo el ruido ambiental causa que el locutor modifique la forma en que habla, lo que se conoce como efecto Lombard [89].
  - *El interlocutor:* la gente adapta su forma de hablar dependiendo a quién lo hace: a un niño, a otro adulto, a un sis-

tema de diálogo, etc. También se producen modificaciones de la pronunciación por empatía, adaptando la forma de hablar a la del interlocutor.

Además de estas variaciones de pronunciación también existen factores externos que se suman a la señal de voz y modifican la onda resultante agregando variaciones de diversa índole. El medio acústico donde se registra el habla es muchas veces un factor clave en la variación de rasgos del habla: ruidos ambientes, reverberaciones e interferencias solamente son algunos de los elementos que provocan tales variaciones.

También se pueden agregar a esta lista factores como calidad de los micrófonos, atributos de los canales de transmisión, placas de adquisición de señales, que influencian los rasgos de la señal resultante.

Aparte del problema de variabilidad de la señal de habla, otro factor que abona la complejidad del RAH está vinculado con la segmentación de las palabras. A diferencia de lo que ocurre cuando leemos, cuando escuchamos no tenemos silencios, que al igual que los espacios en blanco en la lectura nos indiquen cuándo termina una palabra y comienza la siguiente. Los sistemas de reconocimiento deben encargarse de separar las secuencias de sonidos además de clasificarlas.

A continuación se expondrá brevemente la evolución histórica de los sistemas de RAH, sus principales paradigmas metodológicos y se detallará el estado del arte actual en el tema.

1.3.1 Antecedentes y evolución histórica del RAH

Desde sus albores, la investigación en RAH estuvo motivada por la visión de contar con máquinas que puedan comunicarse con los seres humanos de manera natural. Se pueden encontrar algunos hitos en la historia del desarrollo tecnológico del RAH:

- ***Década de* 1920*: reconocimiento de una sola palabra.***

  Probablemente la primera máquina en *"reconocer"* el habla fue un juguete comercial, un perro llamado *"Radio Rex"* comercializado en esta década. El juguete podía moverse gracias a un sistema de resortes que se activaba cuando recibía suficiente energía acústica del rango de 500 Hz, frecuencia que corresponde aproximadamente a la primera formante de la vocal *"e"* presente en el nombre del perro. Por lo tanto este parecía responder ante la pronunciación de su nombre. Sin embargo era incapaz de rechazar muchos otros sonidos conteniendo suficiente energía en la banda de frecuencia citada.

- ***Década de* 1950*: reconocimiento de dígitos aislados (*10 *palabras) dependiente del locutor.***

  Si bien se realizaron muchos trabajos vinculados con el análisis del habla durante las décadas de 1930 y 1940, el primer sistema de reconocimiento de palabras puede considerarse el presentado

por Davis en 1952, que para los laboratorios Bell desarrolló un reconocedor de dígitos empleando las frecuencias de los formantes de las vocales que contenían.

- ***Década de** 1960**: reconocimiento de palabras aisladas (**100 **palabras), múltiples locutores.***

  En la década de 1960 los reconocedores de dígitos llegaron a obtener buena precisión para múltiples locutores. Empleando algunas propiedades acústico-fonéticas del habla, muchos reconocedores fueron capaces de reconocer pequeños vocabularios de palabras aisladas. Durante la década de 1960 también se puede destacar la incursión de varios institutos y laboratorios japoneses en la investigación en RAH, desde ese entonces se mantienen en la vanguardia principalmente en cuanto a la transferencia desde el ámbito académico hacia el comercial de estas tecnologías.

- ***Década de** 1970**: reconocimiento de habla continua, vocabularios medianos (**100 − 1000 **palabras) y múltiples locutores.***

  Durante esta década muchos sistemas consiguieron reconocer vocabularios medianos para múltiples locutores. Tal vez el mayor logro de esta época haya sido proponer los primeros sistemas para el reconocimiento de habla continua. La tendencia en este lapso estaba dada por sistemas basados en conocimiento humano, codificado generalmente en forma de reglas. Hacia fines de los 60 y principios de los 70 se producen algunas modificaciones importantes de paradigmas en el área. El primero de ellos está vinculado a la propuesta de algoritmos de extracción de características a partir de la señal acústica. Por un lado el modelado de rasgos espectrales, incentivados por el surgimiento de la transformada rápida de Fourier ([27]), la aplicación de procesamiento cepstral ([132]), y el desarrollo de LPC ([6]). Por otro lado se proponen técnicas de *warping* (expansión y/o compresión temporal de señales) que permiten manejar y compensar diferencias entre las señales de entrada y los patrones almacenados debidas por ejemplo a diferentes velocidades de habla o longitud de segmentos. El algoritmo natural para resolver este problema es el de programación dinámica, que desde 1968 se reinventa varias veces. Finalmente la tercera innovación importante de este período es el surgimiento de los modelos ocultos de Markov (HMM)[1]. A mediados de los 70 se comienza a aplicar HMM en sistemas de reconocimiento del habla, en 1975, Jelinek para IBM y Baker en Dragon Systems, de manera independiente emplean HMM para el modelado de la señal de habla bajo un encuadre Bayesiano. Ambos sistemas son muy parecidos, sin embargo en la decodificación Baker emplea el algoritmo de Viterbi (empleado en los sistemas actuales de reconocimiento).

modelos ocultos de Markov

- ***Década de** 1980**: reconocimiento de habla continua, vocabularios grandes (**1000 − 10000 **palabras) y múltiples locutores.***

1 En esta tesis se optó por utilizar HMM para abreviar *modelos ocultos de Markov*. A pesar de ser un acrónimo en Inglés, se privilegió esta forma por su extendido uso en la literatura.

En este período los investigadores comenzaron a tratar el problema de reconocimiento del habla continua, independiente del locutor y gran vocabulario. Se observa un cambio en la metodología pasando de sistemas basados en conocimiento a sistemas estadísticos basados en HMM desarrollados en la década anterior. Uno de los motivos de la amplia difusión de HMM en la comunidad del habla fue la instauración de programas de investigación en el área promovidos por la Agencia de Proyectos de Investigación Avanzada del Departamento de Defensa de Estados Unidos (DARPA) que se extienden hasta la década del 90.

Estos proyectos se caracterizaron por brindar a los investigadores corpus comunes para sus desarrollos, objetivos de medición de resultados preestablecidos, lo que brindó la posibilidad de comparación objetiva entre el desempeño de las distintas metodologías propuestas. Dentro de este enfoque a mediados de los 80 se lanza la tarea de Resource Management, que comprende la trascripción de habla leída (oraciones construidas a partir de un lexicón de 1000 palabras) independiente del locutor. Posteriormente se propone la tarea Wall Street Journal, que también corresponde a habla leída pero en este caso de oraciones extraídas de ese periódico, comenzando con un vocabulario de 5000 palabras pero llegando a sistemas de 60000 palabras.

- ***Década de* 1990*: reconocimiento de habla continua, vocabularios grandes (*1000 − 10000 *palabras), múltiples locutores y condiciones adversas.***

  En la década del 90 los objetivos pasaron del habla leída a tareas con habla más natural, es así como DARPA lanza la tarea de *Broadcast News*, que consiste en transcripciones de noticias radiales incluyendo pasajes complicados como son las entrevistas en la vía pública, y posteriormente los dominios *CALLHOME* y *CALLFRIEND*, que consisten en comunicaciones telefónicas entre amigos. Es decir, que esta época está marcada por el desarrollo de sistemas de reconocimiento para aplicaciones reales sobre tareas limitadas, como aplicaciones para información de vuelos, tareas de dictado, etc; y muestra una tendencia hacia aplicaciones con habla telefónica, que como resultado generó un mayor interés en temas tales como procesamiento del habla en condiciones adversas, y habla espontánea.

- ***Década de* 2000*: reconocimiento de habla continua (conversacional), vocabularios grandes (más de* 20000 *palabras), múltiples locutores.***

  En esta década comienzan a extenderse las aplicaciones comerciales de sistemas de RAH, debido a los niveles de desempeño alcanzados y a la capacidad de cómputo de los sistemas computacionales.

Como se comentó, a mediados de la década de 1980, y por iniciativa de DARPA se iniciaron una serie de programas de investigación en el área del RAH. Como parte de dichos programas se fueron definiendo

tareas de reconocimiento específicas y para cada una se brindaba a cada institución participante el acceso a datos para entrenar y evaluar el desempeño de sus sistemas.

En 1989 se realizó la primera competencia formal (que continúan realizándose actualmente), comparando sistemas de distintas instituciones. Cualquier grupo de investigación que participa en una evaluación de este tipo debe probar su sistema de reconocimiento sobre el mismo conjunto de datos de evaluación, ajustándose a reglas de mediciones predefinidas, y posteriormente enviar los resultados de reconocimiento para que el Instituto Nacional de Estándares y Tecnología (NIST), una agencia del gobierno de los Estados Unidos califique y publique los resultados. La figura 11, presenta un resumen de la evolución y situación actual en el campo de reconocimiento del habla de acuerdo a resultados reportados por el **NIST** en mayo de 2009.

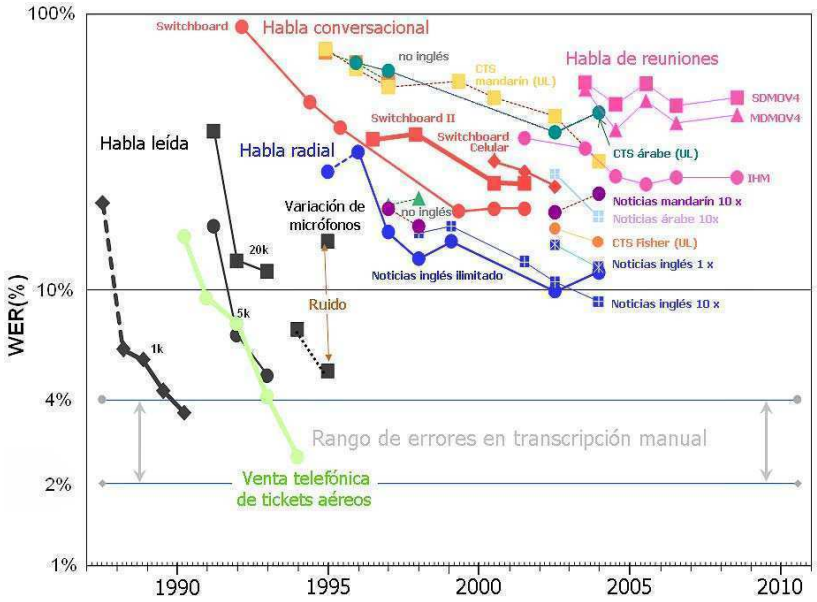

Figura 11: Evolución del desempeño de reconocedores del habla (en términos de WER, graficados en escala logarítmica) para varias tareas según evaluaciones efectuadas por el NIST y publicadas en mayo de 2009

La métrica para cuantificar los desempeños en la figura 11 es la tasa de error por palabras (WER por sus siglas en Inglés), que se define como el porcentaje de palabras reconocidas de manera incorrecta sobre el total de palabras, y se determina empleando un conjunto específico de datos de evaluación.

Analizando los datos presentados se puede ver que el problema de RAH se ha atacado de manera progresiva, comenzando con máquinas capaces de reconocer unos pocos sonidos, y llegando a sistemas más sofisticados capaces de reconocer habla continua y natural. En las primeras etapas de investigación en RAH, se acotaba la cantidad de variabilidad de cada palabra usando habla aislada, lo que por supuesto reduce el grado de interacción entre palabras, y además provoca en los locutores una tendencia hacia una articulación más cuidadosa.

Si bien este tipo de interacciones simplificaba la tarea de los sistemas de reconocimiento, era a expensas de los usuarios: reducción de la velocidad de interacción y necesidad de adaptarse a una forma de habla poco natural. A medida que mejoraba la tecnología se incursionó en modalidades de habla cada vez más naturales: de palabras aisladas al habla conectada y desde esta a oraciones leídas cuidadosamente para finalmente llegar a habla conversacional o espontánea [164].

Durante la historia del desarrollo de sistemas de RAH, que brevemente ilustró esta sección, se pueden encontrar muchas propuestas metodológicas y de algoritmos para resolver el problema de RAH. A continuación se presentarán de manera sucinta los paradigmas principales.

1.3.2 Aproximaciones al RAH

Se pueden distinguir básicamente dos aproximaciones para el desarrollo de sistemas de RAH:

- **Aproximación Basada en Conocimiento:**

  La idea detrás de esta aproximación se centra en la incorporación directa y explícita dentro de los sistema de RAH de conocimiento proveniente de expertos en el campo del habla, codificado en la forma de reglas.

  Este conocimiento está vinculado principalmente con las propiedades acústico-fonéticas de las señales de habla, por lo que también se la conoce como aproximación acústico fonética. La teoría acústico fonética postula que para cada lenguaje existe un conjunto finito de unidades fonéticas distintivas, que se pueden caracterizar mediante propiedades de la forma de onda y/o espectrograma de la señal de habla.

  La secuencia de procesos que involucra esta aproximación se presenta en el diagrama de bloques de la figura 12.

  El primer paso consiste en la conversión de la señal de voz a una representación paramétrica, generalmente mediante índices que describen sus características espectrales en el tiempo.

  El siguiente paso es la etapa de detección de características, que tiene la finalidad de convertir las mediciones espectrales en un conjunto de parámetros intermedios que describen propiedades acústicas de las diferentes unidades fonéticas. Entre esas propiedades se pueden nombrar nasalidad, fricación, localización de formantes, clasificación sonoro/sordo y relaciones entre energías de alta y baja frecuencia. Otras propiedades están vinculadas con movimientos articulatorios, fonética acústica, fonotáctica, etc.

  El tercer paso es la fase de segmentación y etiquetado. Consiste en buscar regiones estables en la señal que permitan obtener fragmentos temporales discretos, en cada uno de los cuales las propiedades acústico-fonéticas sean representativas de una o varias unidades o clases fonéticas, y luego vincular etiquetas fonéticas a cada región segmentada según sus propiedades acústicas.

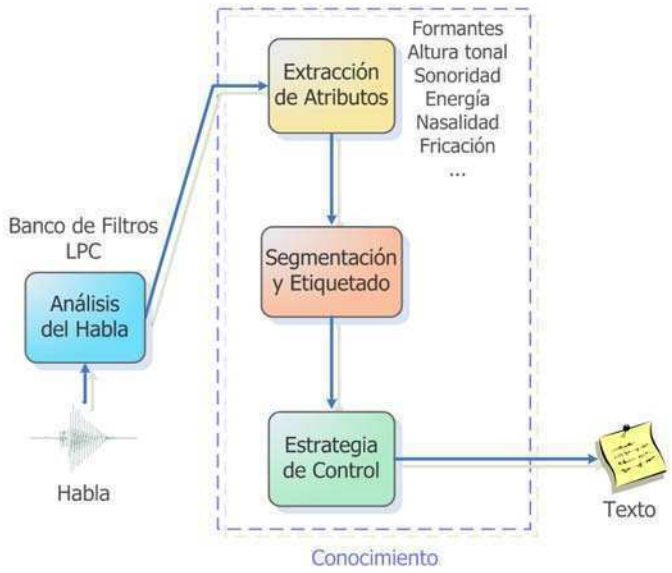

Figura 12: Esquema de sistemas para RAH basado en conocimiento

El último paso consiste en determinar una palabra o una secuencia de palabras válidas a partir de la secuencia de etiquetas fonéticas y de una serie de restricciones impuestas (pertenecer a un vocabulario dado, tener sentido sintáctico y semántico, etc).

Si bien esta aproximación resulta manejable para un conjunto pequeño de palabras aisladas, la tarea pasa a ser intratable para tareas de gran vocabulario [188]. Como resultado, los cálculos requeridos para su funcionamiento son costosos computacionales, mientras que los resultados de reconocimiento obtenidos han sido relativamente pobres. En general, la aproximación basada en conocimiento enfrenta los siguientes inconvenientes:

- Pérdida de generalidad, por depender de la capacidad del conocimiento de expertos del área para extraer y confeccionar las reglas.
- Limitado número y generalidad de reglas, debido a que se deben construir manualmente.
- Inconsistencias al crecer el número de reglas.
- Dificultad para cubrir un gran rango de dominios.

- **Aproximación Estadística:**

  Se basa en el modelado de la señal de habla usando algoritmos estadísticos. Se diferencia de la aproximación anterior en que el sistema puede aprender y extraer automáticamente el conocimiento del conjunto de datos disponibles. Además emplea los patrones de cada segmento de habla sin el paso intermedio de segmentación ni extracción de características.

  Como se puede ver en la figura 13, este tipo de reconocedores presenta dos fases de funcionamiento: entrenamiento y compa-

ración de patrones. Durante la fase de entrenamiento se supone que si se brinda al sistema suficientes versiones de los patrones a ser reconocidos, el algoritmo de entrenamiento debería ser capaz de caracterizar de manera adecuada los rasgos acústicos que permitan individualizar cada clase de manera más robusta y repetible, para el conjunto de datos presentados. Como resultado de esta fase se obtienen modelos o ejemplares prototípicos para cada clase.

El sistema entrenado es empleado luego en la fase de comparación de patrones, en la que un fragmento de habla desconocido se contrasta con cada uno de los modelos aprendidos, seleccionando finalmente la clase que maximiza alguna medida de similitud entre patrones.

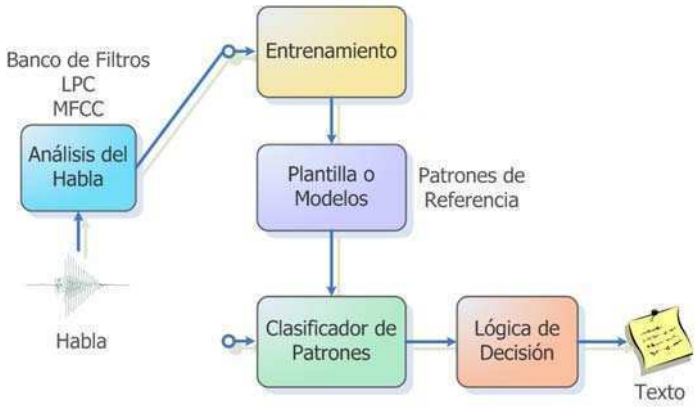

Figura 13: Esquema de sistemas para RAH estadísticos

La aproximación estadística también se puede ver como un caso especial de reconocimiento de patrones .

Sea Ω el universo de patrones que resultan de interés para una aplicación dada y que se puede segmentar en κ categoría de objetos, cuyo valor puede o no conocerse a priori. Sea por otro lado $\mathbf{x}$, un vector de p variables (atributos) definidas sobre objetos de Ω y que en conjunto brindan cierta clase de descripción numérica de dichos objetos. Sea X el espacio de atributos, dominio de variación de $\mathbf{x}$ correspondiente a todos los patrones de Ω. Sean $C_1, C_2, \cdots, C_\kappa$ las correspondientes categorías o clases de los patrones.

El problema de reconocimiento de patrones consiste en determinar, para cualquier patrón $\mathbf{x} \in \Omega$ de categoría desconocida, la clase a la que pertenece. Esencialmente la aproximación general para resolver el problema es encontrar una partición de X en $X_1, X_2, \cdots, X_\kappa$, tal que ante un patrón desconocido $\mathbf{x} \in X_j$, se pueda inferir que el nuevo patrón viene de la clase C_j.

En la figura 14 se representa el proceso de clasificación de patrones: en la primera fase se registra y digitaliza la señal de entrada.

El patrón resultante f es preprocesado, para reducir ruido o realzar algunas propiedades relevantes y se obtiene h. A continuación se extrae un conjunto de atributos de la señal preprocesada y se los ordena en un vector $x \in \Re^p$. Este paso de extracción de características se puede ver como una transformación de cada patrón, tal que la cantidad de datos se reduce al mismo tiempo que se intenta conservar la mayor cantidad posible de información relevante. Asimismo el clasificador impone algunas condiciones durante la selección de los atributos. En general los atributos correspondientes a una clase deberían ocupar una o más regiones compactas dentro del espacio de atributos, siendo deseable que las regiones pertenecientes a las distintas clases queden separadas. Finalmente, en el último paso de este proceso la tarea del clasificador consiste en mapear cada vector de atributos en un entero:

$$x \mapsto C \in \{1, 2, \cdots, \kappa\} \tag{1.5}$$

Donde se distinguen las κ clases diferentes: $C_1, C_2, \cdots, C_\kappa$. Además se puede utilizar una clase especial de rechazo C_0 para aquellos casos en que el vector de atributos no se pueda asociar con suficiente confianza con ninguna de las clases existentes.

Figura 14: Estructura general de un clasificador de patrones

Solamente consideraremos clasificadores consistentes en funciones paramétricas. Tales parámetros se optimizan utilizando una muestra representativa de patrones $\omega = \{x^1, x^2, \cdots, x^\kappa\}$, tal que se minimice cierto criterio de riesgo. La clasificación está basada en una regla de decisión, que a su vez puede sustentarse en una medida de distancia o en probabilidades de clases. Las reglas de decisión óptimas que minimizan las probabilidades de errores, deciden por la clase C_k con máxima probabilidad a posteriori $p(C_k|x)$, que es la probabilidad de la clase C_k dado el vector de atributos x. El problema principal en el entrenamiento del clasificador consiste en encontrar una buena estimación de las probabilidades a posteriori.

Generalmente se usa la regla de Bayes:

$$p(C_k|x) = \frac{p_\lambda \cdot p(x|C_\lambda)}{\sum_{\kappa=1}^{k} p_\lambda \cdot p(x|C_\lambda)} \tag{1.6}$$

Tal que las probabilidades a posteriori p_λ y la función de verosimilitud $p(x|C_\lambda)$ se puedan estimar independientemente. A esta clase de clasificadores se denomina clasificadores estadísticos. A menudo se utiliza un clasificador de distribución normal, donde:

$$p(x|C_\lambda) = p(x|\mu_\lambda, \Sigma_\lambda) = \sqrt{(|2\pi\Sigma_\lambda|)} \cdot e^{\frac{-(x-\mu_\lambda)_t \cdot \sum_{-1}^{\lambda} (x-\mu_\lambda)}{2}} \quad (1.7)$$

Esta expresión requiere que los vectores de atributos sigan una distribución normal multivariada. Los parámetros de las distribuciones: los vectores de medias condicionales μ_λ y las matrices de covarianzas Σ_λ deben estimarse empleando muestras etiquetadas $\omega_\lambda \subset \omega$ a través de un proceso de aprendizaje supervisado.

Si los atributos no están distribuidos de acuerdo a una distribución normal multivariada, se puede aproximar la distribución real mediante mezclas de distribuciones normales, definidas de la siguiente manera:

$$p(x|C_\lambda) = \sum_{\nu=1}^{L_\lambda} \omega_{\lambda\nu} \cdot p(x|\mu_{\lambda\nu}, \Sigma_{\lambda\nu}) \quad (1.8)$$

con $\sum_{\nu=1}^{L_\lambda} \omega_{\lambda\nu} = 1$

Para el proceso de entrenamiento generalmente se dispone de muestras etiquetadas de la forma ω_λ, mientras que las muestras etiquetadas que requiere la ecuación anterior son de la forma $\omega_{\lambda\nu}$. Por esto los parámetros de cada mezcla de funciones de distribución deben hallarse empleando un entrenamiento no supervisado. Para el caso de la estimación de máxima verosimilitud este proceso se realiza empleando el algoritmo EM (expectation maximization).

Otra posibilidad es la de utilizar clasificadores de distribución libre. En general para κ clases estos clasificadores están basados en un vector de funciones discriminantes paramétricas $d_\lambda(\mathbf{x}, \mathbf{a})$ de la forma:

$$\mathbf{d} = (d_1(\mathbf{x}, \mathbf{a}) \cdots d_\kappa(\mathbf{c}, \mathbf{a}))_t \ \text{ con } \mathbf{a} \in \mathfrak{R}^\kappa \quad (1.9)$$

Un vector de características $\mathbf{x}$ se clasifica de acuerdo a la regla de decisión:

$$k = \underset{1 \leqslant \lambda \leqslant \kappa}{\operatorname{argmax}} \left[d_{\lambda(x,a)} \right] \quad (1.10)$$

El problema en el diseño de este tipo de clasificadores es definir una familia apropiada de funciones paramétricas d_λ y optimizar

sus parámetros $a_1, \cdots, a_\kappa$. Sea la siguiente función discriminante ideal δ:

$$\delta = (\delta_1(x), \cdots, \delta_\kappa(x)) \tag{1.11}$$

$$\text{donde } \delta_\kappa(x) = \begin{cases} 1 & \text{si } x \in C_k \\ 0 & \text{en otro caso} \end{cases}$$

Con la ecuación anterior y dadas las muestras de datos etiquetadas, se puede usar como criterio de optimización la esperanza del error cuadrático:

$$\epsilon = E\left\{(\delta - d)^2\right\} \tag{1.12}$$

Se ha demostrado que si la función d es suficientemente compleja, la función d^* que minimiza el criterio de la ecuación anterior es idéntica al vector de probabilidades a posteriori [Nie90a].

$$d^* = (p(\omega_1|x), \cdots, p(\omega_\kappa|x)) \tag{1.13}$$

Esto por lo tanto significa que el clasificador de distribución libre y general minimiza la probabilidad de los errores.

1.3.3 Estado actual y principales desafíos en RAH

En la actualidad la tecnología del RAH se utiliza en muchas aplicaciones comerciales, y su campo de investigación ha alcanzado un considerable grado de madurez. Lo que hace seis décadas comenzó como un sistema de reconocimiento de dígitos aislados para un solo locutor, hoy en día, se ha extendido hasta sistemas capaces de reconocer habla espontánea fluida, de gran vocabulario, y de manera independiente del locutor.

Sin embargo, como ya se mostró, al comparar el desempeño actual de estos sistemas con el de los seres humanos, se observa que aún resta mucho por mejorar, y que el problema no se puede considerar resuelto.

En las aplicaciones tecnológicas actuales las imperfecciones técnicas se reducen o compensan ajustando la interacción del sistema a la medida del contexto o del usuario. Por ejemplo haciendo el sistema dependiente del locutor (como en sistemas de dictado), limitando el vocabulario (como en sistemas de discado por voz), o empleando una sintaxis simple y predecible (como en introducción de datos o sistemas de comando y control).

Por otra parte, en algunas aplicaciones la precisión de reconocimiento no necesita ser perfecta, por ejemplo en la tarea de recuperación de información vía voz, con una precisión de reconocimiento en orden al

70 % es posible obtener tasas de recuperación de información similares a los logrados a partir de transcripciones manuales del habla.

Actualmente la investigación en este tema se centra en reconocimiento de habla espontánea, multi-locutores y vocabularios muy grande, soportando condiciones de múltiples acentos.

1.4 INFORMACIÓN SUPRASEGMENTAL

Se puede observar a la señal de habla desde diferentes escalas o niveles de abstracción temporal.

En el nivel de mayor granularidad encontramos el rango temporal correspondiente al período de muestreo, resultante del proceso de conversión de la señal analógica a su representación digital. Como el rango de frecuencias de muestreo más usuales suele ir de 8000 Hz (para el caso de canales telefónicos) hasta 44100 Hz, el rango de duraciones de la primera escala temporal se puede considerar entre 20 y 125 µs.

Es posible ver como una segunda escala la que surge del período de un ciclo glótico, que va de 5 a 10 ms. Este tiempo también corresponde al intervalo de separación entre ventanas consecutivas de análisis que utilizan la mayoría de los sistemas de RAH, que va de 5 a 20 ms.

En tercer lugar se puede considerar el rango temporal asociado con los movimientos de los articuladores, que va desde $20 - 50$ ms hasta varios cientos de milisegundos. A este rango se lo puede considerar el segmental o fonético. El siguiente nivel está dado por el **nivel suprasegmental o prosódico** , que puede ir desde 200 ms hasta las duraciones típicas de las oraciones, digamos $1 - 10$ s. Posteriormente se puede definir otra escala de evaluación global en que se consideran tendencias, o comportamientos promedios sobre varias oraciones.

nivel suprasegmental o prosódico

1.4.1 Aspectos básicos de la prosodia

Utilizar información suprasegmental dentro de los sistemas de RAH supone en primer lugar aislar la información lingüística útil, de aquella que no lo es. Esto no es una tarea sencilla, ya que se pueden encontrar los siguientes factores que afectan las características de los rasgos prosódicos [181]:

1. **Aspectos condicionados fonéticamente, valores intrínsecos y coarticulación.**

   Parte de las variaciones observadas en el tono, la duración y la intensidad dependen de las secuencias particulares de sonidos emitidos. A estas variaciones se conoce como aspectos de los parámetros prosódicos condicionados fonéticamente.

   Tales aspectos están condicionados por diferencias en los mecanismos fisiológicos involucrados en la producción de cada sonido individual por un lado, y además por restricciones coarticulatorias temporales, debidas al solapado de gestos articulatorios correspondientes a las secuencias de fonemas pronunciados.

a) Las características intrínsecas de los fonemas.

En particular, se sabe que la duración, la frecuencia fundamental, y la intensidad de las vocales están correlacionadas con la altura de la lengua.

Las vocales altas como la *"i"* tienen intrínsecamente un pitch más alto, una duración más corta y una intensidad inherentemente inferior que las vocales bajas como la *"a"*. Las vocales nasales en el Francés son intrínsecamente más largas que aquellas que involucran más a la cavidad bucal. Las vocales tensas en el Inglés son intrínsecamente más largas y más laxas que las vocales cortas.

b) El contexto inmediato.

Las vocales en un contexto consonante sordo, tienen un pitch relativamente más alto, y son más cortas con respecto a las mismas vocales en un contexto sonoro.

La influencia de la sonoridad sobre la duración de la vocal precedente se incrementa cuando tanto la vocal, como la consonante siguiente forman parte de la misma sílaba. La duración de una vocal por lo tanto depende de la localización de los límites silábicos, y de la característica sonoro-sordo de la consonante siguiente.

En general, tales aspectos son independientes del locutor y del lenguaje en cuestión. Pueden contribuir indirectamente a la identificación de los sonido subyacentes, pero no tienen una función lingüística *per se*. En la decodificación automática de la información prosódica, podría ser útil normalizar los parámetros prosódicos para compensar tales variaciones, pero esto solamente se puede llevar a cabo una vez que se ha identificado los segmentos correspondientes.

2. **Aspectos Lingüísticos de la Prosodia: Sintaxis y Ritmo.**

La porción de las variaciones de tono, duración, e intensidad que no está condicionada por restricciones puramente articulatorias es la responsable de transportar el mensaje lingüístico. Estos aspectos prosódicos de origen lingüístico son los de mayor interés para la decodificación automática de frases. Son los portadores por un lado de la información respecto a diferentes unidades lingüísticas (a nivel fonético, silábico, sobre el acento tonal de palabras y límites de palabras, sobre tipos y límites de frases y de oraciones), y por otro lado información respecto a la estructuración acústica de cada unidad en unidades mayores (cómo se agrupan fonemas en sílabas, sílabas en palabras prosódicas, palabras prosódicas en frases).

En general, los aspectos de la prosodia condicionados lingüísticamente, así como la estructura prosódica, son independientes del locutor, y en cierta medida independientes del lenguaje.

Ahora bien, se puede observar que estos parámetros prosódicos están siendo gobernados al mismo tiempo tanto por la organiza-

ción sintáctico-semántica de las frases, como por principios rítmicos del habla.

Como puede haber diferencias substanciales entre las características impuestas sobre las unidades acústicas según se impongan restricciones sintáctico-semánticas o rítmicas, ambas influencias pueden entrar en conflicto.

Las influencias sintácticas resultan en variaciones de duración y F0 para contrastar y marcar límites entre las secuencias de unidades. Las influencias rítmicas resultan en una tendencia de las sílabas y los acentos a sucederse en intervalos temporales regulares. Las duraciones de los segmentos se comprimen o expanden para preservar como un invariante la duración global de las sílabas (isosilabicidad) y para ecualizar los intervalos de tiempo entre dos sílabas acentuadas (isocronía). Al mismo tiempo, las unidades se acortan o alargan para marcar la organización sintáctico-semántica.

La interacción simultánea de esos tres tipos de influencias: aspectos condicionados fonéticamente, tendencias rítmicas y organización sintáctico-semántica, es el motivo principal para explicar las dificultades a la hora de interpretar las causas de un alargamiento u acortamiento de las unidades a nivel silábico, o fonético.

Las tendencias rítmicas podrían ser explotadas de una manera predictiva, para obtener información por ejemplo sobre los límites de las sílabas.

Una de las funciones más interesantes de la prosodia desde el punto de vista del reconocimiento automático del habla es el agrupamiento de palabras relacionadas semánticamente, tales como el sustantivo y el adjetivo que lo precede o lo sigue en una sola unidad, la denominada palabra prosódica, y expresar diferentes grados de agrupamiento entre distintas palabras prosódicas. Sin embargo la segmentación y estructuración del habla a través de la prosodia no permite recuperar la estructura sintáctica completa de las oraciones por ejemplo en el habla conversacional.

Sí puede suceder que al estudiar oraciones aisladas, pronunciadas de manera neutral y cuidadosa, se encuentre que la estructura prosódica mapee exactamente la estructura sintáctica. No obstante en la mayor parte de los casos no hay una correspondencia uno a uno entre estructura prosódica y estructura sintáctica.

Como resultado, la información prosódica decodificada a partir de la señal, debe ser compatible con las hipótesis léxicas y sintácticas.

Debido a que es necesario considerar posibles variaciones intra e interlocutores, son más fáciles de concebir los algoritmos para chequeo de compatibilidad en el nivel léxico y sintáctico (en el proceso de verificación) que algoritmos para predicción (en el proceso de generación de hipótesis).

3. **Aspectos no Lingüísticos: Sentimientos del Locutor.**

Un locutor también puede pronunciar una frase de manera más o menos marcada (en contraste con una pronunciación neutra).

En estos casos el locutor controla las variaciones prosódicas para comunicar eventualmente algún conocimiento sobre su actitud hacia lo que está diciendo: duda, ironía, compromiso, convicción, etc.; o para dirigir la atención del oyente hacia las palabras más importantes de su discurso.

Existe un continuo entre maneras completamente neutras y muy marcadas de pronunciar una frase.

Los correlatos acústicos de algunas de esas intenciones como la duda, o la ironía aún no se han investigado lo suficiente, y para simplificar la decodificación prosódica de una frase, se suele aislar este componente, haciendo que el locutor pronuncie las oraciones de una manera más o menos neutra. Sin embargo el sistema debería ser suficientemente flexible para tratar con cierta cantidad de énfasis (tal como la asignación de acento enfático o focal sobre palabras particulares en una oración).

Éste aspecto de la prosodia es particularmente importante en los sistemas de diálogo donde los locutores tienden a enfatizar palabras clave para desambiguar una pregunta o una respuesta.

4. **Aspectos Paralingüísticos: Diferencias Fisiológicas y Dialectales.**

   La presencia de patrones prosódicos que resulten poco frecuentes para un oyente, pueden servir para informarle acerca de características particulares del locutor, entre ellos: acento patológico, acento extranjero, estado emocional, o condición física. Se debe notar que estos aspectos generalmente no se encuentran bajo el control voluntario del locutor.

   Los dos últimos aspectos de la prosodia (no lingüísticos y para lingüísticos), que son dependientes del locutor, pueden oscurecer los aspectos condicionados lingüísticamente.

   Ninguno de los sistemas actuales de RAH han resuelto adecuadamente el problema de la adaptación automática a particularidades idiosincráticas del locutor. Los sistemas existentes esperan que los locutores difieran muy poco entre sí y que el estilo de habla sea más o menos neutro.

En la figura 15 se muestra un modelo sistemático del proceso de codificación de información en términos de atributos prosódicos del habla, propuesto en [47].

Este modelo considera la prosodia y sus correlatos acústicos, como el resultado de un proceso complejo de múltiples etapas, sujeto a restricciones en cada una de ellas.

La información de alto nivel (información de entrada) se codifica en unidades y estructuras abstractas de un lenguaje particular, a través de la etapa de planificación del mensaje. Esa etapa está guiada por reglas y restricciones que se pueden considerar como la gramática del lenguaje.

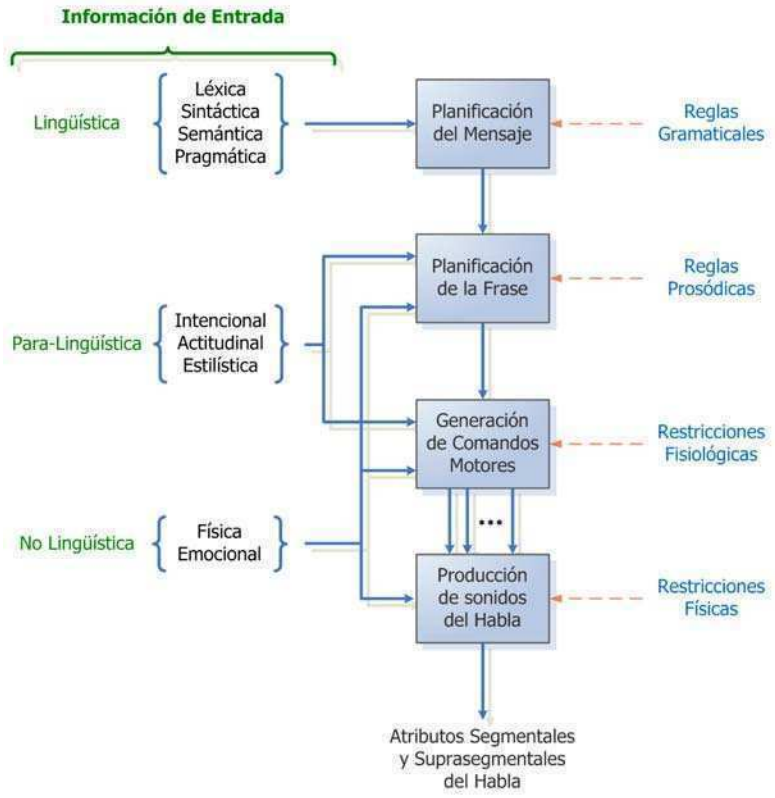

Figura 15: Proceso a través del que se codifica información en los atributos segmentales y suprasegmentales del habla. Los niveles más altos de información de entrada (lingüística, paralingüística y no lingüística) se transforman en sonidos a través de un proceso de múltiples etapas.

En la siguiente etapa se planifica la elocución teniendo en cuenta el fraseo, la acentuación y los principios de utilización de pausas para el lenguaje en cuestión. Además se introduce por primera vez la información paralingüística y no lingüística, determinando por ejemplo el estilo y los agrupamientos de los elementos del fragmento de habla.

La fase de planificación determina los comandos neuro-motores para el control de los mecanismos de producción del habla.

Al final del proceso ilustrado, el contenido de información de la locución se convierte en correlatos acústicos como acentos, frases y pausas. Tanto la etapa de generación de comandos como la de producción del habla están condicionadas por restricciones anatómicas y fisiológicas, como por ejemplo: un repertorio acotado de movimientos articulatorios, un rango limitado de valores de F0, o velocidad de articulación limitada.

Esta superposición de interacciones entre componentes lingüísticos, para-lingüísticos y no lingüísticos hace difícil encontrar una correspondencia clara y unívoca entre las características físicas observables del habla, y la información subyacente contenida en las secuencias.

1.4.2 El rol de la prosodia en la decodificación humana

Naturalmente, se argumenta que los sistemas de RAH deberían tratar de imitar la forma en que los humanos procesamos el habla para llegar a los mismos niveles de desempeño. De esa forma, antes de proponer el empleo de información prosódica en sistemas de RAH se debería cuestionar el uso real de los parámetros prosódicos en los seres humanos. ¿Emplean los seres humanos pistas prosódicas? ¿Cómo y cuándo las utilizan? ¿Cuánto contribuyen las pistas prosódicas a la decodificación del mensaje hablado?

A pesar de la aparente complejidad en la manifestación de los parámetros prosódicos, los oyentes parecen no tener dificultades para decodificar la información transportada prosódicamente, al menos en experimentos controlados.

Resulta difícil evaluar el uso efectivo de los parámetros prosódicos en las conversaciones de todos los días, y también su contribución exacta en la decodificación completa de las frases. A pesar que resulta sencillo diseñar pares de oraciones que sólo puedan ser desambiguadas mediante el uso de parámetros prosódicos (la ubicación de una juntura o la posición del acento en una palabra), los casos en los que los rasgos prosódicos resultan vitales para la comprensión del discurso no son muy comunes.

Es por eso que a veces se dice que los parámetros prosódicos son redundantes con respecto a la información espectral, y se afirma que los parámetros espectrales son suficientes para decodificar unívocamente una frase.

La hipótesis de redundancia también se puede cuestionar. Por un lado se ha demostrado que los oyentes prestan atención a la continuidad prosódica, aún a expensas de la continuidad semántica. Por otra parte, en experimentos perceptuales se encontró que al incrementar gradualmente el grado de compresión temporal de una señal de habla, se produce una disminución sistemática de la inteligibilidad de la misma, sin embargo el debilitamiento de la capacidad de comprensión muestra una menor tasa de disminución cuando las frases se oyen con entonación normal, respecto a si se escuchan con entonación anómala.

En tercer lugar, entender habla sintética que no posee un buen modelo prosódico implica un esfuerzo significativo y conlleva una pérdida de atención por parte de los oyentes. Además descifrar espectrogramas sin el empleo de información como la curva de F0 y de energía resulta más difícil que empleando toda la información disponible.

Dado este conjunto de evidencias, resulta razonable plantear la hipótesis que no solamente los aspectos espectrales sino los temporales y F0 son entradas paralelas a los oídos del oyente, quien las decodifica simultáneamente, y que es necesario el tratamiento simultáneo de todas las entradas para el reconocimiento automático del habla continua.

Esta visión es la aceptada en general por la comunidad psicolingüística.

El rol del acento léxico.

El rol exacto desempeñado por el acento a nivel de palabras en el procesamiento del habla continua, no está establecido con claridad. Sin embargo distintos estudios han investigado el uso potencial de la detección de la sílaba acentuada en el RAH:

En el reconocimiento de palabras aisladas, la determinación de la sílaba que recibe el acento es muchas veces indispensable para diferenciar palabras. Por ejemplo el par papa - papá puede diferenciarse exclusivamente sobre la base de la posición de su sílaba acentuada.

En el Inglés las sílabas acentuadas parecen constituir **islas de confiabilidad** donde las pistas acústicas presentan una mayor robustez. Estudios sobre la lectura de espectrogramas efectuada por expertos, y análisis de confusión fonética en el resultado de decodificadores acústico-fonéticos, confirman el hecho que las vocales acentuadas se reconocen con menor confusión que las no acentuadas. Lo que se puede explicar teniendo en cuenta que las sílabas acentuadas suelen ser más largas.

Esta noción se puede comparar con la noción de **dominancia acústica** descripta para el Francés: las consonantes dominantes son aquellas localizadas en porciones de la oración donde la frecuencia fundamental está en ascenso. Estas consonantes presentan características más similares a las *prototípicas*, que las mismas consonantes en otros contextos. Sin embargo se debe notar que la dominancia fonémica es una noción acústica y se determina por la posición de los fonemas en la estructura prosódica de la frase, mientras que el acento es una noción fonológica.

dominancia acústica

En varios estudios teóricos se mostró la utilidad de detectar los acentos para el acceso multi-nivel al lexicón.

Se ha encontrado que la información de los patrones de acentos puede reducir el espacio de búsqueda en Ingles e Italiano.

En el Inglés, dividir los fonemas de sílabas acentuadas en categorías fonéticas más amplias (manera de articulación), dejando como incógnita las sílabas no acentuadas, resulta casi tan restrictivo como representar al vocabulario mediante seis clases fonéticas. La información acentual por lo tanto representa restricciones potencialmente muy útiles para el reconocimiento de palabras aisladas de gran vocabulario. Empleando un lexicón de 15000 palabras y asumiendo una convención para la descripción de cada sílaba de tres niveles (acentuadas, inacentuadas y reducidas), y una sílaba acentuada por palabra, en [9] mostró que si se conoce el número de sílabas de la palabra, se reduce la clase de palabras candidatas a un 41 % del tamaño del lexicón. Cuando se conoce el patrón acentual, es decir, el número correcto de sílaba y la asignación de acentos correcta para cada sílaba, el tamaño de clase esperada se reduce al 19 % del tamaño del lexicón. Si sólo se conoce el número de sílabas y la ubicación de la sílaba acentuada el tamaño de las palabras compatibles es del 22 % del tamaño original.

Para el Italiano, en [135] se mostró que empleando un lexicón de 12000 palabras, las restricciones impuestas por el conocimiento del número de sílabas y la ubicación de la sílaba acentuada, permite reducir

la cohorte de palabras candidatas a un 4,3 % del tamaño del lexicón completo.

Función de los movimientos de F0.

Tal vez debido a que un aumento en el valor de la frecuencia fundamental se realice tensando las cuerdas vocales, y sus descensos mediante la relajación de las mismas, en muchos lenguajes un ascenso de F0 entre dos vocales sucesivas parece asociarse con la noción de comienzo, y un movimiento de caída de F0 con el de finalización.

Como un primer resultado, generalmente se encuentra una disminución en el F0 al final de cada constituyente sintáctico primario, y un incremento de F0 cerca del comienzo del siguiente constituyente.

A este patrón se conoce como de Ascenso-descenso y se utiliza para demarcar constituyentes. Como segundo resultado, los ascensos y descensos de F0 generalmente aparecen de a pares, provocando un patrón de sombrero en el contorno de la F0, que se utiliza para señalar fenómeno de agrupamiento de palabras.

Por otra parte, el número de ascensos y descensos en la curva de F0 depende de la velocidad de elocución. El número de movimientos pronunciados en F0 disminuyen al aumentar la velocidad del habla. Además, debido a restricciones rítmicas también existe una tendencia que hace corresponder un patrón completo de subida y bajada de F0 para las palabras largas, mientras que las palabras cortas suelen reagruparse en un solo patrón.

Es más común encontrar un patrón de segmentación (patrón de caída y ascenso) entre la frase nominal del sujeto y el verbo de una oración cuando dicha frase nominal es larga en términos de número de sílabas. En consecuencia, para una misma frase y velocidad de habla, cierto tipo de patrón prosódico es más frecuente de encontrar que otros. Sin embargo es necesario integrar la posibilidad de diferencias entre locutores en el sistema.

Los mismos principios básicos gobiernan la determinación de parámetros prosódicos para todos los locutores, pero la contribución relativa de variaciones condicionadas por factores fonéticos, el acento, el ritmo, la sintaxis, la semántica, el estilo y la velocidad son dependientes del locutor.

Cada locutor tiende a ser consistente, aunque se pueden encontrar diferencias regulares para un mismo locutor. El estudio de regularidades dependientes del locutor es muy importante para el RAH: el ajuste del sistema a los hábitos particulares del locutor es un prerrequisito ineludible para la extracción y el empleo completo de la información prosódica en RAH. No se conocen estudios sistemáticos modelando variaciones intra e inter-locutores.

Empleo de pausas y alargamientos como marcadores de límites.

El análisis acústico de habla fluida revela que los locutores insertan un gran número de pausas mientras hablan. Las pausas respiratorias representan sólo una porción del conjunto de pausas encontradas, también se pueden hallar pausas llenas para indicar que un locutor con-

servará su turno de elocución, que está pensando o vacilando. Sin embargo, sin importar el tipo de pausa considerada, todas se presentan en fronteras gramaticales.

El uso de pausas como marcadores de límites primarios tanto entre como dentro de una frase parece ser una invariante independiente del lenguaje.

Por otra parte, en muchos lenguajes como Español, Inglés, Alemán, Francés, Italiano, Ruso, o Sueco, se observa una tendencia hacia el alargamiento de los elementos finales de la estructura lingüística, particularmente la última vocal antes de una pausa, así como alargamiento en elementos finales de palabras y frases. Parece que el alargamiento como la caída de F0 se asocia principalmente con la noción de finalización.

1.4.3 Empleo de Información Suprasegmental en el RAH

A lo largo de la evolución de los sistemas de RAH se puede observar una constante mejora en los índices de desempeño, a partir de la incorporación de más y mejor información acerca de las estructuras lingüísticas, propias del lenguaje a ser reconocido. Sin embargo, muchos de los niveles estructurales del habla que han sido estudiados ampliamente en el dominio de la lingüística, aún no se han incorporado a estos sistemas.

Es por ello que algunos de los principales tópicos de investigación en el campo de RAH rondan en torno al desarrollo de métodos que permitan incorporar fuentes adicionales de conocimiento, a la información espectral de tiempo corto ya utilizada en el proceso de reconocimiento del habla[86].

Entre los niveles lingüísticos que todavía no han sido explotados en los sistemas de reconocimiento estándar se encuentran la prosodia, sintaxis, semántica, pragmática, y contexto del habla. A pesar que se puede observar cierto consenso respecto a que estas fuentes de conocimiento podrían mejorar el desempeño de los sistemas de reconocimiento actuales, hasta el momento no se ha podido cuantificar con claridad sus alcances, y se han reportado muchas dificultades para su modelado, incorporación y adaptación a las arquitecturas típicas de los reconocedores actuales.

Por otra parte, se sabe que la prosodia desempeña un rol importante en el proceso de percepción humana del habla. Es indudable que la prosodia brinda información lingüística, para-lingüística y extralingüística que los oyentes utilizan no sólo para complementar la información léxica sino también para segmentar grupos de palabras, focalizar palabras de contenido, desambiguar significados y filtrar el resto de la información de transporte o función.

Las funciones comunicativas de la prosodia y el hecho que los sistemas de RAH convencionales prácticamente no la aprovechen, ha despertado un gran interés en la comunidad del procesamiento automático del habla al menos hace unas tres décadas [100].

En la literatura se puede ver que no han sido pocos los intentos por aprovechar esta información como factor de mejora en diferentes fa-

ses del proceso de reconocimiento automático. Sin embargo, debido a que la prosodia está vinculada simultáneamente con distintas *capas* de información (lingüística, para-lingüística y no lingüística), sus manifestaciones acústicas son complejas y presentan muchas variaciones, lo que conspira con su aplicación en el proceso de reconocimiento.

A ello se suma la dificultad de integración de información de granularidades temporales diferentes como lo son la segmental y la suprasegmental en los sofisticados esquemas que presentan los sistemas de reconocimiento actuales. Esa integración implica conocer cuándo y dónde integrar el conocimiento prosódico dentro del sistema y cómo combinar las evidencias obtenidas a partir de diferentes fuentes.

En contraste, se pueden encontrar otras áreas de la tecnología del habla en las que se ha empleado con resultados positivos esta fuente de información. Entre ellas, identificación de género [39], identificación de lenguaje [143], identificación y verificación de locutores [2], enseñanza de segunda lengua [5], y especialmente en los sistemas de conversión de texto a habla (CTH), que vienen empleando exitosamente la información prosódica hace varios años, especialmente en relación a la naturalidad de la voz sintetizada [167, 169]. Los estudios y modelos propuestos en el área de CTH brindan una amplia fuente de conocimientos acerca de cómo se manifiesta la prosodia en el lenguaje natural.

Veamos ahora algunos antecedentes que plantean utilizar información suprasegmental en el reconocimiento del habla.

En 1960 y Lieberman mostró que es posible determinar automáticamente los patrones de acento de palabras bisilábicas pronunciadas de manera aislada con un error en el orden de un 1 % [105].

A principio de la década de 1970 hubo una serie de trabajos de Lea que proponían diferentes usos de los parámetros prosódicos en el reconocimiento del habla. Lea sugiere que los parámetros prosódicos se pueden utilizar en todos los niveles dentro del proceso de decodificación de un sistema de reconocimiento automático del habla.

Propone que esta información se puede emplear por ejemplo a nivel acústico-fonético para separar las consonantes oclusivas sonoras y sordas, a nivel léxico para detectar la posición del acento en la palabra, a nivel sintáctico para localizar los límites fonológicos principales de una oración y la estructura de los constituyentes sintácticos, y en el nivel pragmático para localizar la porción enfática de un discurso.

Se puede pensar en varias alternativas para hacer uso de los atributos prosódicos en el proceso de RAH. Una es condicionar el análisis de los atributos acústicos en función de la información suprasegmental, asumiendo que las funciones de transferencia del tracto vocal están relacionadas con la frecuencia fundamental. En [121] se demostró que el valor de F0 influye claramente sobre los valores de los coeficientes cepstrum de vocales y consonantes sonoras (como las nasales). Sin embargo esas relaciones fueron dependientes del locutor, y algunos sonidos no mostraron estar influidos por el valor de F0, por lo que debería investigarse más antes de aplicar esta relación en el modelado acústico.

Otra forma de utilizar la prosodia consiste en detectar los límites prosódicos para condicionar el proceso de reconocimiento. Una forma sencilla de implementar esta estrategia es determinar los límites prosódicos antes del proceso de reconocimiento y usar esa información para segmentar los fragmentos de habla a reconocer. Si el método funciona debería mejorar las tasas de reconocimiento y reducir el tiempo de procesamiento.

Aunque se probaron varias alternativas siguiendo esta estrategia, no se encontraron mejoras significativas. El problema principal viene dado por el pobre desempeño de los detectores de límites de constituyentes, y un alto grado de dependencia del locutor y el tipo de habla. Las tasas de detección de fin de frases podría mejorar utilizando múltiples manifestaciones suprasegmentales como: valles en el perfil de F0, transiciones abruptas en los valores de F0, alargamiento en la duración de fonos, y/o adoptando métodos estadísticos. Sin embargo los resultados en esta materia sugieren que para esta tarea los atributos prosódicos no son suficientes, sino que también es necesaria la información segmental [76]. Se podría pensar en usar la información de segmentación fonética de la primera pasada de reconocimiento para mejorar la detección de los límites de frase y utilizar la información de fronteras de frase en la segunda pasada del reconocedor. Se propusieron diversos métodos de acuerdo a esta idea. El caso más extremo quizás sea el del etiquetado prosódico automático, donde el texto correspondiente a un fragmento de habla se conoce de antemano y el decodificador realiza un alineamiento forzado entre el texto y la señal de entrada. Si bien este caso no es adecuado para construir reconocedores, es un tema de investigación importante para la construcción de grandes cuerpos de datos prosódicos .

Por otro lado se puede encontrar una gran cantidad de trabajos en donde se emplea información suprasegmental en la determinación del tipo de acento de palabra para el Japonés y el tipo de tono para el Chino. Para el caso del Chino, cada sílaba puede tener cuatro significados diferentes dependiendo del tipo de tono, se habla entonces de tonos léxicos. Esto vuelve al proceso de identificación de tonos en una pieza esencial dentro del proceso de RAH.

En el caso de reconocimiento de acento/tono en el habla continua, la información de límites segmentales es indispensable. Usando el F0, energía y sus derivadas como vectores de parámetros acústicos en el marco de HMM, y teniendo en cuenta los efectos de sílabas precedente y siguiente, se obtuvieron resultados próximos al 90 % [195]. No obstante la mayoría de los sistemas de RAH para el chino no incorporan de manera estándar el reconocedor de tonos. Eso se debe en parte a que la tarea de reconocimiento no es tan complicada.

Para el caso del Japonés, el reconocimiento del tipo de acento correcto es un poco más complicado, debido a que cada palabra incluye varias sílabas y su contorno de F0 varía en función de varios factores, como la posición de la palabra en la frase. En [76] se propone un modelo para emplear información de tipos de acentos para obtener una verosimilitud de palabras para el Japonés, a partir del análisis de perfiles de F0 a nivel de moras.

En este sentido, conviene resaltar que la información suprasegmental es altamente dependiente del lenguaje, tanto respecto a sus manifestaciones acústicas, como en la información que codifica. Por ello no suele ser válido extrapolar resultados de un idioma a otro.

También se ha empleado la prosodia en el modelado de duraciones, intentando mejorar el desempeño de los sistemas de RAH. En [25] se intenta aprovechar los patrones de duración de segmentos fonéticos y de pausas para extraer el contenido lingüístico de una frase. Se propone un modelo de duraciones que opera de manera jerárquica a nivel fonológico, fonémico, silábico y morfológico. Se analizan dos modelos estadísticos de duraciones, uno basado en duraciones relativas entre unidades subléxicas, y otro en duraciones absolutas normalizadas con respecto a la velocidad de elocución. Empleando estos modelos en el reconocimiento fonético se reporta una mejora relativa de la clasificación de hasta un 7,7 %.

En [76] se introduce un esquema para utilizar la prosodia en el control del haz de búsqueda durante la fase de decodificación de los reconocedores. En la estrategia propuesta se aprovechan dos elementos encontrados en el reconocimiento del habla. Por una parte, el hecho que lingüísticamente hay un mayor grado de certezas respecto a la identidad léxica hacia el final de la palabra que en el comienzo de la misma; y que por otra parte durante el reconocimiento va aumentando la verosimilitud acústica de una palabra cuanto mayor cantidad de tramas de entrada se tengan como evidencias, el tamaño del haz de búsqueda requerido para contener la hipótesis óptima va disminuyendo hacia el final de las palabras. De esta forma se introduce un esquema de control del factor de poda, haciendo el espacio de búsqueda más amplio luego de un límite de constituyente prosódico, el que va disminuyendo hacia el límite siguiente.

En [44, 11] se propone usar la prosodia para derivar modelos de pronunciaciones dinámicos, intentando mejorar las tasas de reconocimiento del habla.

En la actualidad, si bien se puede encontrar un gran número de artículos desarrollados sobre el tema, la mayoría de los sistemas comerciales no hacen uso efectivo de la prosodia en el reconocimiento automático del habla.

Quizás una de las pocas excepciones en este sentido sean los trabajos vinculados con el proyecto **Verbmobil**, donde se emplearon los límites de frases entonativas para acotar el espacio de búsqueda (límites muchas veces marcados por pausas), y se utilizó el ascenso final de la curva de F0 para determinar la modalidad de las oraciones. Sin embargo sus autores afirman que la contribución más importante de la prosodia en este proyecto estuvo dada en la fase de comprensión del habla, más que en la de reconocimiento [131].

Finalmente, un estudio reciente publicado en [57] investiga los errores exhibidos por dos sistemas de RAH convencionales aplicados al reconocimiento de habla conversacional. En ese estudio se busca establecer las posibles causas de tales errores, indagando qué palabras tienden a ser mal reconocidas y por qué. Los autores concluyen que las palabras con mayores tasas de error incluyen a:

- las que presentan características prosódicas extremas,
- las que aparecen al comienzo de un nuevo turno de diálogo o como marcadores de discurso,
- *pares de doble confusión*: palabras similares desde el punto de vista acústico, y además con probabilidades similares dentro del modelo de lenguaje.
- disfluencias o palabras adyacentes a disfluencias,
- diferencias entre locutores.

En lo referente específicamente a los efectos prosódicos reflejado en el desempeño de los reconocedores, se indica que para el rango de valores típicos de la mayoría de los atributos acústico-prosódicos como: valor medio y rango de F0, valor promedio de intensidad, jitter y ritmo de habla, la influencia sobre los errores de reconocimiento es pequeña. Sin embargo, los efectos son significativos cuando estos atributos presentan valores extremos superiores o inferiores a los típicos, salvo para el caso de duración, que muestra menores tasas error de reconocimiento cuando sus valores superan la duración media.

Estos resultados indican que se podría esperar un aumento en las tasas de reconocimiento si los sistemas actuales pudiesen adaptarse a las variaciones prosódicas tanto intra como inter-locutor. Esta adaptación haría necesaria la representación explícita de aspectos prosódicos de la señal de habla en el reconocimiento.

Debido a que aún falta recorrer mucho camino para que los sistemas de RAH alcancen las capacidades de reconocimiento del ser humano [107], se propone incorporar en los mismos información suprasegmental, que permita caracterizar límites de frases y palabras, contenido léxico y rasgos sintácticos, diferencias entre elementos acústicos correspondientes al habla de aquellos que no lo son, con la hipótesis que esta información puede ser relevante para aumentar la robustez de los reconocedores especialmente en condiciones adversas, como ancho de banda del canal de transmisión acotado, ruidos de fondo, eventos acústicos del locutor que no corresponden a fragmentos de habla. Todos estos elementos se encuentran presentes en el corpus a emplear.

Se toma como base para el desarrollo de este trabajo los aportes que en el área se pueden encontrar en [119, 120, 86].

Por otra parte, pueden encontrarse en forma aislada y con diferentes grados de desarrollo, diferentes estudios relacionados con estas tres etapas en publicaciones recientes [193, 156, 157, 162]. Sin embargo, dista aún de ser propuesto un sistema que realice todas las asociaciones necesarias de forma automática y a partir únicamente de la señal de voz.

Como antecedentes del estudio de la prosodia en Español se pueden citar [141, 160, 108], y para el Español hablado en la Argentina [65, 63, 175].

1.5 OBJETIVOS DE LA TESIS

Para finalizar este capítulo se resumen los objetivos de la Tesis, que guían las etapas de su desarrollo.

Objetivos Generales:

1. Investigar la utilización de la información suprasegmental en el Reconocimiento Automático del Habla para el Español de Argentina
2. Evaluar nuevas alternativas de procesamiento y utilización de la informacióon prosódica en el RAH, fundamentadas en estudios de la percepción humana

Objetivos Específicos:

1. Estudiar la manifestación acústica de los atributos suprasegmentales en relación a fuentes de información léxica que se puedan emplear en el RAH
2. Extraer automáticamente rasgos acústicos distintivos que permitan caracterizar frases y/o palabras a partir de los atributos suprasegmentales y puedan ser incorporados en sistemas de RAH
3. Incorporar información suprasegmental en sistemas de RAH basados en HMM
4. Evaluar el aporte de esta fuente de información contrastando los sistemas propuestos con sistemas de RAH estándar

Finalmente, cabe destacar que la variante del Español Argentino tiene rasgos idiosincráticos que lo diferencian de las demás variantes de este idioma. Por ejemplo se han encontrado en el Español hablado en Buenos Aires rasgos típicos del Italiano, originados de la interacción entre estos idiomas durante corrientes inmigratorias de principios del siglo XX [26] que lo diferencian del resto de los países iberoamericanos.

La inexistencia de antecedentes en el uso de información suprasegmental en RAH para el Español de Argentina, supone que la presente tesis representará un aporte novedoso al área de reconocimiento automático del habla.

2 RAH BASADO EN MODELOS OCULTOS DE MARKOV

ÍNDICE

Prácticamente todos los sistemas de reconocimiento del habla actuales presentan arquitecturas basadas en modelos ocultos de Markov.

El predominio de los HMM en este campo se puede atribuir a múltiples razones, entre ellas: su sólida base matemática, la disponibilidad de un extenso cuerpo de investigación y experiencia en su aplicación a la tarea de reconocimiento del habla, su naturaleza estadística que permite realizar aprendizaje a partir de datos, minimizando la necesidad de conocimiento experto para el desarrollo de sistemas, la disponibilidad de algoritmos eficientes a la hora de construir y utilizar los reconocedores.

Los HMM son capaces de modelar las dos fuentes de incertidumbre que presenta la señal de habla, y que se encuentra en el corazón del problema de reconocimiento: las variabilidades espectrales y las variabilidades temporales. Asimismo permiten realizar la tarea de segmentación y clasificación de manera conjunta sobre la base de la teoría de decisión estadística, de forma tal que se minimiza la pérdida media de precisión por decisión.

A pesar de la adopción generalizada de esta metodología, también se pueden encontrar detractores que ven a las simplificaciones y restricciones que impone este modelo como uno de los principales factores que explican el menor desempeño de los sistemas de reconocimiento automático con respecto a los niveles de desempeño mostrado por los seres humanos.

Teniendo en cuenta que el objetivo de esta tesis es estudiar la utilización de información suprasegmental en sistemas de RAH estándar, este capítulo presenta los aspectos fundamentales del reconocimiento del habla basada en HMM.

2.1 ARQUITECTURA BÁSICA

Como se detalló en el capítulo 1, los sistemas de RAH basados en HMM siguen una aproximación estadística al reconocimiento del habla. En la figura 16 se muestra la arquitectura básica de un sistema de RAH basado en HMM.

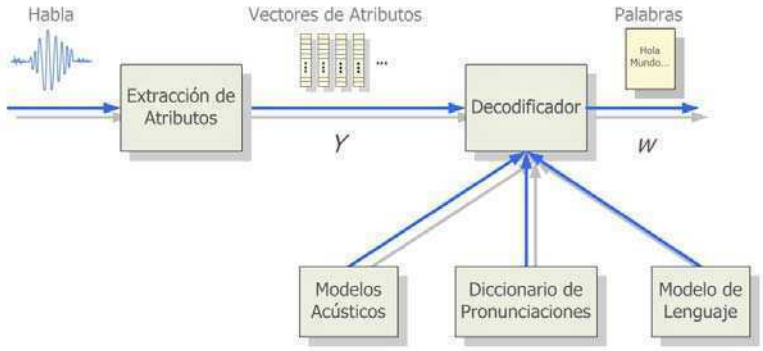

Figura 16: Esquema de un sistema de RAH basado en HMM. Adaptado de [51]

extracción de atributos

El módulo de extracción de atributos convierte la señal de habla digitalizada en una secuencia **Y** de vectores acústicos de tamaño fijo.

Sobre la base de dichos vectores de observaciones, el decodificador intenta determinar la secuencia de palabras **W** que puedan explicar mejor la secuencia de observaciones **Y**, o en otros términos, la secuencia de palabras más compatible con tal secuencia de observaciones. Para ello hace uso del *modelo acústico* y del *modelo de lenguaje*.

modelos acústicos

Los modelos acústicos más simples representan fonos. Si se construye un sistema de reconocimiento con dichas unidades acústicas, durante la fase de entrenamiento se utilizan fragmentos de habla con su correspondiente transcripción ortográfica para estimar el conjunto de parámetros de cada modelo de fonos del lenguaje.

Una vez entrenados los modelos, se puede construir el modelo acústico para cualquier palabra, concatenando los modelos de fonos que indique el *diccionario de pronunciaciones*.

modelos de lenguaje

El modelo de lenguaje comúnmente viene dado por N-gramas, aprendidos de grandes corpus de texto. En este tipo de modelos la probabilidad de cada palabra está condicionada por sus $N-1$ predecesoras.

decodificador

El decodificador obtiene la secuencia de palabras más probable realizando un proceso de búsqueda sobre todas las secuencias de palabras

posibles, haciendo uso por lo general de algún mecanismo de poda para excluir las hipótesis altamente improbables y conseguir un procesamiento más eficiente.

Cuando se llega al final de una frase, se puede obtener como resultado la secuencia de palabras más probables, o una red con una representación compacta de las hipótesis más probables.

A continuación se describe cada uno de los módulos anteriores con mayor detalle.

2.1.1 Extracción de Atributos

Este paso de procesamiento tiene como objetivo obtener una representación compacta de la señal de habla, minimizando la pérdida de información contenida en aquellos detalles que permiten discriminar entre palabras, y tratando de satisfacer los supuestos que imponen los modelos acústicos. Por ejemplo, si los modelos acústicos representan las distribuciones de estado-salida mediante una matriz de covarianza diagonal y gausiana, entonces los atributos deberían diseñarse para que tengan una distribución gausiana y no estén correlacionados.

La primera operación dentro del bloque de extracción de características consiste en el **preénfasis**. *preénfasis*

El espectro de magnitud de los sonidos sonoros presenta un declive de alrededor de 6 dB/octava. Esto se debe a la combinación de dos factores: por un lado una pendiente de -12 dB/octava que exhibe la fuente glótica, y por el otro un ascenso de $+6$ dB/octava debido al efecto de radiación generado por los labios.

La operación de preénfasis busca compensar dicha pendiente espectral, haciendo pasar a la señal por un filtro pasa-altos tipo FIR de primer orden (6 dB/octava) para ecualizar su espectro.

Dicho filtro generalmente tiene la forma:

$$H(z) = 1 - a\, z^{-1}, \quad 0{,}9 \leqslant a \leqslant 1{,}0 \tag{2.1}$$

Empleando la ecuación 2.1, la señal de salida del bloque de preénfasis ($\widetilde{s}(n)$) está relacionada con la señal de entrada ($s(n)$) mediante la siguiente expresión:

$$\widetilde{s}(n) = s(n) - a\, s(n-1) \tag{2.2}$$

El valor del coeficiente a empleado habitualmente es de 0,97.

Si bien de acuerdo a los argumentos anteriores solamente sería necesario aplicar la compensación a los segmentos sonoros, en la práctica se la aplica a toda la señal de entrada, debido a que no genera problemas para el análisis de regiones sordas.

Volvamos ahora a la parametrización de la señal de habla. Durante la locución, y producto de restricciones físicas, la configuración del tracto vocal suele modificarse de manera bastante lenta, y se la puede considerar constante en intervalos de alrededor de $10 - 20$ ms. Esto

por su parte hace posible considerar también a los sonidos del lenguaje estacionarios por tramos.

Para representar el habla de manera compacta la idea es codificar cada segmento estable de la señal con un solo conjunto de parámetros, asumiendo hipótesis de estacionaridad al momento de estimar los mismos. Obviamente se obtiene una mayor compactación si se emplea el segmento estable más grande posible. El tiempo máximo en que se puede considerar a esta señal estable viene determinado por la máxima velocidad con la que los articuladores pueden cambiar su configuración de manera significativa.

Esto también se puede pensar considerando que la duración máxima de ese intervalo de descomposición no debe superar la duración mínima de las unidades acústicas utilizadas, ya que de otro modo estas unidades podrían pasar desapercibidas durante el reconocimiento. Por lo tanto, desde el punto de vista de estacionaridad es conveniente que los segmentos sean cortos, para que las propiedades de interés de la señal de habla no se modifiquen.

Por otra parte es conveniente que los segmentos de análisis empleados sean suficientemente largos tal que permitan una estimación espectral confiable y de buena resolución en frecuencias, además de resultar conveniente desde el punto de vista de la generación de una representación compacta, como ya fue mencionado.

Para evitar artefactos en el cálculo de parámetros espectrales se aplica una función de ventaneo sobre la señal de habla, como se muestra en la figura 17. La aplicación de estas ventanas permite minimizar las discontinuidades de la señal al principio y final de cada trama y evitar así la aparición de espurios de alta frecuencia al realizar el análisis espectral.

Se puede apreciar que las ventanas de análisis consecutivas se solapan, ello permite medir correctamente las transiciones abruptas de la señal de habla.

Debido a que el análisis del habla se efectúa de forma sincrónica (a intervalos fijos de tiempo), durante el segmentos sonoros las ventanas de análisis deben ser suficientemente largas, de manera tal que la estimación espectral sea insensible a la posición relativa del ciclo glótico (se necesita al menos de un ciclo glótico completo en la parte central de la ventana).

De todo lo anterior, la elección de las duraciones para cada segmento de análisis debe surgir de una solución de compromiso entre las necesidades de compactación, estacionaridad, y confianza en la estimación de los parámetros.

En la práctica, las ventanas de análisis empleadas son de 25 a 30 ms de duración y los vectores de atributos suelen calcularse cada 10 ms. Cada vector de atributo caracteriza el tercio central de la ventana de análisis y a esos intervalos se conoce como trama o *frame*.

Una vez segmentada la señal de habla, se parametriza cada segmento empleando algún tipo de codificación que extraiga la información de la envolvente espectral de tiempo corto de dicha señal, de manera análoga a la operación que se lleva a cabo en la cóclea, como se explicó en el capítulo 1.

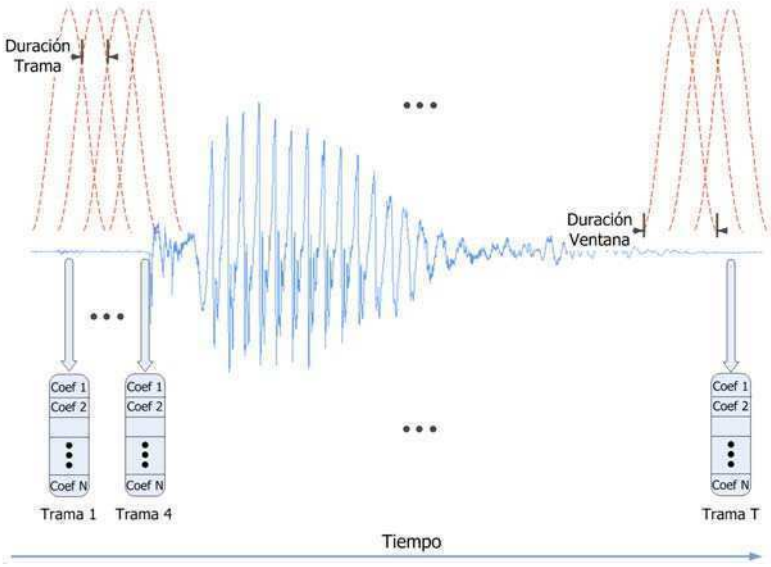

Figura 17: Parametrización de la señal de habla correspondiente a la sílaba *ta*. Se fragmenta la señal de habla empleando ventanas de Hamming solapadas y para cada una se obtienen sus parámetros espectrales de tiempo corto.

coeficientes Mel cepstrum

Uno de los esquemas de codificación más simples y populares está basado en coeficientes cepstrales en escala mel (conocidos como MFCC por sus iniciales en inglés) [32]. Para obtener los coeficientes mel-cepstrum del segmento de habla $s(t)$ correspondiente a una ventana de análisis, en primer lugar se obtiene la transformada discreta de fourier $s(\omega)$, posteriormente se utiliza un banco de filtros solapados y espaciados uniformemente sobre la escala de mel (escala no lineal que se corresponde con el comportamiento perceptual humano), para obtener los componentes de energía en cada banda de frecuencia $\widetilde{s}_k$, con $k = 1, 2, \ldots, K$. Finalmente los coeficientes MFCC $\widetilde{c}_n$ se obtienen aplicando la transformada discreta coseno (DCT) al logaritmo de $\widetilde{s}_k$:

$$\widetilde{c}_n = \sum_{k=1}^{K} (\log \widetilde{s}_k) \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right], \quad n=1,2,\ldots,L \tag{2.3}$$

Donde L es la longitud cepstral elegida.

coeficientes PLP

Otra técnica de codificación muy habitual es la predicción lineal perceptual (PLP) [72].

La idea de la técnica de PLP es adaptar la información espectral de tiempo corto de la señal de habla de acuerdo a algunas propiedades encontradas en el mecanismo de audición humano, y después aproximar el espectro resultante empleando un modelo autorregresivo de tipo todos polos.

Específicamente emplea tres conceptos psicofísicos de la audición para obtener una estimación del espectro auditivo:

- La resolución espectral en bandas críticas

- La curva de equi-sonoridad
- La ley de potencia de intensidad-sonoridad

El primer paso para obtener estos coeficientes consiste en realizar un análisis espectral de la señal de habla y transformarlo a escala Bark. El espectro de potencia obtenido se suaviza, convolucionándolo con el espectro de potencia de una curva de enmascaramiento que simula el comportamiento de las bandas críticas. Con esta operación se reduce su resolución espectral, permitiendo aplicarle un submuestreo que en la práctica está dado en intervalos de 1 Bark aproximadamente.

Posteriormente se aplica un filtro de pre-énfasis multiplicando el espectro de potencia de baja resolución con una curva que simula la curva de equi-sonoridad.

Finalmente se comprime en amplitud el espectro resultante del paso anterior, empleando una función de compresión con forma de raíz cúbica, lo que aproxima el comportamiento de ley de potencias auditiva reflejando una relación no lineal entre la intensidad del sonido y su sonoridad percibida. Los dos últimos pasos además de simular los comportamientos psicofísicos mencionados, reducen la variación espectral y de amplitud del espectro, y permiten que el modelo autorregresivo posterior pueda tener un orden bajo.

El espectro de baja resolución resultante se encuentra listo para un modelado de tipo todos polos: una transformada DFT inversa brinda una función de autocorrelación que conduce al cálculo final de los coeficientes de predicción de tal modelo y finalmente a su representación equivalente en término de coeficientes cepstrales.

En la práctica los coeficiente PLP pueden mejorar los resultados obtenidos con coeficientes MFCC, especialmente en ambientes con ruido. Para estos casos también se puede encontrar combinado el análisis PLP con un filtrado temporal de trayectorias cepstrales para obtener los denominados **parámetros RASTA** , una representación más robusta para el reconocimiento del habla bajo diversas condiciones ambientales [73].

parámetros RASTA

Además de los coeficientes espectrales se suele agregar los coeficientes de regresión de primero y segundo orden, denominados coeficientes delta y delta delta , o coeficientes de velocidad y aceleración respectivamente. Este recurso heurístico intenta compensar las suposiciones de independencia condicional empleada por los modelos acústicos. Si el vector de atributos original y estático es $\mathbf{y}_t^s$, entonces el parámetro delta $\Delta\mathbf{y}_t^s$ viene dado por:

coeficientes delta y delta delta

$$\Delta\mathbf{y}_t^s = \frac{\sum_{i=1}^{n} w_i \left(\mathbf{y}_{t+i}^s - \mathbf{y}_{t-i}^s\right)}{2\sum_{i=1}^{n} \mathbf{w}_i^2} \tag{2.4}$$

donde en este caso n es la longitud de la ventana y w_i son lo coeficientes de regresión.

Por su parte los coeficiente delta delta ($\Delta^2\mathbf{y}_t^s$) se obtienen de igual forma, pero empleando las diferencias de los coeficientes delta.

Finalmente los vectores de atributo $\mathbf{y}_t$ se obtienen concatenando todos los coeficientes mencionados:

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{y}_t^s \\ \Delta \mathbf{y}_t^s \\ \Delta^2 \mathbf{y}_t^s \end{bmatrix} \tag{2.5}$$

El resultado final es un vector de atributos con aproximadamente 40 coeficientes, parcialmente decorrelacionados.

2.1.2 Modelo Acústico

Los modelos acústicos desempeñan un papel primordial dentro de los sistemas de reconocimiento del habla. Tienen por función establecer una correspondencia en términos probabilísticos, entre una secuencia de vectores de atributos acústicos, de origen físico y observables, y las unidades elementales del habla, entidades de naturaleza abstracta.

Considere que para cada palabra de un lexicón se cuenta con un dispositivo o modelo capaz de generar los patrones de atributos que la representan. Cada vez que se activa uno de estos modelos produce un conjunto de vectores de características que representa una instancia de la palabra correspondiente. Si el modelo es suficientemente bueno, las estadísticas de un número muy grande de esos vectores serán similares a las estadísticas obtenidas de las pronunciaciones humanas para la misma palabra.

Supongamos por el momento que las palabras se pronuncian de manera aislada, tal que se conoce el principio y el final de cada una y la tarea de reconocimiento consiste solamente en determinar la identidad de la misma. Durante el reconocimiento se puede considerar que la mejor palabra candidata para un fragmento de habla es aquella cuyo modelo sea el más susceptible de generar la secuencia observada de vectores de atributos. Formalmente se intenta encontrar la palabra w que permite maximizar la probabilidad a posteriori $P(w|Y)$, es decir, la probabilidad que se haya pronunciado la palabra w dado que se observó el conjunto de atributos Y [79]:

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} \{P(\mathbf{w}|\mathbf{Y})\} \tag{2.6}$$

Es decir que se debe calcular la probabilidad a posteriori que un modelo dado haya generado un conjunto determinado de vectores de observaciones. Para ello no es necesario que el modelo produzca los vectores de atributos, sino usar las propiedades conocidas de cada modelo para los cálculos de probabilidades. En la práctica resulta complicado estimar el término de la derecha de la ecuación 2.6, sin embargo se puede hacer uso de la regla de Bayes para obtener una expresión equivalente:

$$P(w|\mathbf{Y}) = \frac{P(\mathbf{Y}|w) \cdot P(w)}{P(\mathbf{Y})} \tag{2.7}$$

La ecuación 2.7 indica que la probabilidad a posteriori de una palabra dadas las observaciones, es equivalente a la probabilidad de las observaciones dada la palabra, multiplicada por la probabilidad de la palabra, y dividida por la probabilidad de las observaciones.

Llevando la ecuación 2.7 a la ecuación 2.6, se puede ver que el término $P(\mathbf{Y})$ no depende de la palabra que se esté considerando como candidata, solamente actúa en la expresión como un factor de escala. Si el objetivo es encontrar la palabra w que maximice $P(w|\mathbf{Y})$ se puede ignorar dicho término sin afectar la maximización.

Así, es posible reescribir la ecuación 2.6 de una forma alternativa:

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmax}} \{P(\mathbf{Y}|\mathbf{w}) \cdot P(\mathbf{w})\} \tag{2.8}$$

La verosimilitud $P(\mathbf{Y}|\mathbf{w})$ se denomina *modelo acústico* y el término $P(\mathbf{w})$ *modelo de lenguaje*.

Ahora bien, considere que el modelo de cada palabra está compuesto por una serie de estados, y que en cada instante de tiempo se encuentra en uno de dichos estados posibles, cada uno de los cuales pueden estar asociados con una o más tramas de entrada. En general se asume que el modelo pasa de un estado a otro en intervalos regulares de tiempo, específicamente estos intervalos tienen igual duración que las tramas del análisis acústico. Además, como las palabras presentan variabilidad temporal, se admite que el modelo pueda permanecer en un mismo estado durante tramas sucesivas, y que pueda saltearse algún estado de la secuencia, permitiendo así modelar por ejemplo un alargamiento vocálico o pronunciaciones más rápidas respectivamente. La forma de este modelo se muestra en la figura 18.

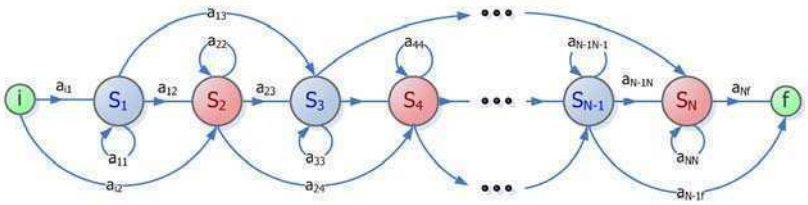

Figura 18: Grafo que representa las transiciones de estados para un modelo de palabra simple. Adaptado de [79]

probabilidad de emisión

Para cada vector de atributos proveniente del bloque de extracción de características acústicas, el modelo efectúa una transición de estados de manera probabilística, de acuerdo al conjunto de probabilidades de transición asociadas con el estado en que se encuentre en ese momento. Este es un primer proceso estocástico. Por otra parte, cada vez que un estado se activa emite una secuencia de vectores de atributos, cuyos valores están gobernados por una función probabilística denominada **probabilidad de emisión** , que conforma un segundo proceso estocástico. Es decir, el modelo descripto es doblemente estocastico: hay un proceso aleatorio dado por la emisión de vectores de

características, subyacente a otro proceso probabilístico, el de transición entre estados en el modelo.

En la práctica se recurre a ciertas simplificaciones que permiten hacer más tratable la matemática asociada con el modelo de la figura 18. Entre ellas se considera que las probabilidades de las distintos cursos de acción alternativos en un instante dado t dependen solamente del estado en que se encuentre el modelo en ese instante y no del valor particular de t. Así, la salida del modelo depende de la identidad del estado actual, pero es independiente de la secuencia de estados previos por los cuales ha atravesado el modelo antes de llegar a tal estado. De acuerdo a esas suposiciones se dice que el modelo opera como un proceso de Markov de primer orden, y que la secuencia de estados es una cadena de Markov de primer orden.

En este tipo de modelos el poder observar los vectores de atributos emitidos no basta para determinar la secuencia de estados que los generó. Por lo tanto, los valores reales de los atributos observados son funciones probabilísticas de los estados, y los estados en sí permanecen ocultos para el observador. De ahí que se los denomine modelos *"ocultos"* de Markov. En la siguiente sección se detallan los aspectos fundamentales de los HMM.

Hasta aquí se hizo la suposición que se de contaba con un modelo acústico por cada palabra del lexicón. Si bien ese suele ser el caso al construir sistemas de RAH de palabras aisladas, e incluso sistemas para habla conectada de vocabularios pequeños, la aplicación de esa estrategia para el reconocimiento de habla continua y de gran vocabulario no es conveniente ni viable.

Se pueden distinguir al menos dos desventajas al emplear modelos acústicos de palabras completas para el reconocimiento de habla continua [142]. En primer lugar, para obtener modelos confiables de palabras completas es necesario contar idealmente con varias realizaciones de cada palabra en los diversos contextos fonéticos que pueda aparecer. Debido a que las variabilidades acústicas a nivel de palabra se dan principalmente al comienzo y final de las mismas, para modelar esa variabilidad adecuadamente se hace necesario disponer de varias muestras de cada palabra con todas las variantes fonéticas como palabras vecinas precedentes y subsecuentes. Considere el problema de reconocimiento para un vocabulario de 1000 palabras, en el que cada una de ellas tiene en promedio 100 contextos fonéticos precedentes diferentes e igual cantidad de contextos fonéticos subsecuentes distintos. Para tener 10 ejemplares de cada una de esas variantes se necesitarían $10 \times 100\,1000 \times 100 = 100$ millones de secuencias de habla diseñadas cuidadosamente!

Claramente grabar y procesar esa cantidad de muestras es costoso e impráctico aún para tareas de reconocimiento dependiente del locutor.

En segundo lugar, emplear modelos de palabras completas es muy redundante. Aún cuando se contara con la cantidad de muestras comentada anteriormente, esta estrategia no aprovecharía al máximo la información contenida en los datos, ya que no tiene en cuenta que palabras distintas pueden compartir sonidos constituyentes similares. Esta es la razón por la cual los sistemas de RAH de gran vocabulario

en vez de emplear modelos de palabras completas, utilizan modelos de sub-palabras.

Se pueden utilizar varios tipos de unidades de sub-palabras diferentes como modelos acústicos:

- **Unidades Fonémicas:** se utiliza el conjunto básico de fonemas del idioma a reconocer, pero generalmente se introducen algunas modificaciones. Esas modificaciones se deben a que en la definición de unidades fonéticas se realiza un agrupamiento de elementos de acuerdo a similitudes lingüísticas, mientras que para el modelado acústico interesa agrupar las unidades por similitudes acústicas. Para el español se pueden considerar alrededor de 40 unidades fonémicas.
- **Unidades Silábicas:** nuevamente para definir el conjunto inicial de unidades se utiliza la definición lingüística de sílaba (núcleo vocálico más contextos inicial y final opcionales constituídos por una consonante, grupo consonántico o vocal en el caso de diptongos y triptongos), para luego modificar ese conjunto de acuerdo a agrupamientos por similitudes acústicas. En el español hay alrededor de 10000 sílabas.
- **Unidades Demisilábicas:** en este caso se considera unidades similares a las silábicas pero segmentadas a partir de algún punto del núcleo vocálico. Se pueden considerar alrededor de 2000 demisílabas en el español.
- **Unidades Acústicas:** se definen a partir de la segmentación del habla usando criterios objetivos de similitud acústica. Estas unidades pierden una relación clara con respecto a unidades lingüísticas. Para el inglés se encontró que un conjunto entre 256 y 512 unidades permiten modelar un rango amplio de vocabularios. Si bien este conjunto de unidades es interesante desde el punto de vista teórico, ya que es el que busca exactamente lo que pretende el modelo acústico: determinar cuáles son las distintas unidades a partir de rasgos acústicos, al estar desvinculado del nivel lingüístico hace dificil crear lexicones.

Todas las unidades alternativas que se mencionaron permiten representar cualquier palabra del lexicón, sin embargo cada una tiene sus ventajas y desventajas a la hora de emplearlas como unidades en el modelo acústico. El bajo número de unidades fonémicas distintas hacen sencillo entrenar y operar con estos modelos, sin embargo son extremadamente sensibles al contexto, es decir de la identidad de sus fonos precedentes y siguientes. En el otro extremo se encuentran las sílabas, que son las unidades de mayor duración y menor sensibilidad al contexto, sin embargo hay tantas sílabas diferentes que es prácticamente equivalente a trabajar con modelos de palabras completas. Otra gran ventaja de las unidades fonémicas es la sencillez con la que se las puede emplear para crear lexicones, o diccionarios de pronunciaciones.

Modelos de Probabilidades de Observaciones

Dependiendo de si las señales empleadas como entradas sean valuadas discretas o continuas, las probabilidades de observación para cada es-

tado de un HMM puede ser una función probabilidad de masa (fpm) valuada discreta, o una función densidad de probabilidad (fdp) valuada continua, respectivamente. Como se verá, los modelos discretos también se pueden utilizar para representar el espacio de un proceso valuado continuo, a través de la cuantización de este espacio en subespacios discretos.

Consideremos primero el modelo de densidad de observaciones discretas. Se pude asumir que existen M vectores de centroide discretos asociados con el estado S_i de un HMM: $[\mu_{i1}, \mu_{i2}, \cdots, \mu_{iM}]$, con una fpm $[P_{i1}, P_{i2}, \cdots, P_{iM}]$. Generalmente esos vectores de centroides al igual que sus probabilidades, se obtienen utilizando técnicas de agrupamiento sobre un conjunto de señales de entrenamiento asociadas con cada estado, como se detallará más adelante.

Si el número total de vectores de atributos válidos es muy grande, por ejemplo como sucede si se admiten muchos valores distintos para cada uno de tales atributos, es muy probable que una gran cantidad de tales vectores válidos no se encuentren dentro del conjunto de entrenamiento. En consecuencia, todas las probabilidades de observaciones asociadas a esos vectores de atributos serán nulas. Posteriormente si se observa alguno de esos vectores como entrada durante el funcionamiento del sistema, su reconocimiento podría resultar imposible.

Por otro lado, debido a las probabilidades de ocurrencia de los sonidos del habla se puede encontrar que algunos vectores espectrales son más frecuentes que otros, y por lo tanto el espacio multidimensional de características no se ocupa de manera uniforme.

Estas dos circunstancias hacen conveniente aplicar técnicas de **cuantización vectorial** para lograr una codificación más eficiente de las señales y evitar la falta de muestras de entrenamiento para algunos vectores de observaciones. La cuantización vectorial consiste en seleccionar del conjunto original de vectores válidos un subconjunto pequeño pero significativo, que a partir de ese momento serán los único valores que se admitan para las señales representadas. Posteriormente se reemplaza el valor original de cada vector de entrada por el más *parecido* de ese subconjunto.

cuantización vectorial

Ello permite aproximar el conjunto original de vectores de atributos por un número mucho menor de elementos, lo que hace posible entre otros beneficios: un entrenamiento más confiable de los modelos estadísticos, menores requerimiento de almacenamiento como de tamaño del cuerpo de datos para el entrenamiento, mayor eficiencia en algunas fases de procesamiento. En contraparte la aproximación implica pérdida de información e implica cálculos adicionales para llevar el vector original a la nueva representación.

Las técnicas de cuantización vectorial también se pueden utilizar para la representación de señales continuas mediante fpm.

La figura 19-a muestra una partición y cuantización del espacio de señales empleando 6 centroides.

Como se dijo, el proceso de cuantización vectorial supone una pérdida parcial de la información contenida en los datos originales. En muchas aplicaciones esa pérdida puede provocar una seria degradación en el desempeño de los reconocedores. Para evitar las limitaciones que

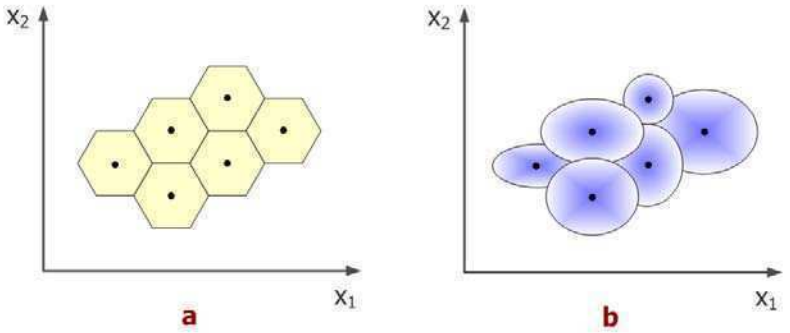

Figura 19: Modelado de un espacio de señales estocásticas empleando **(a)** una función de probabilidad de masas valuada discreta y **(b)** una función de densidad de probabilidad valuada continua compuesta por una mezcla de Gaussianas.

surgen de esas distribuciones discretas se puede utilizar alguna distribución paramétrica para codificar los atributos.

Al elegir qué tipo de distribuciones emplear en esa representación es conveniente elegir alguna que sea compatible con el comportamiento de las señales representadas. En muchos procesos naturales se observa que las magnitudes de las variables pueden ser aproximadas por la distribución normal o Gaussiana, cuyos parámetros independientes son el valor medio (μ) y el desvío estándar (σ). Empleando esa distribución, la función de densidad de probabilidad para la variable x viene dada por:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)} \tag{2.9}$$

Para el modelado de procesos valuados continuos, el espacio de señales asociado con cada estado es segmentado en un número de particiones. Si las señales dentro de cada agrupamiento se modelan con una distribución uniforme, entonces cada uno de esos grupos puede ser descripto por el vector de centroide y la distribución de probabilidad. En este caso los modelos de observaciones para cada estado consisten en M centroides de agrupamientos y la fdp asociada: $[\mu_{ik}, P_{ik}]$ con $i = 1, 2, \cdots, N$ y $k = 1, 2, \cdots, M$, como se muestra en la figura 19-b.

Aunque modelar las probabilidades de los atributos como funciones continuas ofrece ventajas significativas, para simplificar la explicación de los fundamentos de estos modelos se asumirán inicialmente sólo distribuciones de probabilidades discretas, y posteriormente se extenderán los resultados a observaciones continuas.

2.1.3 Modelo de Pronunciaciones

El modelo de pronunciaciones es una fuente de conocimiento que permite conectar las palabras del lexicón con los modelos acústicos. Éste modelo se encarga de definir, seleccionar y modelar la representación de símbolos que se pueden emplear para describir las realizaciones acústicas de las palabras.

Generalmente se especifica como un diccionario de pronunciaciones (por lo cual también se lo conoce con ese nombre). Está conformado por una lista de palabras seguidas por sus pronunciaciones aceptables, especificadas en términos del conjunto de modelos acústicos del reconocedor [44].

La definición de la pronunciación para una palabra en este diccionario tiene tres efectos: determina el rango de duraciones para una palabra en particular, determina el número de estados pseudo-estacionarios en la palabra, y permite asociar los estados de una palabra a estados de otras palabras que presentan una naturaleza acústica similar [69].

Como se vio durante la descripción del problema de reconocimiento del habla en el capítulo 1, cada palabra del lenguaje puede ser pronunciada de maneras alternativas. El empleo de modelos de pronunciaciones también ofrece la posibilidad de representar la variabilidades encontrada en el interior de las palabras. En este caso, las variabilidades de pronunciación simplemente se agregan al diccionario.

De acuerdo a la fuente empleada para determinar cuáles son las pronunciaciones alternativas, se puede hacer una distinción entre métodos basados en datos y aquellos basados en conocimiento [164]. A diferencia de la aproximación basada en conocimiento donde se asume que la información de las pronunciaciones alternativas ya está disponible por ejemplo en a partir de estudios lingüísticos efectuados sobre una población determinada, en la basada en datos la idea es obtener dicha información directamente de las señales, requiriendo generalmente de una etapa previa de etiquetado fonético de tales datos. Ese etiquetado fonético se puede obtener de forma manual o semi-automática (a través de un reconocedor de fonemas y alineamiento forzado)[90].

Se debe considerar que el material bibliográfico empleado por los métodos basados en conocimiento generalmente brindan la información de variaciones fonéticas pero sin detallar datos cuantitativos sobre la incidencia de cada variante en una población, por lo que esa información finalmente se debe obtener a partir de algún cuerpo de datos acústicos.

Por su parte, si se considera la forma de extraer las variantes de pronunciaciones a partir de las fuentes mencionadas, se pueden distinguir entre métodos manuales y automáticos. Entre las alternativas para la extracción automática de estas variantes se han empleado reglas [151], redes neuronales [50], conversores de grafemas a fonemas [16], optimización empleando el criterio de máxima verosimilitud [80], árboles de decisiones [147].

La razón por la cual se usan múltiples pronunciaciones alternativas para una misma palabra en este modelo es aumentar las chances que el reconocedor seleccione como hipótesis a la palabra correcta. De esta forma se brinda al sistema un modelo de lo que debe buscar, más

ajustado a lo que acústicamente puede observar. Consecuentemente, se busca que esta descripción más rica acústicamente conduzca a disminuciones en las tasas de error de reconocimiento. Sin embargo, al agregar variantes de pronunciación en el lexicón también se están introduciendo posibles nuevos errores; ya que aumenta la confusión acústica. Al agregar más cadenas acústicas para una palabra, aumentan las chances que se encuentren otras palabras en el lexicón con una secuencia de sonidos parecida. Esto se puede minimizar haciendo una selección adecuada de las variantes, tal que la reducción de errores ya existentes sea mayor que el número de nuevos errores ocasionados. Se pueden encontrar diferentes criterios para determinar qué conjunto de variantes maximizan las ganancias de desempeño, por ejemplo: frecuencia de aparición de la variante, medidas de máxima verosimilitud, grado de similitud fonética entre las variantes agregadas y las existentes.

Pero además de las variabilidades intrapalabras el habla continua presenta variabilidades de pronunciación entre palabras. Una forma simple de modelar este tipo de variabilidad y seguir empleado la arquitectura ordinaria de reconocedores es introducir en el diccionario de pronunciaciones palabras múltiples, y tratarlas como cualquier palabra convencional. De esta manera las variaciones en las interfases de las palabras múltiples se pueden incluir explícitamente en el diccionario de pronunciaciones.

Sin embargo esta aproximación es capaz de contemplar una pequeña proporción de la variabilidad referida, ya que resulta eficiente en caso que se agreguen al diccionario un número reducido de palabras múltiples. Generalmente se suele optar por agregar solamente las palabras múltiples cuyas ocurrencias sean más frecuentes. También se han propuesto otros métodos para considerar la variabilidad de pronunciación entre palabras como [8, 118, 35].

2.1.4 Modelo de lenguaje

El desempeño de los sistemas de reconocimiento automático del habla de gran vocabulario depende de manera crítica del conocimiento lingüístico embebido mediante modelos de lenguaje.

Los modelos de lenguaje estadísticos tienen por objetivo brindar una estimación de la probabilidad de secuencias de palabras W para una determinada tarea de reconocimiento. Si se asume que W se especifica mediante una secuencia de la forma:

$$W = w_1\, w_2\, \cdots\, w_q \tag{2.10}$$

Entonces puede parecer razonable que P(W) se pueda calcular mediante la expresión:

$$\begin{aligned} P(W) &= P(w_1\, w_2\, \cdots\, w_q) \\ &= P(w_1)\, P(w_2 \mid w_1)\, P(w_3 \mid w_1\, w_2)\, \cdots \\ &\quad \cdots\, P(w_q \mid w_1\, w_2\, \cdots\, w_{q-1}) \end{aligned} \tag{2.11}$$

Desafortunadamente, resultaría imposible estimar de manera confiable las probabilidades condicionales de la ecuación 2.12 para todas las

palabras y longitudes de secuencias de un lenguaje dado. En la práctica es conveniente emplear un modelo de palabras de N-gramas, en el que se aproxime

$$P(w_j|w_1w_2\ldots w_{j-1}) \;=\; P(w_j|w_{j-N+1}\;\cdots\;w_{j-1}) \tag{2.12}$$

es decir, basado en las N-1 palabras precedentes.

Aún con esa simplificación, las probabilidades de N-gramas de la ecuación 2.12 son difíciles de estimar de manera confiable para un valor de $N > 3$.

En la práctica, a menudo resulta conveniente usar modelos de pares de palabras que especifican cuáles de estos son válidos en el lenguaje empleando funciones binarias:

$$P(w_j|w_k) = \begin{cases} 1 & \text{si } w_kw_j \text{ es válido} \\ 0 & \text{en otro caso} \end{cases} \tag{2.13}$$

Dentro de las gramáticas estadísticas se pueden encontrar los N-gramas de clases de palabras.

También existen modelos de lenguaje alternativos a los estadísticos que incluye a los modelos de gramáticas formales (gramáticas dependientes o independientes del contexto). Estos últimos tienen una forma de procesamiento de los constituyentes del lenguaje natural más adecuada que las N-gramas de palabras. Sin embargo resultan más dificil de integrar dentro de la decodificación acústica de los HMM.

Modelado Estadístico del Lenguaje

En el modelado estadístico del lenguaje se emplean grandes corpus de texto para estimar P(W). Por razones prácticas, la probabilidad de secuencias de palabras P(W) se aproxima de la siguiente manera:

$$P_N(W) \;=\; \prod_{i=1}^{Q} P(w_i|w_{i-1}, w_{i-2}, \cdots, w_{i-N+1}) \tag{2.14}$$

A la ecuación 2.14 se conoce como modelo de lenguaje de N-gramas. Las probabilidades condicionales $w_i|w_{i-1}, w_{i-2}, \cdots, w_{i-N+1}$ se pueden estimar empleando las frecuencias relativas encontradas en un corpus de texto suficientemente extenso:

$$\hat{P}(w_i|w_{i-1}, \cdots, w_{i-N+1}) = \frac{F(w_i, w_{i-1}, \cdots, w_{i-N+1})}{F(w_{i-1}, \cdots, w_{i-N+1})} \tag{2.15}$$

En la ecuación 2.15 F es el número de ocurrencias de la secuencia de palabras que aparece en el argumento, dentro del corpus de entrenamiento empleado. Por otra parte, para que la estimación de 2.15 sea confiable $F(w_i, w_{i-1}, \cdots, w_{i-N+1})$ tiene que registrarse de manera sustancial dentro de dicho corpus. Para que todas las secuencias posibles dentro de un lenguaje sean registradas dentro de un corpus, este debería ser prohibitivamente grande. En la práctica existe un límite al tamaño de los mismos, lo que implica que existan secuencias de palabras para las que no se encuentren ocurrencias dentro del mismo ($F(w_i, w_{i-1}, \cdots, w_{i-N+1}) = 0$).

suavizado de N-gramas

Una forma de tratar este problema es realizando un **suavizado de las frecuencias de N-gramas** , propuesto originalmente en [85].

Existen varias alternativas para realizar el suavizado de N-gramas. Considerando un modelo de trigramas (N = 3), el suavizado se efectúa interpolando las frecuencias relativas de trigramas, bigramas y unigramas:

$$\hat{P}(w_3|w_1,w_2) = p_1 \times \frac{F(w_1,w_2,w_3)}{F(w_1,w_2)} + p_2 \times \frac{F(w_1,w_2)}{F(w_1)} + p_3 \times \frac{F(w_1)}{\sum F(w_i)} \tag{2.16}$$

En 2.16 los coeficientes p_i son no negativos y satisfacen la restricción: $p_1 + p_2 + p_3 = 1$ y $\sum F(w_i)$ es igual al tamaño del corpus. Los pesos p dependen de los valores de $F(w_1,w_2)$ y $F(w_1)$ y se pueden obtener aplicando el principio de validación cruzada [85].

Perplejidad del Modelo de Lenguaje

Una vez construido el modelo de lenguaje a partir del corpus de entrenamiento, se puede emplear el concepto de fuente de información para tratar de establecer cuál es el desempeño esperado del modelo en la tarea de reconocimiento.

Además del concepto de fuente de información, para establecer esa medida de desempeño se deben emplear diversos conceptos del área de teoría de la información como: entropía, entropía estimada y perplejidad.

entropía de la fuente

Consideremos una fuente de información que produce secuencias de palabras (símbolos) $w_1, w_2, \cdots, w_Q$, cada una de las cuales escogidas de un vocabulario $\bar{V}$ cuyo tamaño es $|\bar{V}|$, de acuerdo a alguna ley estocástica. Se puede definir la **entropía de la fuente** como:

$$H = -\lim_{Q\to\infty}\left(\frac{1}{Q}\right)\left\{\sum P(w_1,w_2,\cdots,w_Q)\times\log\left(P(w_1,w_2,\cdots,w_Q)\right)\right\} \tag{2.17}$$

En la ecuación 2.17 $P()$ es la probabilidad que la fuente genere la secuencia que aparece como argumento, dados la ley estocástica de emisión de símbolos, y la sumatoria es sobre todo el conjunto de secuencias $w_1, w_2, \cdots, w_Q$ posibles. Si las palabras en la secuencia de símbolos son generadas por la fuente de manera independiente, entonces:

$$P(w_1,w_2,\cdots,w_Q) = P(w_1)\cdot P(w_2)\cdot\cdots\cdot P(w_Q)) \tag{2.18}$$

y la ecuación 2.17 se puede reescribir como:

$$H = -\sum_{w\in\hat{V}} P(w)\cdot\log\left(P(w)\right) \tag{2.19}$$

A la ecuación 2.19 se suele denominar como entropía de primer orden de la fuente. La cantidad H en la ecuación 2.17 se puede considerar como la información promedio de la fuente cuando genera una palabra w. De manera equivalente, una fuente de entropía H presenta un contenido de información similar a una fuente genérica que produce palabras de manera equiprobables seleccionadas de un vocabulario de 2^H elementos.

Si la fuente es ergódica, sus propiedades estadísticas se pueden caracterizar completamente mediante una secuencia suficientemente larga que ésta produzca. En esas condiciones la entropía de la ecuación 2.17 es equivalente a:

$$H = -\lim_{Q\to\infty}\left(\frac{1}{Q}\right)\log\left(P(w_1, w_2, \ldots, w_Q)\right) \tag{2.20}$$

En otras palabras, se puede calcular la entropía a partir de secuencias típicamente largas de palabras generadas por la fuente. La longitud de esta secuencia (el corpus) idealmente debe aproximarse a infinito, lo que obviamente es imposible. Generalmente se calcula H a partir de una secuencia finita pero suficientemente larga:

$$H = -\left(\frac{1}{Q}\right)\log\left(P(w_1, w_2, \cdots, w_Q)\right) \tag{2.21}$$

Una interpretación interesante de H desde la óptica de reconocimiento del habla es considerarla como el grado de dificultad promedio que el reconocedor encuentra cuando va a determinar una palabra de la misma fuente. Esta dificultad o incertidumbre está basada en la probabilidad $P(w_1, w_2, \ldots, w_Q)$ la cual es usualmente desconocida a priori para los lenguajes naturales, y por lo tanto debe ser estimada.

Una forma de estimar H es empleando $P(W) = P(w_1, w_2, \cdots, w_Q)$ a partir del modelo del lenguaje. Por ejemplo si se emplea el modelo de lenguaje de N-gramas $P_N(W)$ de la ecuación 2.14 se puede reescribir la ecuación 2.21 para estimar H de la siguiente manera:

$$H = -\left(\frac{1}{Q}\right)\sum_{Q}^{i=1}\log\left(P(w_i|w_{i-1}, w_{i-2}, \cdots, w_{i-N+1})\right) \tag{2.22}$$

En general,

$$H = -\left(\frac{1}{Q}\right)\sum_{Q}^{i=1}\log\left(\hat{P}(w_1, w_2, \cdots, w_Q)\right) \tag{2.23}$$

donde $\hat{P}(w_1, w_2, \cdots, w_Q)$ es una estimación de $P(w_1, w_2, \cdots, w_Q)$. La cantidad H_p es una entropía estimada, calculada da partir de una secuencia suficientemente larga basada en un modelo de lenguaje.

Si la fuente es ergódica y $Q \to \infty$ entonces $H_p \geqslant H$. Intuitivamente esto se puede verificar de manera simple por el hecho que el conocimiento de la probabilidad real $P(w_1, w_2, \ldots, w_Q)$ es lo mejor que

puede usar un reconocedor y ninguna otra estimación de las probabilidades de un modelo de lenguaje puede hacer la tarea de reconocimiento más sencilla. Como H_p es una indicación de la dificultad de reconocimiento, acotada inferiormente por H, un modelo de lenguaje que logre un H_p menor (más cercano a H) es considerado un mejor modelo que otro modelo de lenguaje que conduzca a un H_p mayor.

perplejidad

Asociado con H_p existe una cantidad denominada **perplejidad** (a menudo denominada factor de ramificación de palabra promedio del modelo de lenguaje), definido como:

$$B = 2^{H_p} = \hat{P}(w_1, w_2, \cdots, w_Q)^{-1/Q} \tag{2.24}$$

Se debe notar que H_p es la dificultad o incertidumbre promedio de cada palabra, en base a un modelo de lenguaje. Cuando el reconocedor utiliza este modelo de lenguaje para reconocer el habla, la dificultad a la cual se enfrenta es equivalente a la de reconocer un texto generado por una fuente que elige palabras de un vocabulario de tamaño B independientemente de todas las demás palabras que son equiprobables.

Otra forma de ver la perplejidad es considerarla como el número promedio de posibles palabras que siguen a cualquier cadena de $(N-1)$ palabras en un corpus grande basado en un modelo de lenguajes de N-gramas.

La perplejidad es un parámetro importante en la especificación del grado de sofistificación en una tarea de reconocimiento, a partir de la incertidumbre de la fuente hasta la calidad del modelo de lenguaje.

2.2 MODELOS OCULTOS DE MARKOV

Como se comentó en la sección previa, los modelos acústicos de los sistemas de RAH actuales están basados en modelos ocultos de Markov. A continuación se detallarán los principales aspectos teóricos de estos elementos y su empleo en el reconocimiento del habla.

2.2.1 Definición

Un HMM (λ) es un autómata estocástico que se define mediante el siguiente conjunto de parámetros:

- **N**, *número de estados del modelo*.

  Si bien cada uno de esos estados $\mathbf{S} = \{S_1, S_2, \ldots, S_N\}$ son ocultos, generalmente se les puede hacer corresponder un significado físico. Por otra parte se designa al estado ocupado por el modelo en el tiempo t como q_t

- **M**, *número de posibles símbolos de observaciones por estado*.

  Estos símbolos corresponden a las salidas físicas del sistema que se está modelando, y se designan como: $\mathbf{Y} = \{y_1, y_2, \cdots, y_M\}$.

- $\mathbf{A} = \{a_{ij}\}$, *distribución de probabilidades de transición entre estados*.

  Considerando que el modelo se encuentra en el estado *i* durante el instante t, a_{ij} es la probabilidad que pase al estado *j* en el instante siguiente, de manera formal:

$$a_{ij} = P\left[q_{t+1} = S_j \mid q_t = S_i\right], \quad 1 \leqslant i, j \leqslant N \tag{2.25}$$

  Al definir estas probabilidades se asumen procesos de primer orden e invariantes en el tiempo, lo que implica que la probabilidad de alcanzar un estado determinado depende solamente del estado previo y que estas probabilidades son independientes del tiempo t:

$$\begin{aligned} a_{ij} &= P\left[q_{t+1} = S_j \mid q_t = S_i\right] \\ &= P\left[q_{t+\tau+1} = S_j \mid q_{t+\tau} = S_i\right] \\ &= P\left[q_{t+1} = S_j \mid q_t = S_i, q_{t-1} = S_k, \ldots\right], \\ &\quad 1 \leqslant i, j \leqslant N; \ 1 \leqslant t \leqslant T \end{aligned} \tag{2.26}$$

  Esos supuestos son necesarios para hacer a los modelos tratables computacionalmente, sin embargo no se ajustan a lo observado en el habla continua.

  Por otro lado, como las probabilidades de transición obedecen restricciones estocásticas estándar, se cumplen las siguientes propiedades:

$$a_{ij} \geqslant 0 \quad \forall\, i, j \tag{2.27}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \forall\, i \tag{2.28}$$

- $\mathbf{B} = \{b_j(k)\}$, *distribución de probabilidades de observarciones*.

  Donde $b_j(k)$ es la probabilidad de emisión del símbolo y_k al activarse el estado S_j en el instante t. La expresión matemática de esta probabilidad es:

$$b_j(k) = P\left[y_k \text{ en } t | q_t = S_j\right], \quad 1 \leqslant j \leqslant N; \ 1 \leqslant k \leqslant M \tag{2.29}$$

  Se puede notar que la ecuación 2.29 asume independencia entre las observaciones subsecuentes, lo que no sucede en el caso del habla.

- $\boldsymbol{\pi} = \pi_i$, *distribuciones de probabilidades para los estados iniciales del modelo*.

  Donde π_i denota la probabilidad que el estado S_i se active al inicio de la secuencia:

$$\pi_i = P[\, q_1 = S_i \,], \quad 1 \leqslant i \leqslant N \tag{2.30}$$

  Al igual que en la ecuación 2.28, en este caso se satisface:

$$\sum_{i=1}^{N} \pi_i = 1 \tag{2.31}$$

Por simplicidad se suele usar una notación compacta para indicar el conjunto completo de parámetros que definen a un HMM:

$$\lambda = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}) \tag{2.32}$$

Como se comentó al introducir los modelos acústicos, los HMM definen un proceso doblemente estocástico. Por un lado un procedimiento aleatorio discreto que genera para cada instante t una secuencia de símbolos de estados $q_t \in \mathbf{S}$ a partir de las distribuciones de probabilidades dadas por $\mathbf{A}$ y $\boldsymbol{\pi}$.

El otro proceso estocástico viene dado por la emisión de un símbolo u observación y_k cada vez que se activa un estado. Las probabilidades de emisión son específicas para cada estado y obedecen las probabilidades expresadas en $\mathbf{B}$.

HMM discretos
HMM continuos

Como se adelantó en la primera sección de este capítulo, esas observaciones pueden ser discretas, en cuyo caso los modelos se se denominan **HMM discretos** , o pueden ser valuadas continuas, denominándose en este caso **HMM continuos** . En este último caso generalmente se utilizan mezclas de distribuciones normales como se definió en la ecuación 2.9.

HMM semi-continuos

Además de esos dos modelos, otra estrategia consiste en utilizar un híbrido de los modelos anteriores, denominados **HMM semi-continuos** , que se pueden ver como HMM continuos que utilizan un conjunto de L distribuciones independientes de cualquier estado. Estas distribuciones se colocan en una bolsa común y cada estado modela su función densidad de probabilidad como la combinación de las L distribuciones. Solamente deben especificar el coeficiente de ponderación para cada una de las mezclas.

2.2.2 Problemas Básicos de los HMM

Para que los HMM sean útiles en la práctica se deben resolver tres problemas básicos:

1. **Problema de Evaluación**

   Dada la secuencia de observaciones ($\mathbf{Y} = y_1, y_2, \cdots, y_T$) y el modelo ($\lambda$), este problema consiste en determinar $P(\mathbf{Y}|\lambda)$, la probabilidad que la secuencia observada sea generada por ese modelo. A este problema también se lo puede denominar como problema de clasificación o reconocimiento, y ver de la siguiente forma: dados varios modelos competidores y una secuencia de observaciones, cómo elegir el modelo más compatible con esas observaciones.

2. **Problema de Estimación**

   Dada la secuencia de observaciones ($\mathbf{Y}$) el problema radica en determinar los parámetros del modelo ($\lambda = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$), que maximicen la probabilidad $P(\mathbf{Y}|\lambda)$. La resolución de este problema permite establecer un método para optimizar los parámetros del modelo tal que conduzca a una mejor descripción de los procesos físicos observados.

3. **Problema de Decodificación**

   Dada la secuencia de observaciones ($\mathbf{Y}$) y un modelo (λ) este problema consiste en determinar la secuencia de estados $\mathbf{q} = q_1, q_2, \cdots, q_T$ que explique mejor las observaciones empleando algún criterio de optimalidad. Este problema se vincula con la recuperación de la parte *oculta* del modelo.

Como se verá en la sección siguiente, en el marco de construcción y utilización de HMM en el reconocimiento del habla estos tres problemas se resuelven de manera eficiente bajo un mismo esquema probabilístico.

2.3 RECONOCIMIENTO DE HABLA CON HMM

Una vez definidos formalmente los HMM, veamos cómo se los utiliza en el reconocimiento del habla.

En la figura 20 se muestra la forma que tienen los HMM empleados en el RAH. Estos HMM se denominan modelos de izquierda a derecha

modelos de izquierda a derecha

y satisfacen las siguientes restricciones:

$$a_{ij} = 0 \quad \forall\ j < i \tag{2.33}$$

$$\pi_i = \begin{cases} 1 & \text{si } i = I \\ 0 & \text{en otro caso} \end{cases} \tag{2.34}$$

Los modelos como los mostrados en la figura además presentan dos estados especiales: un estado inicial *I*, y un estado final *F* que no están asociados con la emisión de ningún vector de atributos, por los que se los denomina también estados no emisores, y poseen un conjunto restringido de probabilidad de transiciones.

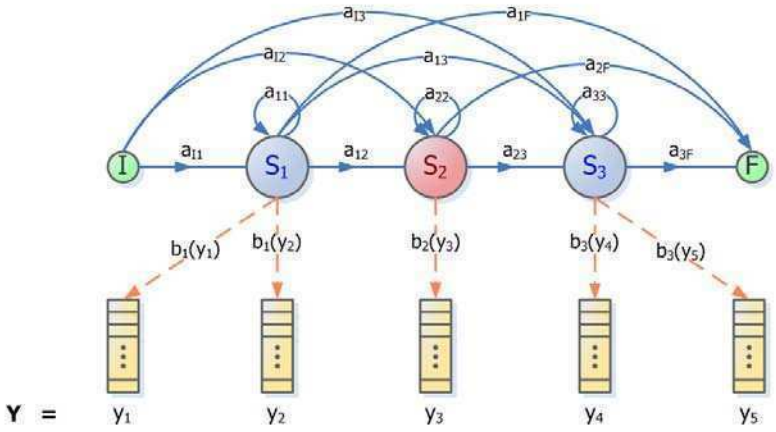

Figura 20: Esquema de HMM de izquierda a derecha, empleados en el RAH. Se pueden observar dos estados especiales: el inicial (I, y el final (F), que a diferencia de los demás son estados que no emiten vectores de observaciones.

El estado inicial se utiliza para especificar las probabilidades de transición desde el inicio del modelo a todos los posibles estados iniciales con emisión, mientras que el estado final contempla las probabilidades de transición desde cualquier posible estado final con emisión hasta el final del modelo. En consecuencia cada modelo debe comenzar en el estado I y terminar en el F, por lo que para generar T observaciones el modelo debe pasar por $T + 2$ estados.

La utilización de esos estados no emisores es un método conveniente para modelar el hecho que hay modelos más susceptibles que otros de estar asociados con la primera y última trama de una secuencia. Además facilitan la concatenación de modelos como se verá más adelante.

Para los cálculos asociados con los HMM se suelen utilizar esquemas de reticulados (*trellis* en la literatura anglosajona) como el presentado en la figura 21. Las retículas de estados y tiempos como la presentada en la figura 21 permiten observar todos los estados que componen un HMM y al mismo tiempo determinar todos los caminos posibles para las secuencias de estados al transcurrir el tiempo.

Debido a que para cada HMM tanto el conjunto de estados como los parámetros de cada uno son invariantes en el tiempo, la retícula presenta una estructura regular y repetitiva. Se puede ver además que para los HMM de izquierda a derecha como el de la figura 20 el reticulado debe comenzar divergiendo del primer estado y terminar por convergir en el último. Sin embargo, en el trellis que representa esa situación hay muchas secuencias de estados posibles. Cada una de esas secuencias de estados tiene su probabilidad a priori, que se puede obtener multiplicando las probabilidades de transición de los estados que la componen. Por ejemplo, para la secuencia de estados $\mathbf{q} = q_I, q_1, q_2, q_2, q_3, q_3, q_F$ la probabilidad es $P = \pi_1 a_{11} a_{12} a_{22} a_{23} a_{33} a_{3F}$. Además como cada estado en general tiene

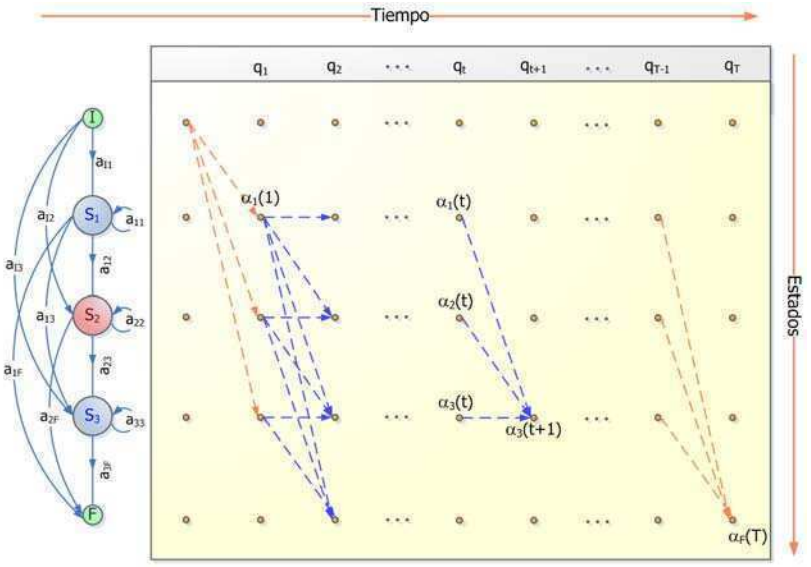

Figura 21: Retícula empleada en los cálculos de distintos algoritmos aplicados sobre HMM. La retícula ilustrada corresponde al modelo λ de la figura 20.

un conjunto de probabilidades de emisiones diferente, las distintas secuencias de estados modelarán diferentes secuencias de observaciones.

Para un modelo λ, una secuencia de observaciones $\mathbf{Y}$ pudo haber sido generada por cualquiera de las secuencias de estados posibles representada en su retícula, con una probabilidad igual a $P(\mathbf{Y}|\lambda)$.

2.3.1 Algoritmos Básicos para HMM

A continuación se presentarán los algoritmos característicos para el RAH empleando modelos ocultos de Markov. Cada uno de ellos permite resolver alguno de los problema básicos asociados con los HMM.

Algoritmo de Avance-Retroceso

Consideremos el **problema de evaluación** presentado en la sección anterior. Se puede reescribir la probabilidad que interesa encontrar de la siguiente manera:

$$P(\mathbf{Y}|\lambda) = \sum_{\forall \mathbf{q}} P(\mathbf{Y}, \mathbf{q}|\lambda) = \sum_{\forall \mathbf{q}} P(\mathbf{Y}|\mathbf{q}, \lambda) \cdot P(\mathbf{q}|\lambda) \tag{2.35}$$

Donde se ha utilizado la propiedad para las probabilidades condicionales: $P(A, B) = P(A|B) \cdot P(B)$.

Para el primer término de la última igualdad en la ecuación 2.35, si se asume que los vectores de atributos se generaron de manera independiente para cada estado, la probabilidad de las observaciones dada una secuencia de estados particular de duración T es el producto

de las probabilidades de emisión individuales para cada estado de la secuencia especificada:

$$P(\mathbf{Y}|\mathbf{q},\lambda) = \prod_{t=1}^{T} b_{q_t}(y_t) \tag{2.36}$$

Por otra parte, considerando el segundo término de la ecuación 2.35, dado un modelo λ, la probabilidad de una secuencia de estados en particular ($\mathbf{q} = \{q_1, q_2, \cdots, q_T\}$) está dado por la productoria de las probabilidades de transición:

$$P(\mathbf{q}|\lambda) = a_{Iq_1} \cdot \left(\prod_{t=1}^{T-1} a_{q_t q_{t+1}} \right) \cdot a_{q_T F} \tag{2.37}$$

Utilizando las ecuaciones 2.36 y 2.37 se puede reescribir la ecuación 2.35 de la siguiente manera:

$$P(\mathbf{Y}|\lambda) = \sum_{\forall \mathbf{q}} a_{Iq_1} \cdot \left(\prod_{t=1}^{T-1} b_{q_t}(y_t) \cdot a_{q_t q_{t+1}} \right) \cdot b_{q_T}(y_t) \cdot a_{q_T F} \tag{2.38}$$

La forma más sencilla de obtener $P(\mathbf{Y}|\lambda)$ consiste en enumerar cada posible secuencia de estados válida para el modelo λ que tengan longitud T (el número de observaciones que contiene $\mathbf{Y}$), y calcular para cada una el término que aparece dentro de la sumatoria de la ecuación 2.38. Finalmente el resultado de la probabilidad buscada se obtiene sumando cada uno de dichos términos.

HMM egódicos

En el caso de **HMM egódicos** , es decir aquellos en los que ninguno de los elementos de la matriz de probabilidades de transición **A** sean nulos, el cálculo directo de las probabilidades de la ecuación 2.38 tiene un orden de $O(N^T)$. Involucra alrededor de $2T \cdot N^T$ operaciones, ya que en cada ($t = 1, 2, \cdots, T$), hay N estados alcanzables (por lo tanto N^T posibles secuencias de estados); y para cada una de esas secuencias de estados se requieren alrededor de $2T$ operaciones para cada término del sumando en la ecuación 2.38. Para $N = 5$ estados, $T = 100$ observaciones, se deben calcular en el orden de $2 \cdot 100 \cdot 5^{100} = 10^{72}$ operaciones! Obviamente, la aplicación directa de la ecuación 2.38 es inviable.

Sin embargo se pueden calcular las probabilidades indirectamente utilizando una relación de recurrencia. A ese procedimiento se denomina **algoritmo de avance-retroceso** (algoritmo *forward-backward* en inglés) .

algoritmo de avance-retroceso

variable hacia adelante

Para este algoritmo se define el valor $\alpha_j(t)$ denominada **variable hacia adelante**, para indicar la probabilidad que el modelo haya producido los primeros t vectores de atributos observados, y al mismo tiempo de encontrarse en el estado S_j al finalizar esa secuencia, es decir, en el instante t:

$$\alpha_j(t) = P(y_1, y_2, \cdot, y_t, q_t = S_i|\lambda) \tag{2.39}$$

Se puede calcular $P(\mathbf{Y}|\lambda)$ contenida en la ecuación 2.35 empleando la variable $\alpha_j(t)$ de manera inductiva mediante el algoritmo de avance, presentado en la figura 22.

Algoritmo de Avance

- INICIALIZACIÓN:

$$\alpha_j(1) = a_{Ij} \cdot b_j(y_1) \tag{2.40}$$

- ITERACIÓN:

  Calcular iterativamente todas las demás variables empleando la expresión de recurrencia en términos de los valores $\alpha_i(t-1)$, para todos los estados previos posibles i:

$$\begin{aligned} \alpha_j(t) &= P\left(y_1, y_2, \cdots y_t, q_t = S_j|\lambda\right) \\ &= \left(\sum_{i=1}^{N} \alpha_i(t-1) \cdot a_{ij}\right) \cdot b_j(y_t) \quad \text{para } 1 < t \leqslant T \end{aligned} \tag{2.41}$$

  Aquí nuevamente se asume independencia entre observaciones subsecuentes.

- FINALIZACIÓN:

  Finalizar el proceso calculando:

$$P(y_1, y_2, \cdots, y_T|\lambda) = \alpha_F(T) = \sum_{i=1}^{N} \alpha_i(T) \cdot a_{iF} \tag{2.42}$$

Figura 22: Algoritmo de Avance aplicado al problema de evaluación.

Este algoritmo demanda $N(N+1)(T-1)+N$ multiplicaciones (orden $O(N^2T)$). Para el mismo ejemplo que el visto previamente ($N = 5$ y $T = 100$), aplicar este algoritmo supondrían 3000 operaciones elementales, un valor 69 órdenes de magnitud menor que con el planteo original.

Análogamente se puede introducir una variable $\beta_j(t)$ denominada **variable hacia atrás**:

variable hacia atrás

$$\beta_j(t) = P\left(y_{t+1}, y_{t+2}, \cdots, y_T|q_t = S_j, \lambda\right) \tag{2.43}$$

Asumiendo que en el instante t el modelo λ se encuentra en el estado S_j, la variable hacia atrás $\beta_j(t)$ representa la probabilidad que a partir de ese momento se emita la porción restante de la secuencia de observaciones. Es decir, la secuencia parcial que comienza en el instante $t+1$ y se extiende hasta completar la secuencia $\mathbf{Y}$ en el instante final T.

También se puede resolver inductivamente el problema de evaluación empleando β, como se detalla en la figura 23:

Algoritmo de Retroceso

- INICIALIZACIÓN:

$$\beta_i(T) = 1 \quad \text{para} 1 \leqslant i \leqslant N \tag{2.44}$$

- ITERACIÓN:

  Calcular iterativamente hacia atrás la expresión de recurrencia:

$$\beta_i(t) = \sum_{j=1}^{N} a_{ij} \cdot b_j(y_{t+1}) \cdot \beta_j(t+1) \quad \text{para } 1 \leqslant t < T \tag{2.45}$$

- FINALIZACIÓN: Finalizar el proceso calculando:

$$P(y_1, y_2, \cdots, y_T|\lambda) = \sum_{i=1}^{N} a_{1i} \cdot b_i(y_1)\beta_1(i) \tag{2.46}$$

Figura 23: Algoritmo de Retroceso aplicado al problema de evaluación

Al igual que el algoritmo de avance, el algoritmo de retroceso resuelve el problema de evaluación con una complejidad $O(N^2T)$. Las operaciones necesarias pueden ser calculadas empleando una estructura de retícula similar a la presentada en la figura 21.

Algoritmo de Reestimación de Baum-Welch

Hasta este momento se consideró que los parámetros de los HMM ya adoptaban sus valores óptimos, valores compatibles con lo observado en un número muy grande de ejemplos de habla. A continuación se tratará la forma de resolver el segundo problema básico de los HMM introducidos en la sección anterior, el **problema de estimación**. Es decir, el problema que consiste en derivar el valor de los parámetros de un modelo a partir de muestras de entrenamiento. De los tres problemas básicos, éste es el que mayores dificultades presenta.

entrenamiento

Dado un conjunto de segmentos de habla disponibles como ejemplos de lo que podría emitir cada HMM, se puede considerar al **entrenamiento** como el problema que busca determinar el valor de los parámetros de tales modelos que permitan representar mejor el comportamiento estadístico observado en los datos. Esto se logra maximizando la probabilidad que dicho conjunto de datos hayan sido generados por tales HMM.

Considere que como conjunto de datos de entrenamiento se tiene secuencias de observaciones $\mathbf{Y}$, cada una asociada con un HMM λ correspondiente. Durante este proceso se busca el conjunto de parámetros de

λ que para todas sus secuencias $\mathbf{Y}$ asociadas maximice la probabilidad $P(\mathbf{Y}|\lambda)$. Formalmente el problema se podría especificar como:

$$\left(\hat{\boldsymbol{\pi}}_\lambda, \hat{\mathbf{A}}_\lambda, \hat{\mathbf{B}}_\lambda\right) = \underset{\forall(\boldsymbol{\pi}_\lambda, \mathbf{A}_\lambda, \mathbf{B}_\lambda)}{\operatorname{argmax}} P\left(\mathbf{Y}|\boldsymbol{\pi}_\lambda, \mathbf{A}_\lambda, \mathbf{B}_\lambda\right) \tag{2.47}$$

Debido a que la probabilidad condicional de la ecuación 2.47 se conoce como verosimilitud, el criterio de entrenamiento que maximiza esa probabilidad se conoce como *criterio de máxima verosimilitud* .

criterio de máxima verosimilitud

Si se pudiera conocer la correspondencia entre cada frame de entrenamiento con el estado de un modelo específico, calcular las probabilidades de máxima verosimilitud asociadas con cada modelo sería sencillo. Por un lado las probabilidades de transición se calcularían a partir de la estadística de las secuencias de estados, y por el otro, las probabilidades de emisión se podrían estimar a partir de las estadísticas de los vectores de atributos asociados con cada estado.

Sin embargo la naturaleza *"oculta"* de los estados contenidos en los HMM hace imposible conocer tal correspondencia. Es por ello que no hay un método analítico para calcular de manera directa los valores óptimos de estos parámetros, y se debe recurrir a algoritmos heurísticos para encontrar una buena aproximación.

Uno de esos algoritmos es el **algoritmo de Baum-Welch** .

algoritmo de Baum-Welch

La idea básica de este algoritmo es emplear un procedimiento de optimización iterativa que partiendo de un conjunto de valores iniciales para los parámetros de cada modelo, permita ir mejorando la estimación previa, empleado para ello los datos de entrenamiento. En este proceso se conserva la topología del HMM, es decir, que para todos los a_{ij} y los π_i que tengan valores nulos, esos valores no se modifican.

El proceso de optimización para cada modelo λ es básicamente el siguiente: en primer lugar se inicializa el modelo con algunas estimaciones provisorias de sus parámetros. Con el modelo así inicializado se calculan las probabilidades para cada uno de sus estados y para cada trama de las secuencias de observaciones correspondientes a los datos de entrenamiento asociados con λ. Finalmente se analiza la estadística de esas secuencias de estados, y los vectores de atributos asociados con cada uno de ellos, para reestimar los parámetros del modelo: sus probabilidades de transiciones y de emisión respectivamente. Empleando ahora los valores de parámetros reestimados, se vuelve a repetir el procedimiento.

Veamos ahora cómo se formaliza este procedimiento.

Supongamos por el momento que para entrenar el modelo λ disponemos de una sola palabra aislada, y que la misma comprende la secuencia de vectores de atributos $\{y_1, y_2, \cdots, y_T\}$; donde la primera trama se asocia con y_1 y la última con y_T. Además asumamos que se cuenta con un conjunto de estimaciones iniciales (groseras) para los parámetros de λ.

Bajo estas condiciones es posible calcular la probabilidad que el modelo emita el conjunto completo de los T vectores de atributos observados, y que se encuentre en el estado S_j en el instante t utilizando el algoritmo de avance-retroceso y el procedimiento descripto en la figura 24:

Algoritmo para Reestimar los Parámetros de λ Contando con una sola Muestra de Entrenamiento

- INICIALIZACIÓN:

$$\beta_i(T) = a_{iF} \tag{2.48}$$

- ITERACIÓN:

  Calcular iterativamente hacia atrás la expresión de recurrencia:

$$\beta_i(t) = \sum_{j=1}^{N} a_{ij} \cdot b_j(y_{t+1}) \cdot \beta_j(t+1) \quad \text{para } 1 \leqslant t < T \tag{2.49}$$

- FINALIZACIÓN: Finalizar el proceso calculando:

$$P\left(y_1, y_2, \cdots, y_T, q_t = S_j|\lambda\right) = \alpha_j(t) \cdot \beta_j(t) \tag{2.50}$$

Figura 24: Algoritmo para reestimación de parámetros empleando una sola muestra de entrenamiento

Utilizando la probabilidad así calculada para cada trama de la palabra se pueden reestimar los valores de probabilidad de transiciones y de emisión asociados con los parámetros de λ.

Sin embargo, la suposición de contar con una sola muestra de entrenamiento por modelo no es viable si lo que interesa es que tales modelos generalicen el comportamiento observado en múltiples instancias y realidades acústicas. En consecuencia, durante el entrenamiento suele haber varias muestras de una misma palabra correspondientes a un modelo dado. En esta situación es improbable que todas esas instancias presenten el mismo número de tramas ni vectores de características, tampoco que las duraciones relativas entre segmentos sea la misma para cada ejemplo. Por lo tanto es de esperar que la versión del modelo λ (inicializado de alguna manera) se ajuste mejor a ciertos ejemplos que a otros. Esto no es conveniente ya que artificialmente se estaría introduciendo un sesgo para favorecer ciertas evidencias, en detrimento de otras instancias igualmente válidas. Es decir, se debe buscar un mecanismo para que durante la reestimación se garantice el empleo de todas las muestras disponibles con el mismo factor de ponderación.

Como se acaba de ver, el producto $(\alpha_j(t) \cdot \beta_j(t))$ que permite reestimar el valor de los atributos de cada modelo suponiendo una sola palabra de muestra, representa la probabilidad conjunta de estar en el estado S_j durante el instante t y generar **un conjunto** particular de vectores de observaciones **correspondientes a un ejemplo**. Esta solución no permite especificar diferentes instancias.

Para poder utilizar en el entrenamiento de un modelo distintas muestras de una misma palabra, con posibles diferencias en sus vectores de observaciones, se necesita una expresión con la probabilidad condicio-

nal de ocupar el estado S_j dada una secuencia de vectores de observaciones.

Es así que se define la variable $\gamma_j(t)$ que indica la probabilidad de estar en el estado S_j durante la trama t, dados los vectores de atributos para un ejemplo de la palabra. Esa cantidad se puede derivar del producto $(\alpha_j(t) \cdot \beta_j(t))$ usando la regla de Bayes. Además es posible probar que el resultado implica una simple normalización de ese producto, empleando la probabilidad del modelo de generar las observaciones:

$$\begin{aligned} \gamma_j(t) &= P\left(q_t = S_j | y_1, y_2, \cdots, y_T, \lambda\right) \\ &= \frac{P\left(y_1, y_2, \cdots, y_T | q_t = S_j, \lambda\right) \cdot P\left(q_t = S_j | \lambda\right)}{P\left(y_1, y_2, \cdots, y_T | \lambda\right)} \\ &= \frac{P\left(y_1, y_2, \cdots, y_T, q_t = S_j | \lambda\right)}{P\left(y_1, y_2, \cdots, y_T | \lambda\right)} = \frac{\alpha_j(t) \cdot \beta_j(t)}{\alpha_F(T)} \end{aligned} \tag{2.51}$$

La normalización por $\alpha_F(t)$ asegura que cuando hay varios ejemplos para una misma palabra, todas las tramas de todos los ejemplos tengan la misma contribución durante la reestimación. Veamos ahora cómo se re-estiman cada uno de los parámetros del modelo λ empleando la variable γ. Durante la descripción de las reestimaciones se utilizará como notación para los valores estimados una barra sobre los símbolos correspondientes a cada parámetros, mientras que los mismos símbolos sin barras indican el valor de los mismos parámetros de acuerdo a la estimación previa.

Analicemos cómo reestimar $b_j(k)$, la probabilidad de observar el vector de atributos y_k cuando el modelo se encuentra en el estado S_j. Esta probabilidad se puede calcular como la relación entre la frecuencia de ocurrencia del símbolo y_k cuando el modelo se encuentra en el estado S_j respecto a la observación de cualquier símbolo en ese estado. Es decir, como la probabilidad de observar y_k cuando el modelo se encuentre en S_j, dividido por la probabilidad de estar en el estado Sj. Para tener en cuenta el conjunto completo de muestras de entrenamiento de la palabra considerada, se debe sumar tanto el numerador como el denominador sobre todos las tramas de todos los ejemplos. Así, asumiendo E ejemplos de la misma palabra correspondiente a λ, la re-estimación para la probabilidad de emisión está dada por:

$$\overline{b}_j(k) = \frac{\sum_{e=1}^{E} \sum_{\{t: y_{t_e} = y_k,\ t=1,2,\cdots,T_e\}} \gamma_j(t,e)}{\sum_{e=1}^{E} \sum_{t=1} T_e \gamma_j(t,e)} \tag{2.52}$$

En la ecuación 2.52 se especifica como T_e el número de tramas del e-ésimo ejemplo, como y_{t_e} al vector de atributos asociado con la trama en el instante t de ese mismo ejemplo, y también se usa $\gamma_j(t,e)$ para expresar el valor de $\gamma_j(t)$ para la e-ésima instancia de la palabra. Además el denominador de esa ecuación es la suma de las probabilidades individuales de estar en el estado j para cualquier instante de tiempo, dado el conjunto completo de datos de entrenamiento. Ese valor da una idea del número de frames en los que ese estado está activo u ocupado.

Para estimar los parámetros a_{ij}, se necesita saber la probabilidad de transición entre cada par de estados. Volvamos por un momento a considerar un solo ejemplo para el modelo λ.

Se define $\xi_{ij}(t)$ como la probabilidad que se produzca una transición del estado i al j en el instante t, dado que el modelo genera el conjunto completo de vectores de atributos correspondientes a la palabra disponible como ejemplo:

$$\xi_{ij}(t) = \frac{\alpha_i(t) \cdot a_{ij} \cdot b_j(y_{t+1} \cdot \beta_j(t+1))}{\alpha_F(T)} \quad \text{para } 1 \leqslant t < T \tag{2.53}$$

La ecuación 2.53 se puede utilizar para calcular la probabilidad de transición entre cualquier par de estados emisores entre el instante $t = 1$ y $t = T - 1$. Para el instante final, como no se produce la transición a otro estado emisor, y la única posibilidad es ir hacia el estado F con probabilidad $\xi_{iF}(T)$,

$$\xi_{iF}(T) = \frac{\alpha_i(T) \cdot a_{iF}}{\alpha_F(T)} \tag{2.54}$$

Por su parte para el estado inicial, es necesario calcular la probabilidad de transición a cada uno de los estados emisores. Esa transición desde el estado I solamente se puede dar en el instante $t = 0$, antes que se produzca cualquier emisión, por lo que para ese caso:

$$\xi_{Ij}(0) = \frac{\pi_j \cdot b_j(y_1 \cdot \beta_j(1))}{\alpha_F(T)} \tag{2.55}$$

La probabilidad total de transición entre cualquier par de estados i y j se obtiene sumando los valores de $\xi_{ij}(t)$ sobre todas las tramas para las cuales sea posible dicha transición, y dividiendo esa cantidad por la probabilidad total de ocupar el estado i: γ_i. Asumiendo nuevamente E ejemplos de la palabra, se puede reestimar a_{ij} de la siguiente manera:

$$\overline{a}_{ij} = \frac{\sum_{e=1}^{E} \sum_{t=1}^{T_e - 1} \xi_{ij}(t, e)}{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_i(t, e)} \quad \text{para } 1 \leqslant i, j \leqslant N \tag{2.56}$$

Nuevamente $\xi_{ij}(t, e)$ denota el valor de $\xi_{ij}(t)$ para el e-ésimo ejemplo de entrenamiento. Además se debe observar que la suma en función de t del numerador solo incluye las tramas hasta $T_e - 1$. El último frame no se incluye dado que en el mismo no es posible que se de alguna transición a otro estado emisor, y por definición $\xi_{ij}(T, e)$ es nula para todos los pares de estados emisores. Las transiciones desde un estado emisor hasta el estado F sólo puede ocurrir en el tiempo T_e, por lo que la probabilidad de transición a_{iF} se puede reestimar como:

$$\overline{a}_{iF} = \frac{\sum_{e=1}^{E} \xi_{iF}(T_e, e)}{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_i(t, e)} \quad \text{para } 1 \leqslant i \leqslant N \tag{2.57}$$

Las transiciones desde el estado I solamente pueden ocurrir en $t = 0$, por lo tanto $\xi_I(0, e) = 1$ para todos los ejemplos, por lo cual:

$$\overline{\pi}_i = \frac{\sum_{e=1}^{E} \xi_{Ii}(0, e)}{E} \quad \text{para } 1 \leqslant i \leqslant N \tag{2.58}$$

Se puede ver de las ecuaciones 2.55, 2.57 y 2.58 que si cualquier π_i o a_{ij} son inicialmente nulos, tras las sucesivas iteraciones de reestimación lo seguirán siendo, preservando la topología del HMM.

El algoritmo de Baum-Welch es un caso particular de un método general conocido como **algoritmo de esperanza-maximización** (algoritmo *expectation-maximization* en inglés) o de manera resumida **algoritmo EM**, aplicable a problemas en los que se debe estimar los parámetros de un modelo y los datos observables son incompletos. *algoritmo EM*

Se ha probado que para cada iteración de este algoritmo, el nuevo conjunto de parámetros es al menos tan bueno como el anterior; y en general, las nuevas iteraciones van mejorando el modelo. Si se repite el proceso de reestimación un número suficiente de veces, el modelo convergerá a una solución localmente óptima. Es decir, luego de la reestimación empleando el algoritmo de Baum-Welch, se garantiza que la probabilidad de los datos de entrenamiento dados los modelos con el nuevo conjunto de parámetros es mayor que la probabilidad para el conjunto de modelos previo, excepto si se alcanza el **punto crítico**, que indica la presencia de un óptimo local, y por lo tanto a partir de allí los modelos y las probabilidades no cambiarán al repetir el proceso de reestimación.

En otras palabra, el algoritmo de Baum-Welch converge de manera monótona y en tiempo polinómico (con respecto al número de estados y la longitud de las secuencias acústicas), a puntos de estacionaridad locales sobre la función de verosimilitud.

Es así que el procedimiento de reestimación se puede aplicar repetidamente hasta que la diferencia de probabilidad entre el modelo previo y el actual sea suficientemente pequeña.

Algoritmo de Viterbi

Consideremos ahora uno de los algoritmos más empleados para resolver el tercer problema básico de los HMM: el **problema de decodificación**.

El **algoritmo de Viterbi** es un procedimiento general de búsqueda sincrónica en el tiempo, aplicable a espacios definidos por la retícula presentada en la figura 21, y su funcionamiento está basado en los principios de programación dinámica [14]. *Algoritmo de Viterbi*

Este algoritmo es similar al procedimiento para el cálculo de manera recursiva de la variable hacia adelante α que se describió en el algoritmo de avance, pero reemplaza la sumatoria sobre todos los estados predecesores por una maximización e introduce un puntero ψ en cada estado para guardar la ubicación de su mejor predecesor. Además

utiliza la variable $\delta_i(t)$ para almacenar la probabilidad acumulada del mejor camino hasta el estado i en el tiempo t:

$$\delta_i(t) = \max_{q_1,q_2,\cdots,q_{t-1}} P(q_1, q_2, \cdots, q_t = S_i, y_1, y_2, \cdots, y_t|\lambda) \quad (2.59)$$

Es decir que $\delta_i(t)$ es el mejor puntaje (la mayor probabilidad) que tiene el camino óptimo parcial que comprende las primeras t observaciones y finaliza en el estado S_j.

Por inducción se tiene que:

$$\delta_j(t+1) = \left(\max_{\forall i} \left\{\delta_i(t) \cdot a_{ij}\right\}\right) \cdot b_j(y_{t+1}) \quad (2.60)$$

Como se comentó, el algoritmo almacena en el vector $\psi_j(t)$ un registro de los argumentos que maximizaron la ecuación 2.60 para cada trama t y estado j, y de esta forma puede recuperar la secuencia óptima completa.

El procedimiento completo para encontrar la mejor secuencia de estados se presenta en la figura 25:

El algoritmo de Viterbi realiza una búsqueda exhaustiva en el espacio de estados, es decir que evalúa de manera eficiente todas las secuencias de estados $\mathbf{q}$ contenidas en el espacio de búsqueda. Sin embargo en las aplicaciones de RAH actuales debido al tamaño del espacio de búsqueda se lo suele podar empleando procesos de *búsqueda en haz*.

En cada paso de tiempo t, después del cálculo de 2.63 (paso 2 del algoritmo), solamente se continúan los caminos para los que $\delta_i(t)$ sea mayor al umbral de haz definido como $\theta = \xi \cdot max_j \delta_j(t$ con $0 \leqslant \xi \leqslant 1$.

Una consecuencia de esta poda es que ya no se puede garantizar encontrar el camino óptimo.

Observe por otro lado la variable P_{max} de la ecuación 2.65, que indica la probabilidad que tiene la secuencia óptima de estados del modelo λ de generar la secuencia de observaciones Y. Este valor se puede usar para puntuar cada modelo λ ante una secuencia Y de entrada. Durante el reconocimiento del habla, para cada modelo de palabra candidata se calcula su P_{max}, y la secuencia de Y se etiqueta con la palabra que consigue el P_{max} máximo.

PROBLEMA DE ESTIMACIÓN EMPLEANDO VITERBI El algoritmo de Viterbi también se puede utilizar para resolver de manera aproximada pero eficiente el problema de estimación.

Es posible realizar la reestimación de los parámetros del modelo λ empleando solamente el camino más probable a través de sus estados. De esta manera, los cálculos se simplifican respecto a los que demanda el algoritmo de Baum-Welch, resultando una solución aproximada pero con un menor costo computacional.

A través de este procedimiento, para cualquier trama de entrada, la probabilidad que se ocupe un estado solamente puede ser igual a uno o cero, dependiendo si ese estado está en la secuencia óptima.

Decodificación empleando el Algoritmo de Viterbi

- INICIALIZACIÓN:

$$\delta_i(1) = \pi_i \cdot b_i(y_1) \quad \text{para } 1 \leqslant i \leqslant N \tag{2.61}$$

$$\psi_1(i) = 0 \quad \text{para } 1 \leqslant i \leqslant N \tag{2.62}$$

- RECURSIÓN: Calcular iterativamente las secuencias de estados de máxima verosimilitud y sus probabilidades empleando las expresiones de recurrencia para $2 \leqslant t \leqslant T$ y $1 \leqslant j \leqslant N$:

$$\delta_j(t) = \left(\max_{1 \leqslant i \leqslant N} \{\delta_i(t-1) \cdot a_{ij}\} \right) \cdot b_j(y_t) \tag{2.63}$$

$$\psi_j(t) = \left(\operatorname*{argmax}_{1 \leqslant i \leqslant N} \{\delta_i(t-1) \cdot a_{ij}\} \right) \tag{2.64}$$

- FINALIZACIÓN: Obtener el estado final más probable y el valor de su probabilidad:

$$P_{max} = \max_{1 \leqslant i \leqslant N} \{\delta_i(T)\} \tag{2.65}$$

$$\mathbf{q}^* = \operatorname*{argmax}_{1 \leqslant i \leqslant N} \{\delta_i(T)\} \tag{2.66}$$

- BACKTRACKING:: Obtener la secuencia de estados óptima haciendo:

$$\mathbf{q}^*{}_t = \psi \mathbf{q}^*{}_{t+1}(t+1) \quad \text{para } t = T-1, T-2, \cdots, 1 \tag{2.67}$$

Figura 25: Utilización del Algoritmo de Viterbi en el problema de Decodificación

Una vez que se identifica la secuencia óptima para todos los estados y tramas de una secuencia particular de observaciones haciendo uso de la variable $\delta_j(t)$ de la forma ya detallada, cada frame quedará asignado a un estado, y se podrá saber qué observación se asocia con cada estado y, obviamente, las secuencias de estados. Por lo tanto para la reestimación se necesitará acumular para todas las instancias de cada

palabra las estadísticas de los vectores de atributos que se producen en cada estado activado, así como las transiciones de estados sobre los caminos óptimos. Específicamente, asumiendo E ejemplos de una palabra correspondiendo a un modelo, se deben almacenar:

- $n_j(y_t = k)$: número de tramas para las cuales cada estado j se asoció con cada vector de atributos k.
- n_{ij} : número de tramas en las que se produjo una transición entre cada par de estados i y j.
- n_{Ij} : número de veces que la primera trama estuvo asociada con cada estado j.
- n_{iF} : número de veces que la última trama se asoció con cada estado i.
- n_i : número de tramas en que se ocupó cada estado i.

Con estos valores, las fórmulas de reestimación vienen dadas por:

$$\overline{b}_j(k) = \frac{n_j(y_t = k)}{n_j} \tag{2.68}$$

$$\overline{a}_{ij} = \frac{n_{ij}}{n_i} \quad \text{para todos los estados emisores, } 1 \leqslant i, j \leqslant N \tag{2.69}$$

$$\overline{a}_{iF} = \frac{n_{iF}}{n_i} \quad \text{para todo i tal que, } 1 \leqslant i \leqslant N \tag{2.70}$$

$$\overline{\pi}_i = \frac{n_{Ij}}{E} \quad \text{para todo j tal que, } 1 \leqslant j \leqslant N \tag{2.71}$$

Se puede observar la similitud entre las ecuaciones 2.68, 2.69, 2.70 y 2.71, con respecto a las ecuaciones obtenidas mediante el algoritmo de Baum-Welch: 2.52, 2.55, 2.57 y 2.58 respectivamente.

La diferencia es que en el caso actual los valores de las probabilidades de activación de un estado dependientes de una trama dada, se reemplazan por 1 o 0, dependiendo si los estados relevantes están o no ocupados en el instante de tiempo dado.

Al igual que lo visto para el algoritmo de Baum-Welch, aquí también se puede aplicar repetidamente el procedimiento de reestimación hasta que las diferencias de verosimilitud entre un par de iteraciones sean suficientemente pequeñas.

Por último, la magnitud de la diferencia en los modelos obtenidos empleando Viterbi o Baum-Welch, está supeditada a que la verosimilitud de la secuencia más probable sea mucho mayor que la de las secuencias alternativas, lo que se logra asegurando que las funciones de probabilidades de emisión entre los estados de cada modelo sean suficientemente diferentes.

PROBLEMA DE EVALUACIÓN EMPLEANDO EL VITERBI Finalmente, el algoritmo de Viterbi también puede ser usado para resolver el primer problema básico de los HMM: el problema de evaluación.

La probabilidad $P(\mathbf{Y}|\lambda)$ de observar una secuencia en particular dado un modelo está conformada por la contribución de un número muy grande de secuencias de estados alternativas, como se puede ver en la sumatoria de la ecuación 2.38. Sin embargo debido a las distribuciones de probabilidades asociadas con cada estado, se puede encontrar que el aporte que hacen a la probabilidad total muchas de esas secuencias de estados es muy pequeña.

Para reducir el número de cálculos de dicha ecuación se podría pensar en reemplazar la probabilidad mencionada por una aproximación $\hat{P}$, considerando en su cálculo solamente la secuencia de estados más probable:

$$\hat{P}(\mathbf{Y}|\lambda) = \max_{\forall \mathbf{q}} (P(\mathbf{Y}, \mathbf{q}|\lambda)) \tag{2.72}$$

La probabilidad de tal secuencia se puede calcular empleando el algoritmo de Viterbi, mediante el procedimiento detallado en la figura 26:

Evaluación empleando el Algoritmo de Viterbi

- INICIALIZACIÓN:

$$\delta_j(1) = \alpha_j(1) = \pi_j \cdot b_j(y_1) \tag{2.73}$$

- ITERACIÓN:

  Calcular iterativamente las demás variables empleando la expresión de recurrencia:

$$\delta_j(t) = \max_{\forall \mathbf{i}} \left(\delta_i(t-1) \cdot a_{ij}\right) \cdot b_j(y_t) \quad \text{para } 1 < t \leqslant T \tag{2.74}$$

  Aquí nuevamente se asume independencia entre observaciones subsecuentes.

- FINALIZACIÓN: Finalizar el proceso calculando:

$$\hat{P}(y_1, y_2, \cdots, y_T|\lambda) = \delta_F(T) = \max_{\forall \mathbf{i}} (\delta_i(T) \cdot a_{iF}) \tag{2.75}$$

Figura 26: Utilización del Algoritmo de Viterbi en el problema de Evaluación

La diferencia entre la probabilidad considerando el conjunto completo de secuencias de estados obtenida empleando el algoritmo de avance-retroceso, y la dada por la expresión aproximada de la ecuación 2.75, calculada a través del algoritmo de Viterbi depende de la magnitud de la contribución a la probabilidad total, de la *mejor* secuencia de estados respecto a las secuencias despreciadas.

Al igual que en la aplicación de Viterbi en el problema de estimación, si las *fdp* de los vectores de atributos de todos los estados difieren

substancialmente entre sí, la probabilidad que la mejor secuencia haya generado las observaciones no debería diferir demasiado de la probabilidad completa, incluyendo todas las secuencias de estados. Sin embargo estas diferencias se incrementarían si el mejor camino incluye varias tramas consecutivas con características que se puedan asociar a dos o más estados que tengan *fdp* muy similares. Es por ello que el diseño de los modelos empleados en los reconocedores actuales busca evitar que se produzcan secuencias de estados con *fdp* de emisión similares.

En conclusión, a pesar de las desventajas teóricas de preservar en el cálculo solamente la mejor secuencia de estados, en la práctica la diferencia en los resultados obtenidos utilizando el cálculo completo y la versión aproximada generalmente es exigua, mientras que la reducción en el requerimiento de cómputo es significativa.

2.3.2 Extensiones para Modelos Continuos

Hasta este momento se consideraron solamente HMM discretos, es decir que se asumieron las probabilidades de observaciones como símbolos discretos provenientes de un alfabeto finito. Esto permitió utilizar funciones de probabilidad de masa dentro de cada estado correspondiente a un modelo.

Sin embargo, como ya se estableció al introducir los modelos acústicos, las distribuciones discretas suponen una degradación en la representación de las señales de entrada. Para utilizar densidades de observaciones continuas se deben imponer algunas restricciones sobre la forma de las *fdp* para asegurar que sus parámetros se puedan reestimar de manera consistente.

Usando a notación $N(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ para indicar la densidad de probabilidades del vector observado $\mathbf{y}$ dada la distribución normal con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$, la forma más general de *fdp* para las que se puede encontrar formulado el procedimiento de reestimación está dada por las mezclas finitas dadas por la expresión:

$$b_j(y) = \sum_{m=1}^{M} c_{jm} \cdot N(\mathbf{y}; \mu_{jm}, \Sigma_{jm}) \tag{2.76}$$

El segundo término dentro del sumando de la ecuación 2.76 corresponde a la probabilidad de emisión de la m-ésima componente de la mezcla de Gaussianas para el estado j; mientras que c_{jm} denota el coeficiente de ponderación para la m-ésima componente de la mezcla para el estado j.

Además en la ecuación 2.76 no es necesario que la distribución esté dada por mezclas de Gausssianas, sino que puede corresponder a cualquier función de densidad log-cóncava o con simetría elíptica. Sin embargo por ser las más utilizadas, en lo que sigue se supondrá a N como una función Gaussiana.

Los coeficientes de ponderación c_{jm} satisfacen las restricciones estocásticas:

$$\sum_{m=1}^{M} c_{jm} = 1 \quad \text{para } 1 \leqslant j \leqslant N \tag{2.77}$$

$$c_{jm} \geqslant 0 \quad \text{para } 1 \leqslant j \leqslant N \tag{2.78}$$

tal que la *fdp* se normaliza de la siguiente forma:

$$\int_{-\infty}^{\infty} b_j(y)\, dy = 1 \quad \text{para } 1 \leqslant j \leqslant N \tag{2.79}$$

Se demuestra que la *fdp* así definida se puede utilizar para aproximar tanto como se desee cualquier función de densidad continua y finita.

En [87] se demostró que un estado con densidad de probabilidad modelada como mezclas de Gaussianas es equivalente a un HMM con múltiples estados cada uno de ellos con una densidad de probabilidad dada por una Gaussiana simple.

Asumiendo que se tiene una estimación inicial de los parámetros para las M componentes de todas las mezclas Gaussianas representando la *fdp* para el estado j, se puede utilizar la reestimación de Baum-Welch para hallar nuevas estimaciones de los parámetros c_{jm}, μ_{jm} y Σ_{jm}. Cuando se usan mezclas de Gaussianas se debe ponderar la contribución de cada observación y por una probabilidad específica para la componente de la mezcla m. Análogamente a lo hecho previamente se define una variable $\gamma_{jm}(t)$ para indicar la probabilidad encontrarse en el estado j en el instante t y utilizar la componente m-ésima para generar y_t, dado que el modelo genera toda la secuencia de vectores de observaciones representando un ejemplo de determinada palabra:

$$\gamma_{jm}(t) = \frac{\sum_{i=1}^{N} \alpha_i(t-1) a_{ij} c_{jm} b_{jm}(y_t) \beta_j(t)}{\alpha_F(T)} \tag{2.80}$$

Ahora para E ejemplos de la palabra, sumando los valores de $\gamma_{jm}(t)$ sobre todas las tramas de todos los ejemplos se obtiene la probabilidad total para la m-ésima componente para el estado j. Dividiendo esa cantidad por la suma correspondiente de $\gamma_j(t, e)$ se obtiene la reestimación para el coeficiente de ponderación de esa componente de la mezcla:

$$\bar{c}_{jm} = \frac{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}(t, e)}{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_j(t, e)} \tag{2.81}$$

La ecuación de reestimación para el vector de medias es:

$$\overline{\mu}_{jm} = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_{jm}(t,e)y_{te}}{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_{jm}(t,e)} \tag{2.82}$$

en la ecuación 2.82 se puede observar que se pesa cada término del numerador de la ecuación 2.81 utilizando la observación, obteniendo el valor esperado de la porción del vector de observaciones atribuíble a la m-ésima componente de la mezcla.

Finalmente para la matriz de covarianza la fórmula de reestimación es:

$$\overline{\Sigma}_{jm} = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_{jm}(t,e)(y_{te}-\overline{\mu}_{jm})(y_{te}-\overline{\mu}_{jm})^T}{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_{jm}(t,e)} \tag{2.83}$$

2.3.3 Extensiones para Modelos Semi-Continuos

Al incrementar el número de componentes de mezclas Gaussianas aumenta la variedad de morfologías de las distribuciones que se pueden representar. Sin embargo el número de componentes de mezclas por estado que se pueden utilizar está limitado por la cantidad de datos disponibles. Al aumentar el número de componentes de mezclas también se introducen nuevos parámetros para estimar. Así, inevitablemente se debe resolver una relación de compromiso entre la fidelidad de la representación de la probabilidad de emisión, la cantidad de datos de entrenamiento necesarios y la robustez de la estimación de los modelos estocásticos.

Por otro lado se puede encontrar entre estados internos correspondientes a unidades acústicas diferentes que hay muchas probabilidades de emisión muy parecidas. Una forma sencilla de aprovechar esa redundancia es utilizar un conjunto único de distribuciones Gaussianas para todos los estados de todos los modelos, caracterizando ahora cada estado mediante los pesos de las mezclas correspondientes a cada estado. De esta forma los parámetros de las distribuciones están enlazados entre estados diferentes. Estos constituyen los HMM-semicontinuos presentados anteriormente.

mezclas enlazadas

Cuando se utilizan **mezclas enlazadas**, las probabilidades de emisión $b_j(y)$ para cualquier estado j se calcula de la misma forma que en la ecuación 2.76, pero en este caso los pesos de cada mezcla c_{jm} son específicos para cada estado, mientras que los $b_{jm}(y)$ son mismos para todos los estados.

Usando la nueva definición para la probabilidad de emisión se puede derivar la fórmula para la media de la m-ésima componente (μ_m):

$$\overline{\mu}_m = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\sum_{j=1}^{N}\gamma_{jm}(t,e)y_{te}}{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\sum_{j=1}^{N}\gamma_{jm}(t,e)} \tag{2.84}$$

Mientras que para la covarianza de la misma componente ($\boldsymbol{\Sigma}_m$) la expresión de reestimación tiene la forma:

$$\overline{\Sigma}_m = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\sum_{j=1}^{N}\gamma_{jm}(t,e)(y_{te}-\overline{\mu}_{jm})(y_{te}-\overline{\mu}_{jm})^T}{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\sum_{j=1}^{N}\gamma_{jm}(t,e)} \tag{2.85}$$

Comparando las ecuaciones 2.84 y 2.85 respecto al caso de mezclas no enlazadas de las ecuaciones 2.82 y 2.83, se puede ver que la única diferencia es que además de sumar las contribuciones sobre todas las tramas de todos los ejemplos, ahora también se las suma respecto a todos los estados.

Por su parte la fórmula de reestimación para c_{jm} es igual que en la ecuación 2.81.

Es importante destacar que el enlazado de mezclas de componentes es solamente un caso particular de un concepto más general conocido como **enlazado de parámetros** . En el caso general, se puede enlazar cualquier parámetro de cualquier estado, y el único efecto en la fórmula de reestimación aparece en la naturaleza de las sumatorias y el indexado de los parámetros del modelo.

enlazado de parámetros

2.3.4 Extensión a Secuencias de Palabras

Una suposición que se hizo para simplificar la explicación de los algoritmos fundamentales para el RAH fue que la tarea de reconocimiento correspondía a palabras aisladas. Estos HMM correspondientes a modelos de palabras completas se pueden extender de manera muy sencilla para modelar secuencias de palabras. Se puede considerar que las palabras conectadas se representan usando modelos de alto nivel en los que cada estado corresponde a una palabra completa, y en el que las probabilidades de transiciones vienen dadas por las probabilidades del modelo de lenguaje.

En la figura 27 se muestra la forma en que se disponen cada modelo de palabra para reconocer habla continua:

Como se muestra en la figura 27, se colocan en paralelo K HMM diferentes, cada uno correspondiendo a una palabra y se los conecta empleando sus estados extremos no emisores: S_σ y S_τ. Finalmente se conforma un bucle al permitir la conexión desde el estado S_τ al S_σ. Generalmente el valor de las probabilidades de transición $a_{S_\sigma i}$ se hacen igual a 1/K. El valor de $a_{S_\tau S_\sigma}$ debería ser igual a 1, pero en la

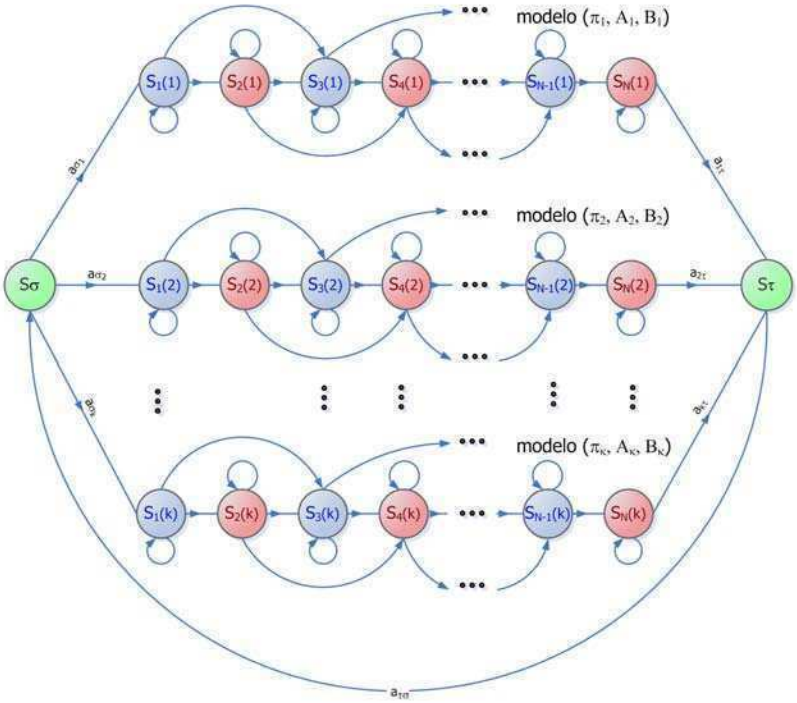

Figura 27: Esquema de HMM anidados empleados en el reconocimiento del habla continua. Los modelos de las clases Ω_k se disponen en paralelo

práctica se hace $a_{S_\tau S_\sigma} = \omega \quad \in \Re$, estimando ω de manera heurística para balancear el número de deleciones e inserciones durante el reconocimiento.

En el caso del reconocimiento de palabras aisladas no interesaba la secuencia de estados en particular dentro del modelo, sino la verosimilitud de cada modelo de palabra de emitir los vectores de atributos observados. Cuando se reconocen palabras conectadas se necesita determinar la secuencia más probable de palabras, tal que a nivel de palabras es necesario el algoritmo de Viterbi.

entrenamiento embebido

Cuando los datos de entrenamiento presentan palabras conectadas pronunciadas de manera natural, se pueden utilizar los mismos algoritmos de entrenamiento que los empleados para palabras aisladas. En general no se necesita segmentar los datos a nivel de palabras antes de comenzar el entrenamiento, sino que basta con una transcripción a nivel de frase para realizar un **entrenamiento embebido** . En este procedimiento se obtiene un modelo compuesto para toda la frase concatenando los modelos de secuencias de palabras requeridos. Si dentro de cada modelo de palabra se utilizan los modelos no emisores al inicio y final de cada modelo de palabra, esa concatenación es sencilla. Implica conectar el estado final de un modelo con el estado inicial de otro. Los parámetros del modelo compuesto se entrenan usando el mismo procedimiento que el utilizado para palabras aisladas. Si un estado ocurre más de una vez en el modelo compuesto (es decir que la frase contiene

más de una instancia para alguna palabra), todas las ocurrencias de ese estado van a contribuir a la reestimación de los parámetros.

2.3.5 Evaluación de Desempeño del RAH

Para comparar el desempeño de los reconocedores automáticos del habla se suelen considerar diversos aspectos entre ellos la cantidad o porcentaje de palabras reconocidas correctamente, o la velocidad del proceso de reconocimiento.

Cuando se reconoce habla conectada se pueden encontrar tres tipos de errores:

- Errores por sustitución (S): cuando se reconoce una palabra de manera errónea.
- Errores por supresión (D): corresponde al caso en que se omite una palabra de entrada
- Errores por inserción (I): que se produce cuando se reconoce una palabra extra.

Sea N la cantidad total de palabras a reconocer, y empleando el conteo de los valores para esos tres tipos de errores, se definen las siguientes figuras de mérito para cuantificar la calidad del reconocimiento:

Tasa de reconocimiento de palabras (R):

$$R = \frac{N - D - S}{N} \times 100\,\% \tag{2.86}$$

Precisión del reconocimiento (P):

$$P = \frac{N - D - S - I}{N} \times 100\,\% \tag{2.87}$$

Tasa de error de palabras (WER) definidas como:

$$WER = 100\,\% - P \tag{2.88}$$

De estas tres medidas, la más utilizada es la tasa de error de palabras, que como en la ecuación 2.88 se denota como WER por sus siglas en inglés.

Debido a que en general no hay una correspondencia uno a uno entre la secuencia de palabra real y la reconocida, se utiliza un procedimiento de alineación basada en programación dinámica antes de poder calcular el WER.

Por otra parte, se suelen utilizar medidas para cuantificar el costo de procesamiento que imponen los reconocedores. La forma más habitual de realizar esta estimación consiste en calcular el tiempo que insume el reconocimiento sobre el conjunto completo de evaluación, dividido el número de ventanas de análisis. Este valor luego se normaliza por la

duración de una ventana. Así se obtiene un tiempo de reconocimiento normalizado (TR), útil para calcular y comparar costos de ejecución para distintos reconocedores. Esta medida es obviamente dependiente del hardware.

Sean T_{REC} el tiempo de reconocimiento promedio para una ventana de análisis y T_{REAL} la duración de la ventana, el tiempo de reconocimiento normalizado viene dado por la ecuación 2.89

$$RT = \frac{T_{REC}}{T_{REAL}} \times 100\,\% \tag{2.89}$$

3 INFORMACIÓN SUPRASEGMENTAL Y PROSODIA

ÍNDICE

En este capítulo se presenta el marco teórico referido a la información suprasegmental o prosódica, fundamentos sobre los que se sustentarán las propuestas formuladas en esta tesis.

El capítulo comienza con la definición de los conceptos básicos asociados con la prosodia, y su relación con la noción de información suprasegmental.

En la segunda parte se describen los atributos prosódicos.

En la tercera sección se presentan los constituyentes prosódicos, un modelo de organización jerárquica de la información prosódica.

Posteriormente se presentan los principales modelos empleados para describir y utilizar la entonación.

Finalmente se describen los métodos computacionales empleados para la estimación automática y procesamiento de la información prosódica a partir de la señal de habla, que serán empleados en el transcurso de la tesis.

3.1 DEFINICIONES BÁSICAS

Al oír un fragmento de habla es posible notar que el tono de la voz se mueve de manera ascendente y descendiente, siguiendo un patrón melódico reconocible. También podemos oír segmentos o sílabas que son reducidas o prolongadas, aparentemente de acuerdo con algún patrón subyacente. También percibimos que algunas sílabas o palabras suenan más prominentes que otras, que las secuencias de palabras pronunciadas son subdivididas por el locutor en frases constituidas por palabras que parecieran formar parte de una misma entidad, y que de igual manera en un nivel de análisis superior, esas frases pueden sonar como si estuvieran relacionadas unas con otras, o por el contrario, como si nada tuvieran que ver entre ellas.

prosodia

Todo este conjunto de fenómenos están relacionados con la **prosodia** del habla.

La palabra prosodia se origina del término griego *prosoidía* (*prós:* agregado y *oidé:* canto). Se la empleaba para indicar el "canto con acompañamiento de un instrumento musical". Se puede constatar que desde su origen existió una estrecha relación entre este término y la "musicalidad" del habla.

En épocas posteriores, ese término se utilizó en el campo de las "ciencia de la versificación y leyes de la métrica", gobernando la modulación de la voz durante la lectura de poesías.

En la actualidad el término prosodia denota algunos atributos del habla que determinan su altura tonal, duración (cantidad), y sonoridad (calidad) de segmentos individuales; y la melodía, ritmo y patrones acentuales en secuencias de segmentos de habla. Estos atributos están correlacionados con magnitudes físicas que se pueden medir en la señal: frecuencia fundamental, duración segmental e intensidad [123].

De lo anterior se puede deducir que los atributos prosódicos se manifiestan tanto a nivel segmental como suprasegmental. Sin embargo, cobran significado de manera contrastiva, es decir por las diferencias de magnitud registradas en unidades vecinas. Debido a esto, en la literatura se suele usar los término información suprasegmental e información prosódica de manera indistinta.

En la fonética moderna el término prosodia se emplea con mayor frecuencia para referirse a aquellas propiedades del habla que no pueden obtenerse a partir del análisis de unidades segmentales de la señal de voz (secuencias cuyas duraciones se encuentran al nivel de fonemas). Ejemplos de estas propiedades son la modulación controlada del tono de voz, las prolongaciones y reducciones de las duraciones de segmentos y sílabas, y la fluctuación intencional de la intensidad. A nivel perceptual, las propiedades mencionadas dan lugar a características

como los patrones percibidos de prominencia relativa entre sílabas, codificada como aspectos melódicos y rítmicos del habla.

3.2 ATRIBUTOS PROSÓDICOS

Si bien como se mencionó, las definiciones actuales de prosodia contemplan a la altura tonal, la sonoridad, la duración y la calidad vocal, como sus principales atributos perceptuales, se pueden considerar algunos atributos prosódicos adicionales.

No resulta fácil hacer una lista exhaustiva de tales atributos, ya que en la propia comunidad de la lingüística no hay concenso al respecto [45].

Aquí se seguirá el criterio empleado por [96] que considera dos grupos de atributos prosódicos: los básicos y los compuestos. La diferencia entre ambas clases está dada en que los atributos prosódicos básicos se pueden determinar a partir de intervalos segmentales, mientras que a los compuestos solamente es posible definirlos contando con fragmentos más extensos de habla.

El conjunto de **atributos prosódicos básicos** está formado por: altura tonal, sonoridad, cantidad o duración, calidad vocal, velocidad del habla y pausas.

Por su parte los **atributos compuestos** que surgen por la variación de los atributos básicos en el tiempo comprenden: entonación, acentuación, frases prosódicas, ritmo y hesitación o balbuceo. Estos atributos están vinculados con unidades suprasegmentales de diferentes duraciones: sílabas, palabras, frases, oraciones, actos de diálogo y turnos de habla.

En el Capítulo 1, específicamente cuando se trató la psicoacústica del habla se explicaron dos de los rasgos prosódicos mencionados: sonoridad y altura tonal. A continuación se describirán los atributos prosódicos restantes.

3.2.1 Cantidad o Duración

El término **cantidad** designa las diferencias abstractas de duración, que se establecen en una lengua con la finalidad de distinguir significados, a través de la distinción léxica.

A diferencia por ejemplo del Alemán o del Latín clásico donde se puede encontrar una oposición binaria de cantidad entre vocales acentuadas cortas y largas, o del Italiano en donde esa oposición se da entre consonantes, o en lenguas como el Finés, el Estonio y el Húngaro que presentan oposiciones tanto vocálicas como consonánticas; el Español no utiliza oposiciones funcionales de cantidad segmental. No obstante, este atributo influye en la calidad de la voz.

La **duración** puede considerarse también un fenómeno segmental, puesto que cada sonido posee una duración característica. Así por ejemplo, es sabido que la fricativas son más largas que las oclusivas, que las sordas son las más largas que las sonoras, etc. Articulatoriamente, la duración se basa en el mantenimiento por más o menos tiempo

de una determinada configuración articulatoria. Por el fenómeno de la coarticulación, dicha configuración (y, consiguientemente, la duración) se ve alterada en función del contexto.

Sin embargo también es una unidad prosódica, ya que se necesita una valoración relativa para determinar la cantidad de un segmento.

La duración de sonidos, sílabas, palabras, enunciados y pausas es un importante indicador de estructuras dircursivas y expresivas, por lo que puede alterar el tempo o velocidad del habla. Como elemento suprasegmental, tanto las sílabas tónicas como las pertenecientes al tonema suelen ser más largas.

3.2.2 Calidad Vocal

El término **calidad de la voz** o timbre comprende atributos relacionados con la estructura espectral de un fono completo. Abarca medidas cuantitativas como *jitter* , *shimmer* , o la energía relativa de los armónicos respecto del F0. Estas cualidades determinan el timbre o grado de ronquera percibida en una voz. Dentro del término calidad vocal también se estudian tipologías de voces por ejemplo susurrada, ronca, inestable, nasalizada, etc.

laringealización

Uno de los fenómenos más estudiados en este ámbito está dado por la *laringealización* , que se corresponden con fragmentos sonoros de habla durante los cuales se distinguen patrones de excitación irregulares. En estos casos el valor de F0 es aperiódico, o mucho más bajo que para el promedio de la frase. Otra característica común es un descenso en la energía, como de su amortiguamiento.

Otro fenómeno de la calidad vocal está dado por situaciones en que el primer y segundo formante de vocales átonas está centralizado, es decir, se aproxima más a la vocal central que su contraparte acentuada.

La vocal central es la que presenta su primer y segundo formantes aproximadamente en el centro del rango cubierto por el conjunto de las vocales.

De acuerdo a la acústica fonética la estructura formántica tiene como correlato el timbre, y es el resultado obtenido por la interacción de tres factores:

- composición espectral
- forma de los transitorios de ataque y extinción
- número y distribución de las regiones formánticas

Si los armónicos de mayor amplitud son los de menor frecuencia, el timbre será grave, mientras que si los formantes de mayor amplitud son los de alta frecuencia, el timbre resultará agudo.

3.2.3 Velocidad del Habla o Tempo

Se puede considerar que la recíproca de las duraciones de los fonos en cierto intervalo de tiempo es un correlato acústico de la velocidad de habla. En Español se calcula que la velocidad normal de elocución oscila entre las 150 y 200 palabras por minuto.

Es posible encontrar un valor medio y desvío estándar característicos en la duración para cada fono, sin embargo hasta ahora no está claro cómo son percibidos estos atributos. Se pueden hacer mediciones objetivas de la velocidad del habla normalizando la duración de cada fono con respecto a sus valores intrínsecos.

En [20, 28] se mostró que un modelo de duraciones dependiente del contexto es perceptualmente más adecuado cuando se consideran los fonos vecinos y la posición en la palabra (al inicio, en el interior, o al final de las mismas).

3.2.4 Pausas

Se pueden diferenciar dos tipos de pausas: vacías y llenas. Una pausa vacía es un segmento relativamente largo de silencio, ruidos respiratorios, o ruido de fondo; mientras que una pausa llena es un lapso también relativamente extenso conteniendo sonidos de características espectrales uniformes.

Una pausa llena puede contener alguna clase de *schwa*, a veces seguido por un sonido nasal. Durante las pausas llenas el contorno de F0 es plano o presenta un descenso poco pronunciado, y se encuentra en un nivel relativamente bajo. A menudo se pueden encontrar las pausas llenas circundadas por silencios.

3.2.5 Entonación

La entonación es quizás uno de los componentes más complejos de una lengua. De manera simplificada se puede definir a la entonación como el uso distintivo del pitch. Debido a que el valor de base del F0 es dependiente del locutor y viene determinado principalmente por el sexo y la edad, la definición de entonación implica que el término se refiere a cambios en el contorno de la altura tonal respecto a cierto contexto, el cual cubre por lo general más de una sílaba.

Una definición más elaborada se presenta en [141]: *"la entonación es la función lingüísticamente significativa, socialmente representativa e individualmente expresiva de la frecuencia fundamental en el nivel de la oración"*.

Como se vió en el capítulo 1, en términos estrictos, la altura tonal o el *pitch* es el correlato perceptual de la frecuencia fundamental o F0; y en la producción del habla, la frecuencia fundamental está determinada por la tasa a la que vibran las cuerdas vocales.

El rango de F0 para un individuo depende principalmente de la longitud y la masa de sus cuerdas vocales. Para los hombres en condición de conversación normal este rango va típicamente entre 80 y 200 Hz, en tanto que para mujeres el rango se encuentra entre 180 y 400 Hz, mientras que para niños pequeños tal rango puede ser mayor aún que el correspondiente a las mujeres.

Dentro de ese rango cada locutor tiene, hasta cierto punto, un control activo sobre la F0, es decir que puede elegir entre hablar con un tono alto o bajo, y es capaz de producir voluntariamente picos y caídas de tono. Sin embargo, muchos detalles del curso del pitch durante el

habla no son controlados por el locutor sino que son efectos secundarios de otros procesos, a menudo relacionados con la producción de sonidos en particular. Por ejemplo, las vocales altas como /i/ y /u/ tienen un pitch intrínsecamente más alto que las vocales bajas como la /a/. En vocales que siguen a consonantes sordas, el pitch comienza más alto que en vocales que siguen a consonantes sonoras. Estos aspectos involuntarios sobre el tono del habla imponen pequeñas perturbaciones a la trayectoria del pitch, y con frecuencia hacen difícil identificar en un análisis visual aquellas variaciones que son responsables de la melodía percibida en el habla.

En el análisis visual de la curva de F0 correspondiente a un segmento de habla fluida, se pueden observar discontinuidades. Estas se deben a interrupciones de la señal de habla durante la producción de consonantes oclusivas sordas como /k/, /p/ y /t/. El lector debería notar que mientras se oye el habla continua uno no advierte esas interrupciones como pausas en la melodía de la frase.

Si bien esas interrupciones de F0 contribuyen a la caracterización de la consonante percibida, como oyentes tenemos la ilusión que el patrón melódico o entonación del habla es ininterrumpido. De hecho, las interrupciones producidas por este tipo de consonantes son recién percibidas como pausas en la entonación cuando son mayores a 200 ms aproximadamente. Este valor está en sintonía con el umbral de integración perceptual obtenido en pruebas psicoacústicas (porciones de silencio de mayor longitud que dicho umbral impiden la integración perceptual de los sonidos de voz precedentes y siguientes).

Por otro lado, cuando en la señal se produce una modificación abrupta en el valor de F0 después de un intervalo de silencio, el oyente percibe que dentro de ese lapso sordo se produjo un ascenso o descenso gradual del pitch. Es como si en este caso la percepción humana inconscientemente completara la porción faltante e interpolara el contorno del pitch. No es sino hasta que el cambio virtual del pitch se eleva más allá de lo normal que esta ilusión se desvanece, y es suplantada por una desintegración perceptual del flujo de habla.

En sonidos complejos cuasi-periódicos, como el de fragmentos vocálicos el pitch se percibe sobre la base de los intervalos de frecuencia entre los armónicos presentes en la señal. Se puede pensar que existe un procesamiento a nivel central que encuentra un divisor común de un conjunto de armónicos candidatos detectados en la señal. Se ha encontrado que desde el primer tercio al primer sexto de los armónicos presentes en la señal son los más efectivos, constituyendo una región de dominancia para el pitch periódico. El armónico más bajo o frecuencia fundamental no necesita estar físicamente presente para que el pitch se pueda percibir. Para ello basta notar que si la frecuencia fundamental fuera necesaria, la voz masculina normal no tendría pitch detectable en comunicaciones telefónicas, donde las componentes frecuenciales por debajo de 300 Hz generalmente son filtradas [129].

umbral diferencial

La percepción humana del tono en señales con una clara periodicidad es sorprendentemente precisa. El **umbral diferencial** (mínima disparidad que se puede resolver como diferente) está en el orden de 0,3 % a 0,5 % . En tanto para el habla natural, la percepción del pich tie-

ne una precisión que varía en función de la claridad en la periodicidad de la señal.

Tal variación en la claridad cubre el rango que va desde la ausencia de periodicidad, como la que se da en períodos silentes de consonantes oclusivas o fricativas sordas, hasta periodicidad bien definida presente en vocales con suficiente sonoridad, producidas con claras vibraciones de las cuerdas vocales y sin cambios abruptos del pitch, pasando por fricativas sonoras de periodicidad pobremente definidas.

A pesar de esta gran variabilidad se puede asumir que durante el habla natural, para la mayor parte de sonidos sonoros la altura tonal se puede determinar con una precisión del 5 % .

Para el estudio de la entonación, las distancias entre pitch son más relevantes que su valor absoluto. Piense que podemos distinguir una misma melodía aún cuando ésta se entone en una escala o rango tonal diferente (por ejemplo entonada por un hombre o una mujer). Por esta razón en el estudio de la entonación es útil medir el pitch en semitonos en vez de hacerlo en Hertz.

La distancia (d) medida en semitonos entre dos frecuencias f1 y f2 se calcula como:

$$d = 12 \log_2 \left(\frac{f1}{f2} \right) \approx 39{,}86 \log_{10} \left(\frac{f1}{f2} \right) \tag{3.1}$$

Un semitono corresponde aproximadamente a una diferencia en frecuencias de aproximadamente un 6 %.

Sin bien la escala en semitonos es adecuada para cuantificar distancias en la altura tonal, no lo es para cuantificar iguales prominencias preceptuales, resultantes de movimientos de pitch en registros distintos (por ejemplo el de hombres y mujeres). Para ese propósito es más adecuado emplear la escala psicoacústica ERB (escala de ancho de banda rectangular equivalente), presentada en el capítulo 1, que se relaciona con la escala lineal de frecuencias de la siguiente forma:

escala ERB

$$E = 16{,}7 \log_{10} \left(1 + \frac{f}{165{,}4} \right) \tag{3.2}$$

donde

$$f = 165{,}4 \left(10^{0{,}06E} - 1 \right) \tag{3.3}$$

Donde E es la tasa ERB: número de ERBs correspondientes a una frecuencia en particular, y f es la frecuencia en Hz.

Al ser más relevante desde el punto de vista perceptual la distancia relativa de altura tonal que sus valores absolutos, resulta conveniente determinar un umbral diferencial a partir del que se perciban dos valores de pitch diferentes como tales.

Se estima que se puede discriminar con precisión las diferencias de pitch a partir de 3 semitonos. Por lo que uno podría pensar que diferencias menores no serían relevantes para el proceso de comunicación

hablado. Sin embargo, se han reportado estudios que indican que diferencias en el orden de la mitad de ese valor, resultan suficientemente confiables a la hora de distinguir prominencias.

Hasta aquí se vio que aparentemente existen muchos detalles caprichosos en las fluctuaciones del pitch que no son controlados de manera activa por el locutor, sino que son efectos colaterales de otros procesos de la producción del habla. También se asumió que tales movimientos involuntarios del pitch no contribuían a la melodía percibida en el habla .

similitud y equivalencias perceptuales

close-copy stylization

Una primera demostración de este fenómeno es la denominada **copia ajustada estilizada** del pitch (*close-copy stylization* en inglés) , que se define como una aproximación sintética del curso natural del pitch que satisface dos criterios: debe ser perceptualmente indistinguible del original, y debe contener el menor número de segmentos de rectas para lograr tal igualdad perceptual.

En [33] se demostró que aquellas curvas aparentemente antojadizas de pitch que aparecen en el habla natural se pueden simplificar empleando segmentos de líneas rectas en el dominio *tiempo-log(*F0*)*, sin provocar diferencias notables desde el punto de vista perceptual entre las curvas de pitch original y sintética.

Ese hallazgo justifica la descripción de la entonación en términos de aproximaciones más simples. Se debe notar que no hay por qué emplear líneas rectas: por ejemplo se puede encontrar el mismo comportamiento si uno emplea como aproximaciones funciones coseno.

Aparentemente la entonación está organizada en términos de patrones melódicos que son reconocibles por los locutores nativos del lenguaje. Así por ejemplo, si se pide a una persona que imite la entonación de una frase oída, ya sea empleando las mismas palabras, palabras diferentes o incluso sin articular palabras, se obtiene una curva de pitch que en término de señales seguramente no será igual a la original. Aunque incluso se puedan oír muchas diferencias, es probable que generen la misma impresión melódica sobre oyentes nativos. Es de aquí que surge el término **equivalencia perceptual**.

equivalencia perceptual

Dos curvas de F0 diferentes son perceptualmente equivalentes cuando una es lo suficientemente parecida, como para ser juzgada por hablantes nativos de esa lengua como una buena imitación melódica de la otra.

Esta equivalencia implica que la misma melodía puede reconocerse en dos realizaciones aunque en ellas existan diferencias notables, de la misma forma que dos palabras pueden reconocerse como iguales a pesar de ser realizaciones diferentes. Esta noción permite construir, a partir de generalizaciones, un inventario de movimientos estándar del pitch, y sus combinaciones generan contornos que son perceptualmente equivalentes a las curvas de pitch que ocurren en habla natural.

declinación

La **declinación** es un fenómeno que aparece en muchos lenguajes. Su forma ininterrumpida está restringida a frases cortas, ya que en elocuciones largas (en especial en habla espontánea), se encuentra interrumpida por reinicios regulares en la declinación. Se cree que el fenómeno de declinación no se encuentra bajo control voluntario del locutor, pero sí la elección del momento en que se introducen los reini-

cios. Otros autores creen que al menos en parte la declinación es voluntaria.

Los locutores prefieren hacer coincidir el instante de reinicios en fronteras importantes de la estructura de constituyente prosódicos. En el análisis de habla continua, estimar la curva de declinación a partir del perfil de F0 no resulta una tarea sencilla. Principalmente por las fluctuaciones del pich, la variabilidad de la magnitud de sus movimientos, y la presencia de reinicios en la declinación.

En [165] se sugiere que la declinación se debería estimar haciendo primero una copia estilizada de la curva de entonación y luego tratando de reemplazar los fragmentos de pitch relativamente bajo por líneas rectas que constituyan una línea de base tentativa. La inclinación de esta línea de base podría entonces dar una estimación de la declinación y generar en la resíntesis el mismo resultado perceptual que el de la declinación original. De acuerdo a ese mismo trabajo, las tres líneas paralelas en la figura 28 se denominan niveles de pitch (*topline*,*midline* y *baseline*), y se emplean como referencias para definir virtualmente todos los ascensos y descensos perceptualmente relevantes.

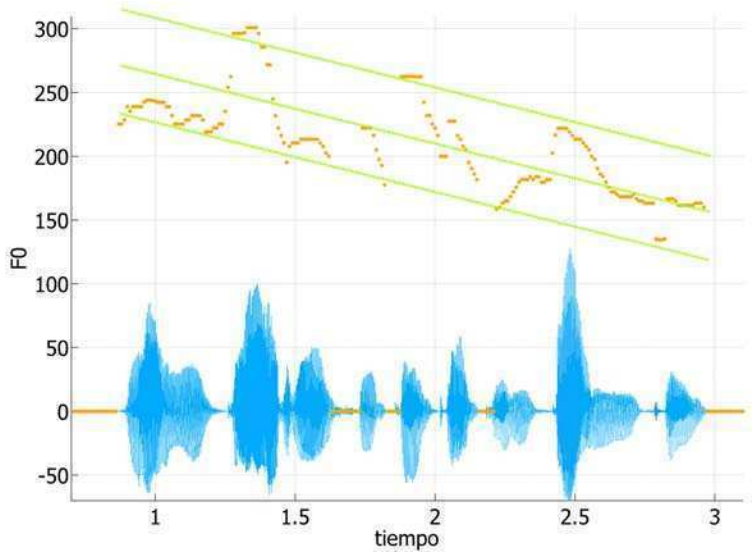

Figura 28: Forma de onda (azul), F0 (naranja), y niveles de pitch (verde) correspondientes a la frase: *La guitarra se toca con pánico*.

Debido a que los niveles de pitch son equidistantes en el dominio *tiempo-log(*F0*)*, se pueden fijar las distancias en semitonos. Para Inglés Británico una distancia de 12 semitonos entre la *baseline* y la *topline* dan resultados satisfactorios, para el alemán esta distancia es de 7,5.

Se debe notar que al fijar estos valores se hace una reducción considerable de la variabilidad de fluctuaciones de pitch. Se puede pensar que para distintos estilos de habla se deberían emplear diferentes distancias entre los niveles de pitch.

Tales cambios sin embargo no deberían modificar la equivalencia perceptual en términos de patrones melódicos reconocibles del lenguaje.

Ahora los movimientos de pitch se pueden describir como movimientos más o menos rápidos de un nivel de pitch a otro. Cada movimiento de pitch estándar se puede caracterizar completamente por su dirección (ascendente o descendiente), tamaño (número de semitonos cubiertos por la variación de pitch), tasa de cambio (en semitonos por segundo), y su ubicación temporal (en milisegundos después del comienzo de la sílaba o antes del fin de sílaba).

El número y la caracterización de los movimientos de pitch perceptualmente relevantes difieren de un lenguaje a otro. Para el Danés hay 10 movimientos perceptualmente relevantes, para el Alemán 11 y para el Inglés Británico 27 [129].

Este número máximo de movimientos está acotado por las posibles combinaciones para cada una de las dimensiones que los caracterizan a nivel de una sílaba. Para la dirección este número es 2 (ascendente o descendiente), para el tamaño se ha estimado en que no se pueden distinguir como separados más de 3 o 4 intervalos, para la tasa de cambio dentro de los límites de una sílaba se cree que es distinguible con respecto a un pitch constante solo un ascenso (y posiblemente un descenso). Además, parecen ser distinguibles respecto a la tasa de cambio sólo ascensos graduales que permanecen durante varias sílabas. Finalmente, con respecto al momento en que se registran estos movimientos, dentro de una sílaba de 200 ms de duración se pueden considerar como separadas a lo sumo tres posiciones distintivas.

Así, el número máximo de movimientos abruptos del pitch, que ocurren dentro de una sílaba parece estar en el orden de $(2x4x1x3 = 24)$. Este número se incrementa a medida que se consideran unidades de análisis mayores que al nivel de la sílaba (en estos casos se deberían contemplar ascensos y caídas graduales).

Perceptualmente, no todos los movimientos de pitch tienen el mismo efecto. Algunas subidas y descensos sirven como realizaciones de acentos tonales, dando prominencia perceptual a determinada sílaba, o confiriendo la idea que la emisión continuará.

Para caracterizar los patrones distintivos que presentan las curvas de altura tonal a nivel global se suelen utilizar por ejemplo tonos bajos o altos, o movimientos descriptos como descensos o cadencias, ascensos o anticadencias, y sus combinaciones: descenso-ascenso y ascenso-descenso, así como un tono conocido como ascenso de continuación (*continuation rise* en Inglés). Este último patrón denota la situación en que el valor de la curva de F0 presenta un ligero ascenso o un nivel superior al F0 de base de la oración.

El valor de F0 de base se define como el mínimo valor que presenta esa variable durante una elocución completa. Si bien el empleo de esa definición hace sencillo su estimación a partir del contorno de F0, tiene la desventaja que en algunos casos se suele utilizar un tono deliberadamente bajo para marcar acentos en la oración, y esos valores extremos no deberían se considerados el F0 de base. Por ello resulta preferible emplear una definición alternativa, y definirlo como el valor de F0 promedio registrado en las sílabas no marcadas de una elocución. Si bien esta última definición puede ser más precisa, su cálculo no resulta simple, y muchos autores optan por aproximar su valor empleando la mediana del pitch para el segmento de habla considerado.

Finalmente, según la utilización lingüística del tono, las lenguas se dividen en tonales y entonativas. Las **lenguas tonales** como el Chino o el Tailandés, utilizan los tonos para distinguir significados. Cumple una función en la distinción léxica.

lenguas tonales

Por su parte las **lenguas entonativas** no utilizan la sucesión de tonos para distinguir significados léxicos, sino para modificar significaciones secundarias como expresividad o intencionalidad. Cumple una función expresiva en la frase.

lenguas entonativas

A este tipo de lenguas pertenecen todas la románicas.

3.2.6 Acentuación

El término acentuación se utiliza para indicar la presencia de algunas sílabas que son percibidas como más prominente que sus vecinas. Es así que el acento es un rasgo suprasegmental que recae sobre una sílaba de la cadena hablada y la destaca o realza frente a otras no acentuadas (o átonas).

En la literatura clásica se vinculó esa percepción de prominencia con la idea que las sílabas tónicas presentaban mayor intensidad que las átonas, y por lo tanto eran producidas con mayor esfuerzo articulatorio. Es por ello que se conocen a las sílabas tónicas como *estresadas* o tensas.

Sin embargo hoy se sabe que la impresión perceptual de prominencia no está asociada solamente con la intensidad sino que es el resultado de la conjunción de varios factores articulatorios:

- una mayor fuerza espiratoria, que genera una mayor intensidad
- una mayor tensión de las cuerdas vocales, que genera una elevación del tono fundamental
- una mayor prolongación en la articulación de los sonidos, que supone un aumento de la duración silábica

Vale hacer una aclaración respecto a la distinción entre los términos *stress* y *accent* que se se emplean con mucha frecuencia en la literatura técnica. De acuerdo a [98] la palabra inglesa *stress* está referida a la prominencia percibida en alguna entidad léxica (palabras o sílabas) del habla continua, mientras que el término *accent* se reserva para referir específicamente a movimientos entonativos de F0. Estos movimientos de F0 pueden servir como uno de los indicadores fonéticos respecto a la localización de la prominencia percibida. Es decir no hay que confundir el término acento como se está tratando en este apartado con el término anglosajón *accent*.

Para confundir más las cosas, en Español también se utiliza el término acento para referirse al **acento ortográfico** , marca tipográfica (tilde) que se coloca sobre alguna vocal con acento léxico.

acento ortográfico

Por otro lado, cuando se observan palabras pronunciadas de manera aisladas, se puede observar que presentan al menos una sílaba acentuada (obviamente en palabras con más de una sílaba). A este acento se conoce como **acento léxico** . A diferencia de otros idiomas como el Inglés o el Alemán, en el Español se puede conocer exactamente cuál

acento léxico

es la sílaba con acento léxico de una palabra aislada empleando un conjunto de reglas ortográficas.

La sílaba con acento léxico se denomina sílaba tónica, y las que no lo tienen átonas. También se habla de *sílabas pretónicas* y *postónicas* para referirse respectivamente a las que preceden o siguen a la tónica.

En el Español todas las palabras aisladas poseen una sílaba tónica, excepto por los advervios terminados en *-mente* que presentan dos, por ejemplo: /**SO**-la-**MEN**-te/ *solamente*. Además las palabras compuestas extensas pueden presentar *acentos secundarios* además del principal.

acento de frase

El último acento principal de una frase se denomina **acento de frase** , y generalmente ocurre en la sílaba acentuada de la última palabra de contenido de una porción de habla.

En muchos casos el acento léxico permite diferenciar palabras. Una gran cantidad de idiomas presentan algunos grupos de palabras, denominados **pares mínimos** , que desde el punto de vista segmental son idénticas pero se diferencian por la ubicación del acento léxico. Por ejemplo, la siguiente figura muestra la forma de onda, energía y F0 de tres palabras similares desde el punto de vista fonético, pero desambiguadas a partir de sus acentos léxicos:

pares mínimos

Algunas lenguas se caracterizan por admitir el acento léxico en una sola posición dentro de la estructura silábica. A estas se las conoce como **lenguas de acento fijo** . Entre ellas, por ejemplo se encuentran el Francés y el Turco en las que el acento recae siempre en la última sílaba de la palabra, en el Checo y Finés en la primera sílaba, mientras que en el Polaco lo hace en la penúltima sílaba. Por su parte hay otro grupo de lenguas en las que el acento puede ocupar distintas posiciones dentro de la palabra y se las denomina **lenguas de acento libre** . El Español, como el Italiano, Inglés y Alemán son lenguas de acento libre.

lenguas de acento fijo

lenguas de acento libre

En un estudio realizado sobre habla continua para el Español se reporta que el 63,44 % de las palabras están acentuadas, mientras que el 36,56 % aparece inacentuadas, [141].

Es decir, si bien que toda palabra aislada de al menos dos sílabas presenta una sílaba acentuada, cuando estas palabras se combinan en el habla continua, algunos de dichos acentos no se expresan (salvo en el caso de habla contrastiva o enfática).

Aquí se puede encontrar la otra noción distintiva de acento, conocido como **acento tonal** .

acento tonal

El acento tonal, aparece en el habla continua de manera relativamente libre, cuando se percibe una palabra con prominencia por encima de las demás. Este acento se aplica dentro del contexto de una frase a determinadas palabras (principalmente las de contenido) para indicar foco o que se trata de información nueva. El correlato acústico tradicional de este acento está dado por el contorno local de la frecuencia fundamental. Por ejemplo, a la pregunta *¿Dónde está la casa?* puede responderse *"la casa está en la montaña"*. En esta respuesta la palabra *montaña* recibe un acento tonal expresado físicamente como un ascenso de la frecuencia fundamental en la sílaba *"ta"*. Debe notarse que el acento tonal se ubica temporalmente en sincronía con el acento léxico de la misma palabra. En el ejemplo anterior la palabra *"casa"* tiene acento léxico en la silaba *"ca"* pero no posee acento tonal.

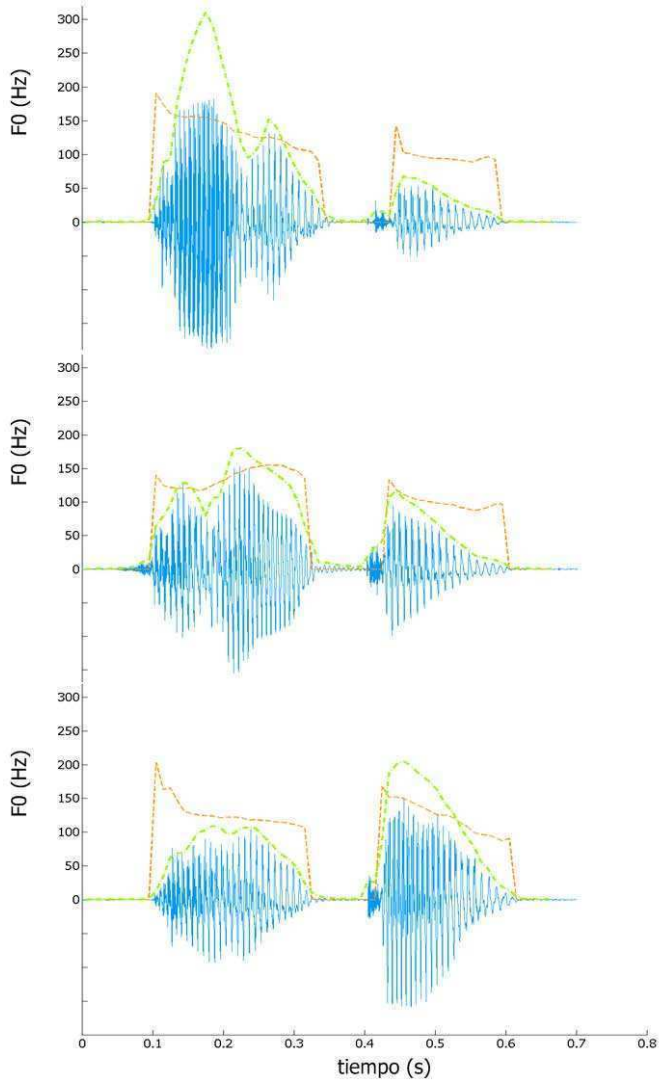

Figura 29: Forma de onda (azul), F0 (naranja), y energía (verde) para tres palabras segmentalmente equivalentes, pero con distinto acento léxico. Arriba: *hábito* (esdrújula), Centro: *habito* (grave), Abajo: *habitó* (aguda).

En el Español los acentos tonales se ubican siempre donde existe un acento léxico [98]. Es decir, los acentos léxicos son reservorios potenciales del acento tonal. Sin embargo existen casos en los que no se respetan las reglas mencionadas anteriormente. Por ejemplo cuando la sincronía entre acento léxico y tonal no es exacta. Existen leves defasajes entre el pico de la frecuencia fundamental y las fronteras de la sílaba acentuada. Y más aún, puede suceder que una sílaba con acento léxico reciba un acento tonal, pero que la frecuencia fundamental no tenga un valor alto o aumentado (denominado H*).

Además en las frases de habla continua es común que el contorno de entonación de la frase predomine sobre los acentos tonales imponiendo un descenso de la frecuencia fundamental en el final de una frase

afirmativa. En esta situación, el acento lexical se manifiesta por un aumento de los parámetros restantes (duración y energía), y el acento tonal recibe la categoría de tono bajo (L*) [26].

Se puede resumir las diferencias entre acento léxico y tonal de la siguiente forma:

- Acento léxico: acento de palabra determinado por reglas de ortografía (abstracto).
- La mayoría de las palabras en Español poseen un acento léxico, que corresponde a la vocal fuerte de la sílaba acentuada.
- Acento tonal: acento manifestado en el habla continua y determinado por factores rítmicos, posicionales, cuantitativos y morfológicos.
- En muchos casos ese acento léxico permite diferenciar palabras, por ejemplo: papa - papá.
- En Español los acentos tonales generalmente se ubican donde existe acento léxico.

De acuerdo a [141], el acento desempeña en la lengua española tres funciones: contrastiva, distintiva y culminativa.

La función contrastiva del acento le permite distinguir entre sílabas acentuadas e inacentuadas, y entre palabras acentuadas como sustantivos, verbos, etc., y no acentuadas tales como preposiciones, conjunciones, etc.

tipologías acentuales

La función distintiva del acento da origen a las **tipologías acentuales** : aguda (u oxítona), grave (oxítona), esdrújulas (proparoxítonas), y en las formas compuestas a las sobreesdrújulas (superproparoxítonas).

A través de su función culminativa el acento aglutina alrededor de la unidad central, otras unidades inacentuadas.

Finalmente, se debe notar que la acentuación no es una medida perceptual binaria, sino que presenta diferentes niveles o graduaciones.

Ritmo

La impresión auditiva de ritmo surge a partir de las distribuciones de duraciones que presentan las unidades del habla.

Tradicionalmente se postuló que los lenguajes presentan una propiedad de isocronía, que se puede definir como la organización del habla en segmentos que se perciben con duraciones iguales o equivalentes. Se podría pensar a estos intervalos como la extensión de los compases en teoría de la música.

ritmo silábico

ritmos de temporización acentual

A partir de la publicación [136] se sugirió la posibilidad de distinguir dos tipos de ritmo en los lenguajes: ritmos de **temporización silábica** , en los que las sílabas generan la impresión de tener siempre la misma duración, el ritmo se puede caracterizar por el número de sílabas por unidad de tiempo; y por otro lado ritmos de **temporización acentual** , donde se percibe a las sílabas acentuadas ocurren a iguales intervalos de tiempo, sin importar el número de sílabas átonas que se presenten

entre ellos. En este último caso el ritmo se determina considerando el número de sílabas acentuadas por unidad de tiempo.

Más tarde en [1] se reivindicó esa hipótesis como una distinción tipológica universal, y se propuso que todos los lenguajes podrían ser ordenados de acuerdo a esa distinción. Sin embargo, un tiempo después se propuso una tercera categoría: los lenguajes con **temporización empleando mora** (unidades sub-silábicas), para idiomas como Japonés o Tamil [[99]].

temporización empleando mora

Si bien esta clasificación tipológica ha sido empleada durante mucho tiempo, existen muchas investigaciones empíricas cuyos resultados refutan este comportamiento, aún para lenguas que fueron presentadas como paradigmas de alguno de los tres grupos propuestos.

Un estudio más reciente [144] se basó en estudios psicolingüísticos que demuestran que los niños son capaces de discriminar entre oraciones tomadas de su lenguaje materno y otro lenguaje de clase rítmica diferente, pero no entre pares de oraciones tomadas de lenguajes de otra tipología rítmica distinta a la nativa. A partir de esa observación los autores intentaron determinar la identidad de los correlatos del ritmo en la señal de habla. Encontraron que dos variables: $\%V$, el porcentaje total del tiempo de la frase correspondiente a las vocales, y δC, el desvío estándar de las duraciones de segmentos correspondientes a consonantes correspondiente a la oración eran suficientes para separar los tres tipos de tipologías acentuales.

En [179] se propuso una teoría para patrones rítmicos en el habla espontánea para el Inglés. En ese estudio se identificaron una jerarquía de patrones rítmicos incluyendo la repetición de acentos de amplitud (beats) además de acentos tonales, ambos a intervalos regulares de tiempo. Los autores establecen que por razones rítmicas, aunque una palabra no sea importante desde el punto de vista semántico puede recibir acento.

Hesitación

Este atributo difiere de los descriptos previamente debida a que no contribuye al significado de una frase sino que indica un conflicto entre la planificación y la producción del habla.

Señala que el locutor se encuentra pensando, pero desea continuar. Por lo tanto es una forma de evitar la interrupción del interlocutor.

Los atributos prosódicos básicos que caracterizan al balbuceo o hesitación son las pausas y el alargamiento exagerado de sílabas y palabras.

3.2.7 Interacciones de los Parámetros Prosódicos

No resulta sencillo determinar la contribución relativa de cada rasgo prosódico observado a través de sus manifestaciones físicas, en las variaciones funcionales de la prosodia. Generalmente una función específica como la demarcación del acento, o la delineación de junturas no está definida por un solo parámetro prosódico, sino por una combinación de éstos. Además la contribución exacta de cada parámetro varía en función del contexto. Ilustremos la complejidad del fenómeno.

Cada función prosódica está marcada por la combinación de rasgos prosódicos, dependientes del contexto.

En lenguajes cuya estructura rítmica viene dada principalmente por el estrés o patrones de prominencia (como el Inglés o Alemán), la sílaba acentuada fonológicamente dentro de la palabra muestra una tendencia de mayor duración, mayor tono (generalmente en ascenso), e intensidad que las sílabas no acentuadas circundantes. Por supuesto que estas son solamente tendencias, ya que por ejemplo el criterio de mayor duración puede fallar. Aún así muchos de los casos de falla generalmente son predecibles, se observa que la vocal acentuada puede ser más corta que las circundantes en los siguientes casos: cuando la vocal acentuada es una vocal intrínsecamente corta y se encuentra rodeada por vocales intrínsecamente largas, cuando la vocal acentuada pertenece a una sílaba con un gran número de fonemas en relación al número de fonemas de sílabas vecinas (cuanto mayor sea el número de unidades en un nivel lingüístico, menor es la duración de cada unidad), o cuando se encuentre cercana a una sílaba que ha sido alargada por ser final de palabra, o de frase.

El criterio de mayor intensidad de la vocal acentuada puede fallar cuando ésta es por naturaleza de baja intensidad, y cuando está rodeada por vocales de mayor intensidad intrínseca.

Generalmente se suele encontrar un movimiento de F0 en cada sílaba acentuada. Dependiendo de la posición de la palabra en la frase, la morfología de la curva de F0 puede ser ascendente o descendiente durante la sílaba acentuada. Cuando dos palabras contiguas se agrupan en la misma palabra prosódica, se da una tendencia de ascenso de la curva de F0 durante la primer palabra y de descenso durante la segunda. Esta tendencia es independiente del lenguaje. Por otro lado, el ascenso de la F0 que generalmente es concomitante con la ocurrencia de una sílaba acentuada, también se puede encontrar en sílabas no acentuadas, como en la realización de un ascenso suspensivo en la vocal final previa a una pausa intermedia. También se puede destacar la tendencia general de un ascenso en la curva de F0 al comienzo de las frases, y luego una declinación progresiva a lo largo de la cual también va disminuyendo el rango de excursión de la F0.

La altura relativa de las sílabas acentuadas y la amplitud de los movimientos tonales se deben interpretar de acuerdo a la posición de la sílaba en la frase. Los mayores movimientos de F0 se esperan en el principio de la frase. Si se encuentran los mayores movimientos en otra parte de la frase, es muy probable que se trate de palabras enfatizadas. Los menores movimientos de F0 se esperan al final de la frase, en caso que esto no ocurra, es probable que la frase esté marcada (puede ser interrogativa).

Todo esto se puede sintetizar en los siguientes puntos:

1. **Interacción de atributos prosódicos.**

   Para una determinación confiable de una función prosódica dada se debe considerar, aún para el caso de palabras aisladas, una combinación simultánea de al menos tres parámetros (tono, duración e intensidad).

2. **Dependencias contextuales.**

Al relacionar las manifestaciones físicas observadas en las señales con las variaciones funcionales de la prosodia se deben considerar la posición de la unidad acústica estudiada dentro de la estructura de frase. Especialmente, se deben utilizar reglas contextuales para realizar esa determinación en habla continua.

3. **Funciones múltiples de rasgos prosódicos.**

   Un evento prosódico como el alargamiento vocálico o el ascenso de F0 podría corresponderse con más de una interpretación.

Normalizaciones Empleando Consideraciones Articulatorias y Perceptuales.

Para analizar los parámetros prosódicos algunos algoritmos intentan sustraer el efecto de las variaciones condicionadas fonéticamente. Tal normalización requiere la identificación de los sonidos, y posiblemente su segmentación silábica.

Una compensación ideal debería considerar al menos los siguientes aspectos:

- Las características intrínsecas de los fonemas y la influencia de los fonemas circundantes.
- El número de fonemas en la sílaba, el número de sílabas en la palabra, y de palabras en la frase.
- La velocidad de elocución.
- La corrección del alargamiento prepausal.

Los sistemas actuales incluyen solo compensaciones parciales de las variaciones condicionadas fonéticamente, justamente aquellas que son más simples de integrar. Es más, no queda aún claro si vale la pena o no aplicar esas compensaciones parciales. De hecho, una compensación completa solamente sería viable en el proceso de verificación de las hipótesis de reconocimiento obtenidas.

También se ha sostenido que las compensaciones de las variaciones producto de las características fonéticas no son suficientes, sino que se debería agregar conocimiento sobre la percepción del habla. Es evidente que tal normalización mediante consideraciones perceptuales debería conducir a una visión más integrada de la contribución relativa de los tres parámetros prosódicos principales. No obstante aún no se logró un nivel de conocimiento suficiente para establecer cómo se perciben y procesan la intensidad, F0 y duración en el habla continua.

3.3 JERARQUÍA PROSÓDICA

El lenguaje representa un mapeo entre sonido y significado. En una visión minimalista, Chomsky simboliza al lenguaje como un sistema computacional que genera representaciones internas que son mapeadas a la interfase senso-motora por un lado y al sistema conceptual-intencional por el otro [23].

En esta visión, muchas de las propiedades del lenguaje hablado no derivan del componente sintáctico del lenguaje sino de las condiciones de interfase entre el procesamiento del núcleo generativo y el sistema de salida. Esto implica que las reglas que gobiernan los cálculos sintácticos pueden estar desligados de los que controlan el habla.

Esa misma conclusión se alcanzó en el campo de la fonología a finales de 1970, estudiando la organización del lenguaje hablado. Se encontró que las reglas sintácticas resultaban insuficientes para explicar la estructura organizativa observada en las frases del habla [104]. En la figura 30 se muestra a manera de ejemplo la organización prosódica de una oración de acuerdo a las terminología propuesta en [127].

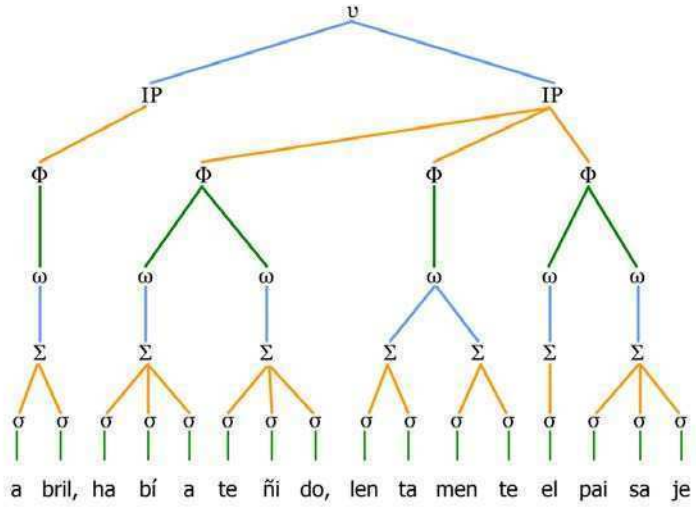

Figura 30: Organización prosódica de la frase: *"Abril había teñido lentamente el paisaje"*

Existen al menos dos elementos que vale la pena considerar en esta figura. La primera es la naturaleza jerárquica de los constituyentes. La segunda es que en ningún lugar se mencionan nociones sintácticas como verbo, sustantivo, frase verbal, etc. Es decir que los constituyentes mostrados en la figura no se encuentran en un dominio sintáctico sino prosódico.

Según la teoría prosódica, la representación mental del habla se divide en segmentos ordenados jerárquicamente: los dominios prosódicos. La figura 31 presenta la jerarquía de los constituyentes prosódicos adaptados de [127] en [158].

Se puede ver en la figura 31 que el menor constituyente considerado es la sílaba. Los dominios prosódicos presentados tienen las siguientes propiedades:

1. Los segmentos constituyen el ámbito de aplicación de reglas fonéticas y fonológicas.

2. Estos segmentos de dominio prosódicos no son necesariamente isomórficos con los segmentos generados por análisis morfosintáctico, o cualquier otro proceso gramatical.

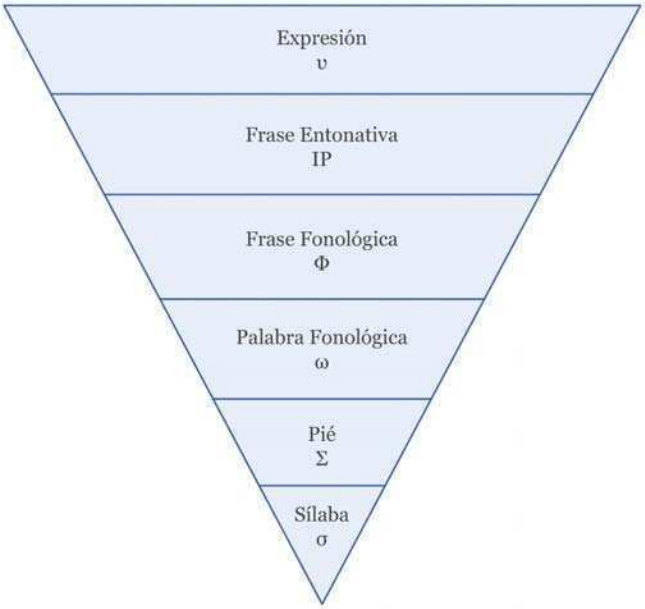

Figura 31: Jerarquía de los constituyentes prosódicos

La primera propiedad puede verse como el proceso de descubrimiento y delineación de constituyentes prosódicos, que están dados por segmentos de habla dentro de los que son válidas ciertas reglas fonológicas; mientras que la segunda implica que dichos segmentos son constituyentes prosódicos independientes, sin una necesaria correlación con otros componentes de la gramática, como la sintaxis o la morfología.

La jerarquía prosódica se sostiene en un conjunto de reglas y restricciones [152, 127]. Dos de esas reglas pertenecientes a la jerarquía prosódica fueron formalizadas en el marco de la **teoría de optimalidad** [140]:

teoría de optimalidad

- Exaustividad: cada constituyente de nivel l está contenido en un constituyente de nivel $l+1$. Por ejemplo una ω está contenida en una ϕ.
- No-recursividad: ningún constituyente de nivel l puede estar contenido en otro constituyente de nivel l.

Varios estudios muestran que esas reglas son aplicables a diferentes niveles de la representación prosódica, tales como [97, 46] para *IP*, [177, 68] para ϕ; y [187] para ω.

Cuando surgió la propuesta de constituyentes prosódicos, se consideraba que la estructura sintáctica era quien dominaba la distribución y división de los estos constituyentes, y especialmente con respecto a las *IP* y ϕ [127], aunque se afirmaba que la velocidad de elocución, el estilo y las emociones podrían provocar reestructuraciones de *IP* en otras *IP* más cortas.

Estudios más recientes muestran que en muchos lenguajes la ubicación de los límites prosódicos responde a otros aspectos además de la sintaxis como una distribución simétrica de los constituyentes, ponderación de los constituyentes, y estructura de la información [30].

A continuación se describirán con mayor detalle cada uno de los constituyentes prosódicos mencionados.

3.3.1 Sílabas (σ)

Desde la fonología, la sílaba puede ser vista como el dominio de ciertos procesos o restricciones fonológicas. Estructuralmente, una sílaba está compuesta al menos por una vocal, y adicionalmente puede tener una o más consonantes. Se puede considerar que las sílabas están organizadas de acuerdo a la estructura mostrada en la figura 32:

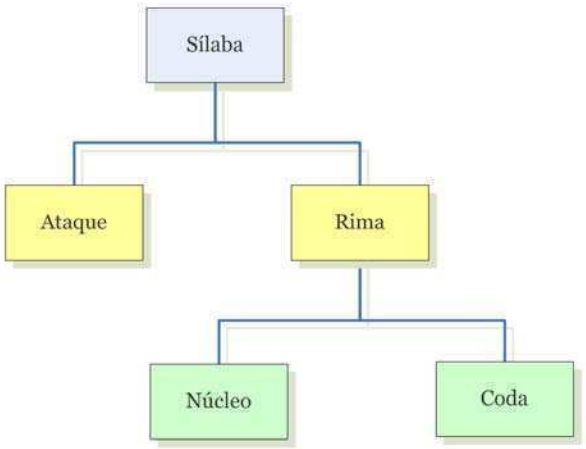

Figura 32: Organización estructural de las sílabas

El núcleo es la parte más sonora, de los segmentos constituyentes de una sílaba. Para el Español (como para la mayoría de los lenguajes) dicho núcleo está conformado por un sonido vocálico, y suele denominarse núcleo vocálico. Dentro de la sílaba la sonoridad suele disminuir a medida que nos alejamos del núcleo, lo que implica que para cada sílaba existe a lo sumo un pico de sonoridad. Es decir que se podría estimar de manera aproximada el número de sílabas de un fragmento de habla empleando el número de picos de sonoridad presentes en la misma.

Se puede realizar una generalización inter-lingüística respecto a la estructura silábica [158]:

- **a.** El ataque puede ser opcional, pero en ningún lenguaje la estructura silábica proscribe la presencia del mismo.
- **b.** El núcleo es obligatorio y el segmento más sonoro.
- **c.** Algunos lenguajes prohíben la coda y otros lo dejan opcional.

- **d.** Algunos lenguajes permiten la existencia de sílabas complejas, admitiendo que cada constituyente de la sílaba pueda ramificarse, es decir tener más de un segmento.

Considerando de manera conjunta **a**, **b** y **c**, queda claro que la sílaba más pequeña es la conformada por una sola vocal. Asimismo de **a** se puede ver que no se conoce lenguaje en el cual el ataque esté completamente ausente de la estructura silábica. En consecuencia la estructura CV (consonante-vocal) se puede encontrar como sílaba en todos los lenguajes. Varios autores han sugerido que la sílaba CV representa la estructura precursora del habla moderna [115].

En psicolingüística se considera a la sílaba como el bloque fundamental tanto en la producción como en la percepción del habla [102, 117].

3.3.2 Pie Métrico (Σ)

El nivel de la jerarquía prosódica correspondiente a Σ se refiere a la unidad suprasilábica menor que una palabra que ayuda a describir los patrones de acentos prosódicos.
Sin embargo, ya en [152] se reconoció que existen pocas evidencias que esta unidad sea un dominio relevante para las reglas fonológicas.

Las sílabas se agrupan entre sí en pies. Cada pie (Σ) se puede considerar conformado por una sílaba relativamente fuerte y cualquier número de sílabas más débiles [127].
La ubicación precisa de la sílaba fuerte depende de factores específicos del lenguaje. El Σ determina la ubicación del acento secundario, que típicamente recae sobre la sílaba fuerte.
Esa estructura de Σ refleja algunas características de los constituyentes prosódicos:

1. Los constituyentes prosódicos de nivel Xp se forman uniendo en un nodo n-ario todos los constituyentes de nivel $Xp - 1$ incluidos en el fragmento de habla delimitado por la definición del dominio de Xp.

2. La relación de prominencias relativas definida para nodos hermanos es tal que a un nodo se puede asignar un valor de fuerte y a todos los demás un valor débil.

3. Una unidad perteneciente a un nivel de jerarquía está completamente incluido en el constituyente superordenado del que es parte.

4. Cada unidad no terminal de la jerarquía prosódica está compuesta por una o más unidades de la categoría inmediata inferior.

5. Un límite en un nivel particular de la jerarquía prosódica implica un límite al menos en un constituyente por cada uno de los niveles inferiores.

Se puede entender 1 considerando que una unidad de nivel no terminal de la jerarquía prosódica está constituido por el arreglo lineal

de unidades del nivel inmediato inferior contenidos dentro de su dominio. Por ejemplo cada Σ está constituido por una o más σ que se encuentren dentro de su dominio.

El principio descripto en 3 excluye dentro de la organización prosódica casos en los que por ejemplo una σ esté contenida en más de un Σ.

El principio 4 se conoce también como *Hipótesis de Estrato Estricto*. Implica que cada Σ está compuesto por σs y no por μs (fonemas). Una consecuencia de esta hipótesis es que los constituyentes prosódicos no pueden mostrar recursión.

El principio 5 es consecuencia de 3 y 4, e implica que un constituyente más alto en la jerarquía prosódica debe ser co-extensivo con al menos una unidad de todos los constituyentes inferiores de la jerarquía. Por lo tanto los límites de grupos clíticos, frases fonológica y frases entonativas también son límites de palabras fonológicas.

3.3.3 Palabra Fonológica o Palabra Prosódica (ω)

En el Español escrito se reconoce una palabra como un texto circundado por espacios y/o signos de puntuación.

Una palabra es el vínculo entre un patrón sonoro y cierto significado evocado en la mente del oyente de un lenguaje. Como se vió en el capítulo 1, en lingüística se considera que la menor unidad que conserva un significado es el morfema. Por ejemplo la palabra *"poemas"* está constituida por dos morfemas: la raíz *"poema"* y el sufijo *"s"* que indica pluralidad. En morfología, una palabra consiste en una raíz y sus afijos.

En la teoría prosódica el constituyente que tiene mayor correspondencia con la palabra morfológica es la palabra fonológica ω. De acuerdo a [127] la palabra fonológica es el menor constituyente de la jerarquía prosódica que refleja una relación íntima entre fonología y morfología.

La contribución precisa de la información morfológica al constituyente prosódico ω varía de lenguaje en lenguaje. Sin embargo lo que se verifica entre los diversos lenguajes es que los ω se corresponden como máximo con una raíz léxica y sus afijos. Esto implica que los límites de ω también delimitan uno o más morfemas. En las teorías de interfases sintáctico fonológicas basadas en límites, los márgenes de las raíces lexicales se alinean con los de constituyentes fonológicos.

Las ω constituyen un concepto relevante para la fonología. Desempeña un rol métrico al describir el acento vinculado a una palabra. Un elemento léxico se puede considerar como una ω si está acentuada prosódicamente.

Con respecto a las variables acústicas que marcan la presencia de una ω en el Español, se pueden encontrar varias opiniones diferentes. Estudios clásicos como [126] argumentan que la intensidad es el indicador más confiable del acento prosódico.

Investigaciones más recientes [52, 141, 53, 42] determinaron que un cambio en el nivel de F0 producido durante el lapso correspondiente a

las sílabas tónicas son los mejores indicadores de la prominencia. Sin embargo, en posiciones nucleares de ciertas frases, donde se observa un valor de F0 reducido, la prominencia se marca por otros medios, como alargamientos de fin de frase. En [24] se formalizó la *regla de acento nuclear*, que se aplica frecuentemente a constituyentes de lenguajes románicos como el Español [138, 82]. Se asume que mientras los movimientos de F0 se asume que son indicadores de ω en posiciones prenucleares, su presencia en posición nuclear está dada por otros correlatos.

Con respecto a los movimientos de F0, la excursión mínima de pitch asociada con la presencia de ω es de 7 Hz [133].

En [141] se da una lista de tipos de palabras acentuadas e inacentuadas, así como estadísticas al respecto. Las acentuadas, principalmente palabras de contenido son aquellas que se esperan estén acentuadas debido a un movimiento del F0 en la sílaba tónica. Por su parte las palabras inacentuadas, principalmente palabras de función son aquellas en que no se espera encontrar movimientos de F0 asociados. Sin embargo factores como la velocidad de habla o el énfasis pueden generar discrepancias con respecto a esas reglas.

Finalmente en [134] encontraron que en los contextos donde el acento prosódico no está indicado por un acento tonal, presenta como correlatos diferencias en duraciones, tilt espectral, y calidad vocal en un grado menor.

En resumen, mientras se usa el F0 como el correlato más prominente del acento y ω, factores tales como duración, intensidad, y tilt espectral, merecen mayores investigaciones respecto a sus asociaciones con los acentos prosódicos.

3.3.4 Frase Fonológica (ϕ)

Las frases fonológicas son similares a las frases intermedias de la fonología entonativa[98]. Ambas se entienden como frases menores contenidas en las *IP*.

Algunos lingüistas argumentan que para el Español ese segundo nivel de frase no es necesario: [160, 12, 88], mientras que por ejemplo en [128] se justifica perceptualmente dos niveles de frases demostrando que el segundo nivel se utiliza para desambiguar semánticamente fragmentos de habla sintácticamente idénticos. En [81, 172, 173] también se reconoce la importancia del segundo constituyente de frases.

De acuerdo a [178] las diferencias entre las ϕ e *IP* están dadas porque las primeras hacen referencia específicamente a frases sintácticas (XPs), como frases nominales (NP), verbales (VP), y adjetivas (AP), mientras que las *IP* están vinculadas con proposiciones sintácticas mayores.

Las frases fonológicas aparecen estrechamente vinculadas y condicionadas por la sintaxis. La estructura sintáctica de un lenguaje está conformada por sintagmas. El **sintagma** es una clase de constituyente sintáctico conformado por un grupo de palabras que a su vez puede contener otros sub-constituyentes, al menos uno de los cuales es un **núcleo sintáctico** . Las propiedades combinatorias de un sintagma se

sintagma

núcleo sintáctico

derivan de las propiedades de su núcleo sintáctico, en otras palabras, *"un sintagma es la proyección máximal de su núcleo"*.

El núcleo sintáctico corresponde a una de las categorías léxicas: sustantivos, verbos, adjetivos, preposición, y es el elemento que determina las características básicas de un sintagma. Por lo tanto, el núcleo sintáctico es el constituyente más importante o de mayor jerarquía que se encuentra de un sintagma. El procedimiento por el cual cualquiera de las categorías citadas se desarrolla en un sintagma se describe mediante la denominada teoría de la $\overline{X}$ (**X-barra**) , según la cual un núcleo se proyecta dos veces y recibe un modificador en cada proyección.

teoría de la X-barra

Si se emplea la letra X para designar un núcleo sintáctico de alguna de las categorías léxicas, el sintagma en que se desarrolla (SX) puede descomponerse como muestra la figura 33.

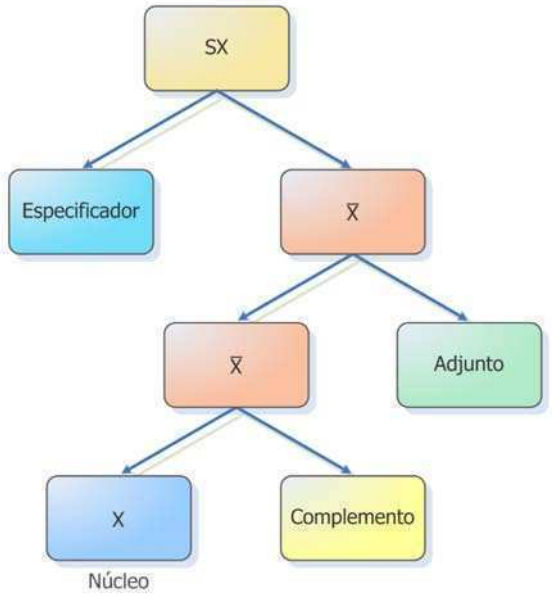

Figura 33: Organización de categorías léxicas de acuerdo a la teoría X-barra

- Un núcleo X se proyecta al nivel $\overline{X}$ al recibir un complemento como modificador.
- $\overline{X}$ puede recibir optativamente un adjunto sin modificar su naturaleza de proyección intermedia.
- $\overline{X}$ se proyecta a SX al recibir como modificador a un especificador.

En [127] se explica la formación de las frases fonológicas. Se propone que el dominio de las ϕ está dado por una cabeza léxica o núcleo (X) y todos los constituyentes del lado no recursivo, hasta donde empieza otro constituyente, fuera de la proyección maximal de X. El lado no recursivo del Español es el izquierdo, el lado recursivo es el derecho (las estructuras se adjuntan de este lado y en forma ilimitada.

Recordando que uno de los principios de la jerarquía prosódica establece que uno solo de los nodos hermanos puede ser fuerte, en [127] se

muestra que la prominencia en el nivel de frase fonológica depende de la sintaxis del lenguaje: los lenguajes de ramificación a derecha como el Español o el Inglés tienen prominencia en el ϕ final, mientras que los lenguajes de ramificación a izquierda, como el Japonés, muestran prominencia en el ϕ inicial. Es decir, la prominencia en el nivel prosódico ϕ refleja y señala una diferencia fundamental en los lenguajes. Esto presenta implicancias para la adquisición del lenguaje: si los niños son sensibles a los ϕ, podrían emplear esa información prosódica para descubrir la estructura sintáctica de su lenguaje nativo.

En [153] se muestra la interrelación entre la frase sintáctica y la frase fonológica, y se propone que el límite derecho de una frase sintáctica (X) debe coincidir con el límite derecho de una frase fonológica (ϕ). En otras palabras, que deben coincidir las posiciones de las fronteras finales tanto sintácticas como prosódicas. En [178] se introduce la regla de *envolventes de frases sintácticas X* (*Wrap-XP*) que completa la propuesta anterior: establece por cada frase sintáctica contenida debe haber una ϕ que sea su continente.

En [139] se observa que otro tipo influencias son incluso más fuertes que las sintácticas. Describe por ejemplo que se suelen respetar patrones de tipo (ωω) dentro de las ϕ, o un balance entre las ϕ contenidas en las IP. Esos fenómenos prosódicos propugnan la euritmia de las emisiones, el principio de regularidad rítmica de las prominencias y de los acentos. Con referencia a la sintaxis, se observa la alineación de la ϕ y del sintagma en el lado derecho de ambas estructuras.

Por en [145] estudiando el dialecto del Español de Lima, Perú, se proponen reglas prosódicas y sintácticas que definen las ϕ: (1) cada ϕ tiene un tamaño *ideal* de dos ω; (2) coincide el límite derecho de una ϕ con el límite derecho del sintagma; (3) cada sintagma está contenido en una ϕ. Observa también la importancia de las limitaciones prosódicas sobre las limitaciones sintácticas en la buena formación de las ϕ, así como en la organización de las ϕ dentro de la IP.

También es posible encontrar en nuestro idioma algunas pistas fonéticas que se correlacionan con los límites de las ϕ, por ejemplo: el instante en que termina un ascenso de continuación de la curva de F0, una mayor duración de las sílabas prominentes, aumentos o disminuciones notables en el rango del pitch, y pausas [38, 81, 30, 139, 173].

Debido a que las ϕ están estrechamente vinculadas a la sintaxis, su comportamiento también depende de la sintaxis del lenguaje considerado. En Inglés por ejemplo la cabeza léxica ocurre en el último ϕ, mientras que en el Japonés se da en el ϕ inicial.

3.3.5 Frase Entonativa (*IP*)

La frase entonativa es el constituyente prosódico que agrupa uno o más ϕ. Estos dominios están definidos por la entonación y son el dominio de contornos entonativos coherentes [127, 155]. Las *IP* están delimitadas por pausas, alargamiento de segmentos finales de frases y patrones de movimientos tonales.

Un rasgo compartido por todos los lenguajes es la presencia de un componente entonativo. Se ha descripto a la entonación como un len-

guaje universal [78]. Por un lado en todos los lenguajes los movimientos tonales transmiten alguna información lingüística o paralingüística, y por el otro los sistemas entonacionales parecen ser compartidos por lenguajes muy diferentes. Por ejemplo, en muchos lenguajes un ascenso en el contorno tonal se emplea para contrastar con un pitch de nivel más bajo, para indicar que la expresión es una interrogación en vez de una declaración.

Cada *IP* está caracterizada por un acento nuclear asociado a una sílaba prominente. El acento nuclear es un patrón tonal que otorga prominencia a la sílaba que alberga. Dicho patrón tonal puede ser un movimiento de pitch, un salto del mismo o un cambio de dirección de su contorno. Además cada *IP* finaliza con un tono de frontera, marcado típicamente por una disminución en el pitch.

Mientras que el patrón global del pitch presenta una tendencia a disminuir en el curso de una expresión, tal declinación vuelve a comenzar en los bordes de las *IP*. Las *IP* son notables perceptualmente, y son responsables de generar puntos naturales de segmentación del habla. Además parecen ser requeridas obligatoriamente en ciertos constituyentes sintácticos [127]. Por ejemplo:

[*Los leones,*] **IP** [*como se sabe,*] **IP** [*son peligrosos*] **IP**

No obstante, como con todos los constituyentes prosódicos, la sintaxis no es suficiente para explicar las *IP*.

Por ejemplo la longitud de una expresión es directamente proporcional al número de *IP* encontrados, e inversamente proporcional a la velocidad de elocución. Consecuentemente, los estilos de habla más lenta conducen a un número menor de *IP*. En consecuencia, la asignación de *IP* en las expresiones puede estar vinculado a causas fisiológicas como la capacidad respiratoria [127].

Por otro lado, no todas las ϕs pueden ser *IP*. Por ejemplo la oración: *"Tres matemáticos de diez derivan un lema"* no se puede dividir en IP de la siguiente forma:

[*Tres matemáticos*] **IP** [*de diez derivan un lema*] **IP**

Es decir que los límites de las *IP* tienden a encontrarse al final de una frase nominal, pero no después de sustantivos en el interior de dichas frases. De hecho en [152] se propone que cada *IP* es una unidad de sentido: dos constituyentes C_i y C_j forman una unidad de sentido si C_i depende de C_j (sea modificador o complemento).

Varios investigadores intentaron brindar un modelo integral sobre cómo se conforman las *IP*, y cómo se las emplea durante la comprensión del habla [190]. Estos estudios muestran que los locutores tienden a ubicar los límites de las *IP* antes o después de los constituyentes sintácticos principales, y que los oyentes utilizan esa pista para determinar dónde se pueden establecer los límites de frases sintácticas.

Finalmente, diversos experimentos conductuales han mostrado un efecto de los límites de *IP* en el procesamientote el habla continua [190, 29].

En Español, el final de una *IP* está marcado mediante un acento de frase dado por un tono L, H o por una pausa.

3.4 MODELOS ENTONATIVOS

En general un modelo de entonación presenta tres componentes: [77]:

- una descripción/representación de la entonación basada en alguna teoría
- un método para el mapeo de esa descripción a morfologías en las curvas de F0
- un método para derivar la descripción de entonación de una curva continua de F0

Los modelos de entonación se pueden agrupar respecto a los siguientes atributos:

1. Nivel de análisis y representación entonativa.

   Sobre esta base se puede hacer una distinción entre los modelos fonéticos o cuantitativos, por un lado y los fonológicos, cualitativos o secuenciales por el otro, pasando por un nivel intermedio, denominado de fonología superficial.

   modelos fonéticos

   Los **modelos fonéticos** describen los atributos entonativos mediante vectores de atributos acústicos o mediante parámetros continuos como duración, pendiente, amplitud, y posición del pico de F0. El valor de F0 obtenido durante la generación del contorno se obtiene empleando algún modelo de regresión sobre dichos vectores de atributos.

   A su vez los modelos fonéticos se pueden clasificar en paramétricos y no paramétricos. Los primeros reciben ese nombre por emplear un conjunto de parámetros para describir los patrones de entonación, mientras que los modelos no paramétricos usan directamente las muestras los valores de F0 para desarrollar los modelos entonativos.

   modelos fonológicos

   Por su parte los **modelos fonológicos** , describen la curva entonativa mediante secuencias de categorías tonales obtenidas a partir de la definición de un inventario de categorías tonales distintivas y de una gramática entonacional. Estos modelos buscan identificar los elementos contrastivos del sistema entonativo, cuya combinación produce los contornos melódicos encontrados en los enunciados admitidos para una lengua.

   A diferencia de los modelos fonéticos, que en primer lugar tienen en cuenta los acústicos de la entonación, los modelos fonológicos consideran desde el inicio los aspectos funcionales, vinculados con información lingüística de un nivel más alto.

2. Forma de procesar la información entonativa

   De acuerdo a la forma en que se modelan los contornos entonativos se distinguen los modelos superposicionales, que obtienen la curva de F0 a partir de la suma de dos componentes de diferentes dominios temporales; y los secuenciales en los que se representan los contornos como secuencias de elementos discretos (acentos tonales, tonos de juntura, conexiones), que se asocian con los elementos de la cadena segmental.

3. Dirección en el análisis entonativo.

   modelos generativos

   modelos analíticos

   De acuerdo a esta distinción se tiene a los modelos generativos y los analíticos. Los **modelos generativos** se corresponden con un procesamiento de arriba hacia abajo (*top-down*), en el que se producen los contornos de F0 a partir de los niveles de información superiores. Por el contrario, los **modelos analíticos** implican un procesamiento de abajo hacia arriba (*bottom-up*) que infiere la información de alto nivel partiendo del contorno de F0.

4. Forma de obtener los codificadores de contornos de F0.

   Según este atributo podemos separar los modelos en basados en datos y los basados en reglas. Los modelos basados en datos emplean técnicas de aprendizaje maquinal para generar los contornos a partir de las entradas simbólicas, mientras que los últimos llevan a cabo la codificación partiendo de reglas especificadas a partir de conocimiento experto.

A continuación se detallarán algunos de los modelos entonativos más empleados:

3.4.1 Modelo Tilt

El modelo Tilt [168] se puede considerar un modelo entonativo fonético secuencial. En este modelo los contornos entonativos se consideran como secuencias de eventos discretos: acentos tonales y tonos de juntura, asociados con elementos de la cadena segmental: sílabas acentuadas y sílabas de final de frase. Las porciones de la curva entre esos eventos se denominan conexiones.

Se utiliza el siguiente procedimiento para modelar los contornos entonativos:

a) Los acentos tonales se etiquetan con la letra **a** y se definen como excursiones de F0 asociadas con las sílabas, que son utilizadas por el locutor para dar algún grado de énfasis a una palabra o sílaba en particular.

b) Los tonos de juntura se indican con la letra **b** y se consideran como el inicio de excursiones de F0 que ocurren en los límites de las frases entonativas y dan al oyente indicios de fin de frase, o de continuidad de una elocución.

c) Las conexiones se denotan con la letra **c**

d) Se utiliza la etiqueta **ab** para describir situaciones en las que un acento tonal y un tono de juntura ocurren tan próximos entre sí que se observa un solo movimiento de F0

De ser necesario se puede extender este conjunto de etiquetas básicas, por ejemplo en lugar de describir un solo tipo de límite de frase se puede hacer una distinción entre límites en ascenso o en descenso e introducir etiquetas correspondientes para indicarlos.

La descripción entonativa resultante es compacta (hay un número limitado de etiquetas) y funcional (tiene en cuenta los aspectos funcionales de la entonación).

Es posible describir completamente cualquier evento entonativo empleando tres tipos de contornos locales: de ascenso, de descenso y un

tercero de ascenso-descenso. De esta forma se puede caracterizar un evento cualquiera mediante cuatro parámetros: amplitud y duración del ascenso, y amplitud y duración del descenso. Si un evento presenta sólo uno de los componentes, por ejemplo de ascenso, tanto duraciones como amplitudes del componente de descenso son nulos.

La duración de ascenso se calcula midiendo la distancia entre el inicio del evento y el pico, de igual manera la amplitud de ascenso es la diferencia entre el valor de F0 en el pico y el valor inicial. Para el descenso se pueden obtener esos valores de manera equivalente: la duración de descenso igual a la distancia entre el fin del evento y el pico, y su amplitud como la diferencia de F0 final con respecto al pico. Por lo tanto la amplitud de ascenso es siempre positiva y la descenso siempre negativa.

De esa manera, el modelo Tilt describe cada evento entonativo mediante movimientos distintivos de la curva de F0, descriptos en términos de parámetros continuos derivados de manera automática de dicha curva.

- La amplitud es la suma de las amplitudes de ascenso y descenso.
- La duración es la suma de duraciones de ascenso y descenso.
- Los parámetros *tilt*, que codifican la forma total del evento entonativo. Se calcula de las amplitudes y duraciones de ascenso ($A_\uparrow$ y $D_\uparrow$) y descenso ($A_\downarrow$ y $D_\downarrow$) empleando la siguiente ecuación:

$$\text{tilt} = \frac{|A_\uparrow| - |A_\downarrow|}{2|A_\uparrow| + |A_\downarrow|} + \frac{|D_\uparrow| - |D_\downarrow|}{2|D_\uparrow| + |D_\downarrow|} \tag{3.4}$$

- Para cada evento se define el valor de F0 inicial y determina su posición temporal, ya sea en relación con el comienzo de la frase o del núcleo silábico con el cual el evento está asociado.
- Posición silábica, que indica la alineación del evento con la sílaba o vocal acentuada. Juntamente con el parámetro *tilt* este atributo codifica las diferencias categóricas entre los acentos tonales.

Para este modelo se han definido métodos automáticos de análisis y síntesis que deben ser entrenados empleando bases de datos anotadas con este tipo de codificaciones. Este procedimiento ha mostrado buenos resultados modelando la entonación del Inglés, pero no se encontraron aplicaciones al Español.

3.4.2 Modelo de Fujisaki

El modelo de Fujisaki [48] es del tipo fonético superposicional, y brinda interpretaciones fisiológicas vinculando los movimientos de F0 con la actividad intrínseca de los músculos laríngeos.

En la figura 34 se muestra un esquema la generación de contornos de F0 empleando este modelo.

Como se puede ver en la figura, este modelo representa los contornos de F0 empleando dos componentes: por un lado los **Comandos de**

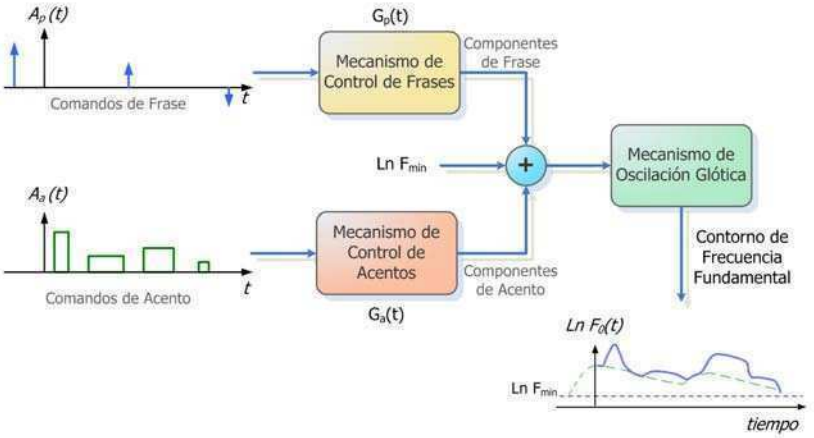

Figura 34: Modelo de Fujisaki para la generación de contornos de F0. La salida del componente de acento (una secuencia de respuestas amortiguadas ante funciones escalón de diferentes amplitudes y duraciones) se suma a la salida del componente de fase (una serie de una o más respuestas asintóticas a funciones impulso), y a una frecuencia fundamental mínima, para generar el perfil de F0 resultante.

frase, que representan el fenómeno de declinación y aportan la morfología global del contorno entonativo. Estos comandos se modelan empleando funciones impulso. Cada nueva frase implica un reinicio de la declinación y se modela con un nuevo comando de frase. Por otro lado los **Comandos de acentos**, que se emplean para reproducir el comportamiento de los acentos tonales. Cada uno de estos comandos está dado por una función escalón con cierta ubicación temporal, amplitud y duración.

Dado el siguiente conjunto de parámetros:

- F_{min} : valor de F0 de base,
- I : número de comandos de frase,
- J : número de comandos de acento,
- A_{pi} : magnitud del i-ésimo comando de frase,
- A_{aj} : amplitud del j-ésimo comando de acento,
- T_{0i} : ubicación temporal i-ésimo comando de frase,
- T_{1j} : instante de inicio del j-ésimo comando de acento,
- T_{2j} : tiempo de finalización del j-ésimo comando de acento,
- α : frecuencia natural del filtro de control de frase,
- β : frecuencia natural del filtro de control de acento,
- γ : cota relativa superior de los componentes de acento.

El modelo utiliza las ecuaciones 3.6-3.7 para obtener la curva de F0 mediante la superposición de los comandos de acento y frase en el dominio logarítmico:

$$\ln(F0(t)) = \ln(F_{min}) + \sum_{i=1}^{I} A_{pi} \cdot G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \left(G_a(t - T_{1j}) - G_a(t - T_{2j})\right) \tag{3.5}$$

$$G_p(t) = \begin{cases} \alpha^2 \cdot t \cdot e^{(-\alpha t)}, & \text{para } t \geqslant 0 \\ 0, & \text{en otro caso} \end{cases} \tag{3.6}$$

$$G_a(t) = \begin{cases} \min\left[1 - (1 + \beta t) \cdot e^{(-\beta t)}, \gamma\right], & \text{para } t \geqslant 0 \\ 0, & \text{en otro caso} \end{cases} \tag{3.7}$$

Donde los parámetros α y β se consideran constantes durante toda la emisión y γ generalmente se adopta como una constante de valor 0,9.

El modelo contempla dos filtros lineales críticamente amortiguados: el *filtro de comandos de frase*, y el *filtro de comandos de acentos*. El componente de frase se genera como la respuesta del filtro de comandos de frase ante una función impulso. Por su parte el componente de acento surge como respuesta a una función escalón del filtro de comandos de acentos. Finalmente el perfil de F0 resultante se obtiene sumando a estos componentes al nivel de base F_{min}, que depende del locutor, así como del tipo y duración de la oración.

3.4.3 Modelo IPO

El modelo IPO (**I**nstituut voor **P**erceptie **O**nderzoek) [165] es un modelo entonativo perceptual, basado en reglas. Los modelos perceptuales en general asumen que el contorno de F0 contiene una serie de movimientos tonales que perceptualmente carecen de relevancia, y por lo tanto si se los elimina mediante un proceso de estilización. El perfil tonal resultante solo contendrá los movimientos tonales relevantes y será perceptualmente equivalente al original.

El proceso de estilización se lleva a cabo manualmente y consiste en reemplazar el contorno original de F0 por el mínimo número de líneas rectas que permitan preservar los movimientos tonales perceptualmente relevantes. El siguiente paso consiste en una estandarización, proceso por el cual se establece un inventario de movimientos de pitch perceptualmente relevantes. Para hacerlo se determinan los atributos comunes de conjuntos de contornos estilizados y se los representa por un movimiento de pitch individual, basando la categorización en aspectos melódicos de la entonación, así como en propiedades acústicas: dirección, rango y pendiente del movimiento tonal y su alineación con el inicio de la vocal acentuada.

De esta forma, cada elemento de ese inventario de categorías distintivas se describe utilizando un vector de atributos relevantes perceptualmente.

El modelo IPO asume que una frase entonativa consiste de tres elementos: prefijo, raíz y sufijo, siendo solamente la raíz obligatoria. Como se dijo este modelo está basado en reglas, las que hacen necesario definir una gramática entonativa para determinar las combinaciones de categorías de movimientos tonales válidas, y las asociaciones entre esas categorías con elementos específicos de la estructura de frase entonativa.

En [53] se estudió la entonación del Español siguiendo este método. Se encontró una serie de patrones característicos a nivel de grupos acentuales y se analizaron los efectos de la declinación en los grupos de entonativos.

3.4.4 Modelo INTSINT

INTSINT (**IN**ternational **T**ranscription **S**ystem for **INT**onation) [77] es un ejemplo de modelos entonativos fonológicos superficiales: se encuentra en u nivel intermedio entre el fonético y el fonológico.

Este modelo supone que la forma y la función de la entonación se debe representar de manera separada. Utiliza un inventario de tonos que describen los contornos entonativos en el nivel fonológico superficial y se interpretan en el nivel fonético como valores de pich de referencia asociados con cada sílaba. El sistema no es completamente fonológico ya que por un lado no tiene en cuenta aspectos funcionales de la entonación, y por el otro no todos los tonos identificados en el contorno resultante se corresponden con tonos fonológicos.

Este modelo en primer lugar realiza una estilización de la curva de F0 empleando un algoritmo denominado MOMEL, que se explicará en detalle en la sección final de este mismo capítulo.

Una vez obtenidas las curvas de F0 estilizadas, se deriva una representación entonativa de alto nivel en término de tonos distintivos discretos, que pueden ser absolutos o relativos, y estos últimos a su vez ser iterativos o no iterativos:

- Los tonos absolutos comprenden las etiquetas: T, M y B (del Inglés *top*, *mid* y *bottom* respectivamente). Estos tonos se escalan respecto al rango tonal que presente el locutor.
- Los tonos relativos y no iterativos H, S y L (del Inglés *higher*, *same*, y *lower*) se escalan con respecto a la posición del tono precedente en la escala frecuencial.
- Los tonos relativos e iterativos U y D (del Inglés *upstepped* y *downstepped*) difieren de los no iterativos por presentar un menor rango del intervalo de F0. Estos tonos se escalan al igual que en el caso anterior con respecto al nivel tonal precedente.

INTSINT utiliza un conjunto de reglas para controlar el mapeo entre los tonos y las categorías tonales, lo que constituye la descripción entonativa en el nivel fonológico. Esas reglas se pueden considerar como

un algoritmo de codificación para la transcripción automática de la entonación. El modelo además incluye una serie de símbolos adicionales para marcar los límites temporales y de amplitud de las unidades de entonación. Por ejemplo se especifica el inicio y fin de las unidades entonativas entre corchetes, se marcan entre $<$ y $>$ el rango de F0 de la unidad entonativa. A través de esas marcas de duraciones y amplitudes, INTSINT facilita el mapeo de fragmentos de la curva de F0 y las etiquetas del modelo.

3.4.5 Modelo ToBI

Su nombre significa índices de tonos y pausas (ToBI: Tones and Break Indices) [13]. El concepto de tonos, hace referencia a la descripción de los perfiles entonativos empleando acentos tonales, acentos de frase y tonos de juntura. Por su parte el término de pausas se refiere a la descripción de los agrupamientos observados en la estructura prosódica.

Este modelo utiliza solo dos categorías de tonos altos (H) y bajos (L), y establece una serie de reglas para asignar tonos de acuerdo a la morfología de la curva de F0 a las unidades de entonación.

El modelo ToBI sigue una aproximación fonológica al modelado entonativo, y se basa en el modelo de entonación para el Inglés Americano propuesto por Pierrehumbert a principios de la década de 1980. Ese modelo a a su vez está fundado sobre la *Teoría Métrica Autosegmental* [98].

La teoría Métrica Autosegmental presenta cuatro preceptos básicos:

1. Linealidad de la estructura tonal: se considera que los perfiles de F0 son secuencias de eventos discretos: acentos tonales y tonos de juntura, asociados fonológicamente con elementos de la cadena segmental: sílaba acentuada y límite de una frase respectivamente. Las porciones del contorno entre dichos eventos se considera que no contribuyen al significado entonativo.

2. Distinción entre acento tonal y prominencia o acento prosódico: se considera que el acento tonal manifiesta la presencia de una sílaba prominente pero no la determina.

3. Análisis de acentos tonales en términos de niveles tonales: los acentos tonales y tonos de juntura se representan mediante niveles tonales fonológicos abstractos: los H y L, y sus combinaciones. Los acentos tonales se definen como un atributo local del contorno de F0, generalmente manifestado por un cambio tonal e involucrando un máximo o mínimo local.

4. Interpretaciones fonológicas de las tendencias globales de F0: Las tendencias globales de la curva de F0 incluye por un lado un descenso paulatino en el rango tonal y por el otro ascensos tonales locales considerados como escalones positivos. En los modelos superposicionales se modela el primer efecto como un fenómeno global de declinación, mientras que los enfoques fonológicos optan por modelar este comportamiento localmente, a través de escalones negativos que se producen en sitios específicos de la onda.

El modelo entonativo de Pierrehumbert definió todos los perfiles entonativos encontrados en el Inglés Americano a través de secuencias de un conjunto finito de categorías tonales, que incluían tonos de juntura (uno inicial optativo y otro final obligatorio), acentos tonales (tonos H, L o sus combinaciones), y acentos de frase (tonos H o L). Además este modelo considera que los acentos tonales pueden ser monotonales o bitonales, no hace distinción entre acentos tonales prenucleares y nucleares (donde el núcleo se considera como el último acento tonal de la frase), y se considera que los tonos de juntura son siempre monotonales. Por otra parte, Pierrehumbert empleó el símbolo * para indicar el alineamiento del tono con la sílaba acentuada, el símbolo % para denotar el alineamiento con los límites de la frase, el símbolo - para indicar la alineación del tono con los límites de palabras y el símbolo + para señalar la concatenación de tonos en la conformación de acentos bitonales.

El modelo ToBI sigue la mayor parte de los supuestos del modelo de Pierrehumbert, diferenciándose respecto al descenso escalonado. En ToBI los acentos tonales relacionados con el descenso escalonado reciben el diacrítico !, así por ejemplo se reemplaza el acento H+L* por H+!H* para indicar que el tono estrellado no es un mínimo local sino un tono alto que fue escalado hacia abajo en comparación al tono alto precedente.

Como se había mencionado ToBI también contiene índices de pausas que se colocan entre cada palabra, para reflejar el nivel de separación en las fronteras de los constituyentes prosódicos. Estos índices representan el valor del grado de juntura percibido entre cada par de palabra, y entre la palabra final y el silencio que indica el fin del enunciado. Los valores que pueden presentar son los siguientes:

- 0: indica el nivel de juntura entre palabras que conforman un grupo clítico.
- 1: marca el límite entre palabras prosódicas.
- 2: caracteriza la juntura entre palabras que teniendo algunas propiedades de los límites de frase no constituyen límites de frase.
- 3: indica límite de frase intermedia.
- 4: corresponde al límite de frase entonativa.

Si bien el modelo ToBI fue propuesto inicialmente como un sistema de etiquetado de corpus prosódicos para el Inglés Americano, se extendió rápidamente a diversos lenguajes como Coreano, Alemán, Griego, Japonés, y también al Español. Actualmente ToBI es el sistema más usado para la transcripción simbólica de la entonación.

A continuación, la figura 35 presenta un esquema que resume los modelos entonativos descriptos.

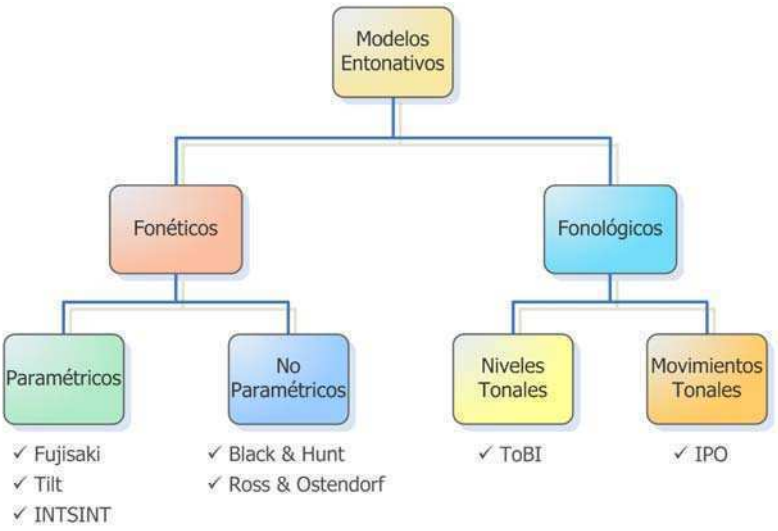

Figura 35: Clasificación de modelos entonativos.

3.5 MÉTODOS COMPUTACIONALES

En esta sección se describen los algoritmos para el procesamiento de la información suprasegmental utilizados durante el desarrollo de esta tesis.

3.5.1 Estimación Automática de F0

Como bien se mencionó tanto en el capítulo 1, como anteriormente en este mismo capítulo, el término pitch se asocia con la percepción auditiva de un tono.

La relación entre pitch y frecuencia se puede establecer pidiendo a un oyente que a través de la manipulación de la frecuencia de un tono senoidal puro, obtenga un pitch equivalente al de la señal en evaluación. En el punto que el oyente no encuentra diferencias perceptuales entre el tono complejo bajo estudio y la señal de referencia manipulada, la frecuencia de esta última se puede definir como el pitch del tono complejo.

La frecuencia fundamental de la señal de habla tiende a estar bien correlacionada con el valor de pitch percibido, por lo que suele utilizarse el valor de F0 para indicar un valor de pitch determinado.

A su vez el valor de F0 está directamente relacionado con la tasa de vibración de las cuerdas vocales, lo que en definitiva hace posible emplear las distintas manifestaciones acústicas de las vibraciones de las cuerdas vocales para estimar un valor de pich determinado.

La figura 36 muestra un oscilograma y el espectro de una vocal. Si bien a partir de esas imágenes puede parecer que la estimación de F0 es una tarea sencilla, existe una serie de fenómenos del habla natural que la convierten en un problema muy complejo [166]:

- Es frecuente que el valor de F0 varíe en cada período glótico.

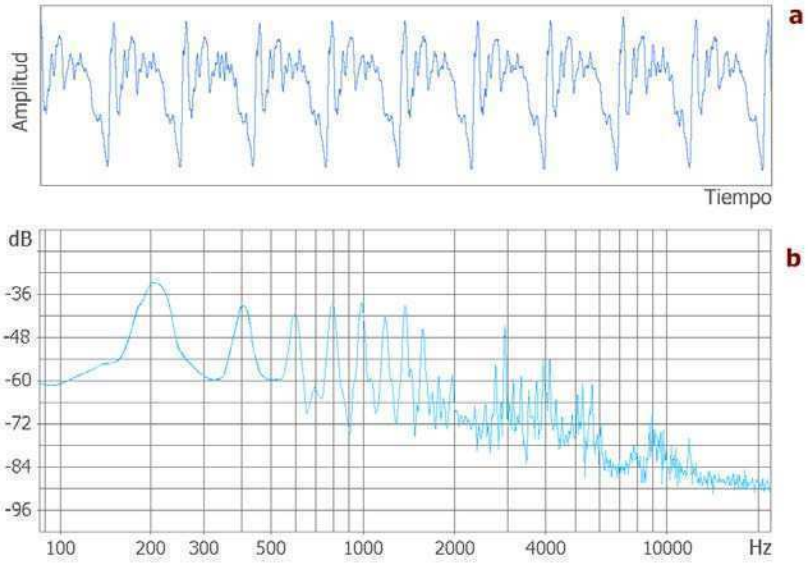

Figura 36: Manifestaciones acústicas de la vibración de las cuerdas vocales. **(a)** Oscilograma correspondiente a 50 ms de la vocal *a*, **(b)** Espectro de un período de la misma señal.

- Los sub-harmónicos del F0 a veces se pueden confundir con sub-múltiplos del F0 real.
- Cuando se presentan sub-harmónicos de magnitud considerable, muchas veces la estimación objetiva más razonable para el F0 está en clara contradicción con la percepción auditiva.
- las resonancias del tracto vocal y el filtrado realizado por el canal de transmisión pueden realzar otros harmónicos por sobre el primero y causar la estimación de múltiplos del F0 verdadero.
- No se puede descartar de plano una transición de una octava, ya que en la práctica a veces suceden.
- La forma de los períodos glóticos suelen ser irregulares en el comienzo y fin de segmentos sonoros, por lo que en esas posiciones la correlación entre períodos adyacentes suele ser baja.
- Es dificil encontrar concenso entre oyentes humanos respecto al inicio y fin de segmentos sonoros.
- El filtrado efectuado por algunas configuraciones del tracto vocal sobre la fuente de excitación en períodos sordos puede generar señales con una periodicidad aparente.
- La amplitud del habla sonora tiene un rango dinámico grande, que va desde una amplitud baja en consonantes oclusivas sonoras a un nivel de amplitud elevado durante la pronunciación de vocales abiertas.
- Resulta difícil distinguir objetivamente entre un ruido de fondo periódico y una voz tipo murmullo.

- Algunos intervalos de habla sonora duran unos pocos ciclos glóticos.

Se puede encontrar una gran cantidad de algoritmos para la determinación del pitch, en [75] se los ha dividido en tres generaciones:

algoritmos de la primera generación

1. Los **algoritmos de la primera generación** estaban diseñados para detectar el primer harmónico o el período de la señal de habla en regiones sonoras. Los primeros algoritmos de esta generación hay se denominan algoritmos del dominio temporal. Muchos de estos algoritmos tratan de reconstruir el ciclo glótico completo a partir de la señal de habla. El poder conocer el instante de cierre glótico es una ventaja de estos algoritmos con respecto a los de segunda generación. Por otro lado los primeros algoritmos del dominio frecuencial de esta generación fueron los basados en cepstrum, la transformada inversa de Fourier del logaritmo del espectro de amplitud. Esta familia de métodos de estimación dominaron el análisis del habla por más de dos décadas debido a su fiabilidad y eficiencia.

algoritmos de segunda generación

2. Los **algoritmos de la segunda generación** se caracterizan por sustentarse en teorías perceptuales del pitch, entre ellas las más importantes son la teoría del pitch virtual de Terhardt [170], y la del procesador óptimo de Goldstein [56]. Según esas teorías el pitch surge de las relaciones harmónicas entre los armónicos inferiores. Los primeros algoritmos de esta clase emplearon la suma espectral harmónica y el producto espectral armónicos para establecer el valor de F0, que más tarde condujeron a otras estrategias como el tamiz harmónico [37], el peine espectral [113] y la suma de sub-harmónicos [74].

algoritmos de tercera generación

3. Los **algoritmos de la tercera generación** no están solamente basados en teorías de percepción del pitch sino también en conocimiento sobre la fisiología del procesamiento auditivo periférico. Estos algoritmos presentan un considerable poder predictivo, que se atribuye al menos en parte a dos características de sus procedimientos: en primer lugar, descomponen la señal de habla en bandas de frecuencia definidos por los filtros auditivos cocleares, y en segundo lugar emplean información temporal en cada canal frecuencial por separado. Esto genera cierta inmunidad ante contaminación inharmónica, y permite tratar de manera adecuada el problema de harmónicos no resueltos.

A continuación se describirá el algoritmo para la estimación de F0 empleado en esta tesis.

Algoritmo RAPT

El algoritmo RAPT está basado en la función de correlación cruzada normalizada (NCCF). Dada una señal de habla no nula, de media cero

s_m, el valor de la función NCCF para el índice de retraso k y el frame de análisis i viene dado por la expresión:

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{e_m e_{m+k}}} \quad \text{con } 0 \leqslant k < K;\ \ 0 \leqslant i < M;\ \ m = iz \tag{3.8}$$

Donde M es el número de frames en la señal, K el número de muestras en la función de correlación, n el número de muestras en la ventana de correlación ($n = w/T$, con T igual al período de muestreo de la señal, y w el número de muestras en la ventana de análisis); z indica la cantidad de muestras de avance entre cada frame de análisis ($z = t/T$, con t representa el intervalo de análisis entre cada frame), y e_j viene dado por la ecuación 3.9

$$e_j = \sum_{l=j}^{j+n-1} (s_l)^2 \tag{3.9}$$

El valor de ϕ se encuentra en el rango $[-1, 1]$. Tiende a 1 para retardos correspondientes a múltiplos enteros del período *real* independientemente de posibles cambios en la amplitud de s, siempre y cuando se mantenga la similitud entre los períodos sucesivos. Para una señal s_n conformada por ruido blanco, $\phi_{i,0} = 1$ y $\phi_{i,k}$ tiende a cero para todo $k \neq 0$ a medida que w se incrementa.

El intervalo de correlación w se puede elegir de manera independiente del rango de F0 bajo consideración. Se lo suele elegir para que se encuentre en la vecindad del F0 buscado (en el orden de un período glótico promedio).

En este algoritmo se aprovechan los siguientes heurísticos para la determinación de la F0:

- Para segmentos sonoros, el máximo local de ϕ correspondiente al F0 *correcto* generalmente es el de mayor valor (excluyendo el correspondiente al valor de retardo nulo).
- Cuando hay varios máximos de ϕ, todos cercanos a 1, generalmente la mejor elección es el correspondiente al menor período.
- Debido a que el perfil de F0 varía suavemente, los máximos locales *correctos* de ϕ en frames de análisis adyacentes se localizan en valores de retardo equivalentes.
- A veces el valor real de F0 varía de manera abrupta, siendo posible que pase a valer el doble o la mitad de su valor previo.
- La sonoridad tiende a cambiar de estado con baja frecuencia.
- Para segmentos sordos, el valor máximo de ϕ en valores de retardos distintos a 0 son considerablemente menores a 1.
- El espectro de tiempo corto de tramas de habla sonoros y sordos generalmente son bastante diferentes.

- La amplitud tiende a aumentar en el inicio de regiones sonoras y a disminuir hacia sus finales.

La figura 37 presenta un esquema resumido del algoritmo:

Algoritmo RAPT

- Generar dos versiones de la señal de habla, una a la tasa de muestro original y otra submuetreada.
- Calcular periódicamente el NCCF de la señal submuestreada para todos los valores de retardo correspondientes al rango de F0 de interés.
- Estimar y almacenar los máximos locales de esa primera pasada de NCCF.
- Realizar una segunda pasada de NCCF pero sobre la señal original, y en la vecindad de los picos *prometedores* obtenidos en la fase previa.
- Buscar nuevamente los máximos locales sobre la NCFF así definida para refinar las localizaciones y amplitudes de los picos, que serán considerados como valores de F0 candidatos para cada frame.
- Emplear programación dinámica para seleccionar para todo el segmento de señal el conjunto de picos de la NCCF correspondientes al valor de F0 buscado, o hipótesis de segmento sordo, empleando en este proceso los heurísticos antes mencionados.

Figura 37: Algoritmo RAPT para la determinación de perfiles de F0.

3.5.2 Estilización de la Curva de F0

Las curvas de F0 se pueden considerar como la combinación de dos componentes: uno macroprosódico que refleja la elección del patrón entonativo que hace el locutor, y otro microprosódico, que es consecuencia de la elección de los fonos utilizados en la frase, y los mecanismos articulatorios para emitirlos. Mientras que el componente macroprosódico es importante para la percepción de la entonación, el microprosódico no lo es. Si bien el componente microprosódico se considera importante en la caracterización acústica a nivel segmental, carece de relevancia perceptual respecto a la entonación percibida.

El proceso de estilización consiste en el reemplazo de la curva de F0 por una función numéricamente más simple, tal que se conserve la información macroprosódica original. Este proceso está estrechamente relacionado con el concepto de *close-copy stylization* introducido más temprano en este mismo capítulo.

A continuación se describirá el algoritmo de estilización utilizado en esta tesis.

Algoritmo MOMEL

Este algoritmo de estilización fue propuesto originalmente como un módulo de procesamiento dentro del modelo entonativo INTSINT, y presenta 5 parámetros independientes:

- hzmin: mínimo valor de Fo admitido como válido
- hzmax: máximo valor de Fo admitido como válido
- A: ancho de la ventana de análisis
- D: umbral de distancia
- R: ventana de reducción

En primer lugar los autores asumen que como entrada al algoritmo se tiene la estimación de la curva de F0, obtenida mediante el algoritmo de peine espectral [112], con un valor de F0 estimado cada 10 ms, y que en las zonas sordas la curva de F0 presenta un valor de 0.

Se puede descomponer las secuencias de procesamiento que realiza el algoritmo en cuatro fases:

1. **Preprocesamiento**:

   Todos los valores de F0 que se encuentren a una altura mayor al 5 % del valor que presentan sus vecinos inmediatos se llevan a cero. Como las regiones no vocálicas en la curva de F0 ya presentaban un valor nulo, el efecto de este preprocesamiento es de eliminar uno o dos valores (alrededor de 10 a 20 ms) del ataque vocálico.

2. **Estimación de los candidatos**:

   Se realizan de manera iterativa los siguientes pasos para cada instante x.

   a) Dentro de una ventana de análisis de longitud A (típicamente 300 ms) centrada en x, los valores de F0 que se encuentran por debajo de un umbral $hzmin$ (de valor por defecto en 50 Hz) o por encima del umbral $hzmax$ (cuyo valor propuesto es de 500 Hz), se consideran subsecuentemente como datos faltantes.

   b) Se aplica una regresión cuadrática dentro de la ventana para todos los valores no neutralizados.

   c) Todos los valores de F0 que se encuentren a más de una distancia D por debajo del valor del F0 estimado por la regresión son neutralizados.

      Se iteran los pasos **b** y **c** hasta que no haya nuevos valores neutralizados.

   d) Para cada instante x se calcula un par $< t, h >$ a partir de los coeficientes de la regresión:

$$\begin{aligned} \hat{y} &= a + b \cdot x + c \cdot x^2 \\ t &= -b/2c \\ h &= a + b \cdot t + c \cdot t^2 \end{aligned} \tag{3.10}$$

Si t se encuentra fuera del intervalo $[x-(A/2),\ x+(A/2)]$, o h se encuentra fuera del intervalo $[hzmin,\ hzmax]$, entonces el punto correspondiente a ese $< t, h >$ se considera como dato faltante.

Los pasos **b**, **c** y **d** se repiten para cada instante x, resultando en un punto $< t, h >$ o valor faltante para cada valor del F0 original.

La regresión modal asimétrica constituye la parte central del algoritmo. Aplica una regresión modal dentro de una ventana móvil, para obtener una estimación óptima de la frecuencia fundamental local, centrada en cada valor de la curva de F0.

Esta fase consiste en encontrar el conjunto de parámetros de una función de tipo *spline* cuadrática, que permitan optimizar su ajuste a los elementos de la serie temporal preprocesada.

Como puede haber más de un valor para los parámetros de la regresión modal, se imponen algunas restricciones que reduzcan las ambigüedades de búsqueda y hagan más sencillo el proceso. Una de esas restricciones consiste en asumir que el único efecto microprosódico es hacer descender la curva macroprosódica, que se propone modelar. Es decir, se introduce la restricción que la función spline cuadrática que se desea encontrar es tal que no hay valores a una distancia mayor a D por arriba de la función, y que el número de elementos de la serie que quedan por debajo de una distancia D de la función es mínimo.

3. **Partición de los candidatos estimados**

   Dentro de una ventana móvil de longitud R (típicamente 200 ms) centrada en cada instante x, se calculan los valores $dt(x)$ y $dh(x)$ como el valor absoluto de las distancias entre las medias de t y h de los candidatos en la primer mitad de la ventana con respecto a las medias de los candidatos de la segunda mitad de la ventana. Luego se obtiene una distancia combinada pesando las anteriores:

$$\begin{aligned} d(x) &= \frac{dt(x)\cdot wd + dh(x)\cdot wh}{wd+wh} \\ wd &= 1/media(td(x)) \\ wh &= 1/media(hd(x)) \end{aligned} \tag{3.11}$$

   Luego se eligen para cada valor de x los límites de las particiones usando las siguientes condiciones:

   - $d(x) > d(x-1)$
   - $d(x) > d(x+1)$
   - $d(x) > mean(d(x))$

4. **Reducción de candidatos**

   Dentro de cada segmento de la partición, se eliminan los candidatos para $dt(x)$ o $dh(x)$ que sean mayores a un desvío estándar de los valores medios correspondientes. Finalmente se calcula el valor medio de los candidatos restantes para cada segmento y

se los adopta como la estimación de los parámetros para dicho segmento.

4 INFORMACIÓN SUPRASEGMENTAL Y CONTENIDO LÉXICO

ÍNDICE

Hasta aquí se han presentado los fundamentos, así como los principales argumentos de esta tesis. En los siguientes tres capítulos se exponen los estudios llevados a cabo con el fin último de aplicar eficientemente información suprasegmental para mejorar el proceso de RAH.

Este capítulo estudia diferentes alternativas para obtener desde los atributos suprasegmentales, información lingüística útil para el reconocimiento del habla. Se intenta determinar qué patrones presentes en el canal de información suprasegmental se puede asociar con algún tipo de información léxica, que se pueda emplear en la reducción de ambigüedad durante el RAH.

El contenido de este capítulo se puede dividir en cuatro partes.

En la primera se detalla el cuerpo de datos empleados para los estudios del presente Capítulo.

Las tres partes restantes se dividen de la siguiente forma:

- **Descubrimiento de agrupamientos en frases entonativas.**

  En esta parte se explora la siguiente proposición: *"Es posible obtener agrupamientos de de frases a partir de sus rasgos suprasegmentales"*.

  De poder identificar estos distintos grupos, se los podría utilizar para construir modelos de lenguajes específicos para cada clase de frase, tal que permitan mejorar el desempeño de un reconocedor estándar. Para evaluar la proposición se buscó determinar agrupamientos (*clusters*) de grupos entonativos empleando información suprasegmental multidimensional y redes neuronales autoorganizadas.

- **Correlación entre número de palabras de contenido y picos tonales.**

  En la tercera sección se examina la siguiente hipótesis: *"Para cada frase existe un correlato entre el número de palabras de contenido, y la morfología de las curvas de* F0*. Específicamente, existe un pico tonal por cada palabra de contenido"*.

  De verificarse esta hipótesis sería posible emplear el perfil de la curva de F0 para determinar la cantidad de palabras de contenido, y usar esa información como una restricción durante la búsqueda de mejores hipótesis, o en el posprocesamiento del reconocedor.

  En esta parte se construyeron diferentes detectores de picos entonativos a partir de varias estilizaciones de la curva de F0, y se evaluó la correlación entre el número detectado de picos y la cantidad de palabras de contenido presentes, de acuerdo a las anotaciones del corpus empleado.

- **Clasificación de acento léxico de palabras finales de frase a partir de rasgos suprasegmentales.**

  Finalmente se explora la hipótesis: *"Se puede encontrar un comportamiento distintivo en los patrones suprasegmentales para palabras agudas, graves y esdrújulas de final de oración"*.

  Esta información podría reducir el número de palabras candidatas durante el reconocimiento a partir de sus acentos léxicos.

  Para estudiar esta hipótesis en primer lugar se analizó el comportamiento del contorno de F0 para palabras monosílabas, agudas, graves y esdrújulas de final de oración. En segundo lugar se efectuó un análisis estadístico de duración, máximo de F0 y máximo de energía para las tres últimas sílabas de cada oración en relación al tipo de acento de la palabra final de las mismas. Se buscó determinar si a partir del comportamiento relativo de estos valores para cada sílaba es posible determinar la tonicidad de la palabra final.

A continuación se describirá cada una de las partes en que se ha dividido el capítulo.

4.1 CORPUS DE DATOS EMPLEADOS

Para el desarrollo de todos los estudios realizados en este Capítulo, así como para los efectuados en el Capítulo 6, se utilizó el corpus oral LIS-SECYT.

Este corpus está conformado por 741 oraciones declarativas, leídas por dos locutores profesionales nativos de Buenos Aires (uno masculino y otro femenino).

El conjunto de oraciones mencionado contiene el 97 % de las sílabas del español, en las dos condiciones de acento (sílaba acentuada y no acentuada), y en todas las variantes posicionales (inicial, media y final) dentro de la palabra.

El 70 % de esas oraciones fueron obtenidas de periódicos que se publican en Buenos Aires. El resto fue creado por maestros de lengua, quienes recibieron la instrucción de elaborar oraciones con palabras que contuvieran las sílabas menos frecuentes.

En esta tesis se estudió solamente el subconjunto de datos correspondientes a la informante femenina, ya que las diferencias entre locutores podrían llegar a obscurecer los patrones prosódicos que se intenta examinar.

Cada una de las emisiones fue etiquetada dos veces por cuatro fonaudiólogas con entrenamiento musical. Cada onda tiene asociada once archivos de etiquetado: cuatro del etiquetado tonal, uno del fonético, uno de datos acústicos, cuatro de categorías sintácticas y uno de clases de palabras [65]. Las marcas tonales y de pausas se efectuaron de acuerdo al método descripto en [49].

En el Cuadro 1 se presenta las frecuencias de oraciones con un número específico de grupos entonativos para la base de datos descripta.

Grupos Entonativos	Número de Oraciones
1	11
2	371
3	319
4	36
5	1
6	3

Cuadro 1: Distribución del número de oraciones en relación a la cantidad grupos entonativos contenidos en el corpus SECyT

Respecto a las condiciones bajo las cuales se registraron estas muestras, las elocuciones fueron grabadas en un ambiente de laboratorio especialmente acondicionado (cámara anecoica), se emplearon micrófonos dinámicos AKG, y digitalización con una frecuencia de muestreo de 16 KHz y resolución de 16 bits por muestra.

4.2 AGRUPAMIENTOS DE FRASES ENTONATIVAS

Una de las formas en que se puede aprovechar la información suprasegmental es a través de las modalidades de una oración. Si bien el reconocimiento del habla estrictamente busca la conversión de habla en secuencias de palabras, también resulta útil saber si la secuencia de palabras pronunciadas son parte por ejemplo de una pregunta o afirmación.

A través de la morfología entonativa es posible caracterizar la modalidad a la que corresponde un fragmento de habla. De esta forma, es posible utilizar las características del perfil de F0 para establecer por ejemplo si el locutor está realizando una interrogación, declaración o

tipos de actos de habla

exclamación. Estas modalidades de elocuciones se conocen también como **tipos de actos de habla** .

La información sobre esas modalidades es especialmente útil para el área de comprensión del habla. Por ejemplo se han utilizado esas clases de actos de habla para determinar las intenciones de los usuarios operando con interfases orales de diálogo [163].

Sin embargo esa información también resulta útil para el reconocimiento del habla [163]. En [91] se emplea un detector de tipos de actos de diálogo a partir de parámetros suprasegmentales, y se adapta un modelo de lenguaje general para cada categoría de esos actos de habla, logrando algunas reducciones en las tasas de error por palabras.

En este apartado se estudia esa misma idea, a saber, obtener grupos de patrones entonativos a partir de la información suprasegmental, con la finalidad de determinar modalidades que puedan emplearse en el proceso de RAH. En esta sección se buscó determinar si se podría obtener de manera automática agrupamientos a nivel de frases entonativas observando los parámetros prosódicos de cada una de dichas frases.

Para ello se partió de la segmentación manual de un conjunto oraciones en frases, es decir se asumió que se contaba con la información de los instantes de comienzo y fin de cada frase entonativa. Posteriormente se parametrizó la información suprasegmental presente en cada una de esas frases y se construyeron **mapas autoorganizados** [95] para buscar ordenamientos naturales en los datos.

4.2.1 Parametrización de Rasgos Suprasegmentales

En primer lugar se determinaron las curvas de F0 correspondientes a cada oración del corpus LIS-Secyt empleando el algoritmo RAPT, descripto en la Sección 3.5.1. Además se aprovechó la estimación de energías por frame que realiza este mismo algoritmo para obtener el conjunto de curvas de energía para cada oración. Al igual que para el caso de F0 las estimaciones de energía se realizaron cada 10 ms, y en este caso se suavizó el resultado obtenido, empleando un filtro de medianas de 50 ms.

Como se mencionó previamente, en esta prueba se supuso que las segmentaciones de las oraciones en sus frases entonativas eran conocidas. En la práctica se utilizaron los archivos de anotaciones grafémicas (*.gra), y de pausas (*.brk) del corpus LIS-Secyt, para obtener tales segmentaciones de los archivos de audio de forma automática.

En la figura 38 se muestra un ejemplo de estos archivos de anotaciones correspondientes a la oración: *"Álvarez, se había animado a contarle un chiste"*. En la misma, el valor de la columna derecha para los archivos **.brk*, indica el grado de pausas percibido entre pares de palabras según la notación ToBI (ver Sección 3.4.5).

Un índice igual a 0 corresponde a los casos en que no se percibe separación entre palabras y de 4 al nivel máximo de separación perceptual, generalmente manifestada a tavés de un silencio entre el par de palabras correspondientes.

Secyt-m1-1.gra		Secyt-m1-1.brk	
Tiempo	Label	Tiempo	Label
0,677722	Álvarez	0,677722	4
1,321935	/p	1,321935	4
1,390507	se	1,392489	4
1,623004	había	1,622957	0
2,046500	animado	2,046500	1
2,736911	a	2,736183	3
2,822149	contarle	2,822785	1
3,453146	un	3,453146	0
3,678731	chiste	3,678301	4

Figura 38: Archivos de anotaciones grafémica (izquierda) y de pausas (derecha), correspondientes a la oración *"Álvarez, se había animado a contarle un chiste"*. La primera columna de ambos archivos indica el tiempo de inicio de la anotación (en segundos), y la segunda su identidad.

La tarea de segmentación radicó así en extraer para cada oración, los fragmentos de las curvas de F0 y energía, de acuerdo a las anotaciones de los archivo de pausas. Específicamente se emplearon los índices de pausas de valor 3 y 4 para marcar los instantes de inicio de cada nuevo intervalo.

A partir de la segmentación de las frases entonativas, se emprendió la tarea de establecer el conjunto de parámetros que permitiesen caracterizar de manera sintética, la información suprasegmental de cada una de ellas.

Se determinó que uno de esos parámetros fueran las duraciones totales de cada frase, ya conocidas a partir de las segmentaciones, y que la información morfológica de las curvas de F0 y energía serían representadas empleando los coeficientes de una aproximación polinómica.

El empleo de regresiones polinómicas para representar curvas de F0 no es nuevo. Uno de los primeros antecedentes en la parametrización de curvas de F0 mediante ajustes polinómicos se puede encontrar en [103], mientras que aplicaciones más recientes incluyen a las descriptas en [4] y [58].

polinomios ortogonales de Legendre

En [58] se propone la familia de polinomios ortogonales de Legendre , para realizar esa parametrización, argumentando que ésta familia de funciones exhibe igual sensibilidad sobre la extensión completa de la curva a representar, y que además, por ser una familia de polinomios ortogonales presentan las siguiente propiedades:

propiedades de polinomios ortogonales

- Se pueden usar de manera reversible: para analizar una curva y para sintetizarla.
- Forman un conjunto completo de funciones. Esto implica que combinando las funciones de la familia se puede aproximar cualquier otra función tanto como se desee.

- Son ortogonales, es decir que las funciones que conforman la familia no están correlacionadas entre sí, y se pueden usar cada una para medir una propiedad diferente de la curva considerada. En el caso más común, los ajustes realizados empleando cada función de la familia pasan a ser estadísticamente independientes entre sí.

Como las propiedades anteriores también resultan convenientes a los fines de este estudio, se eligió utilizar polinomios de Legendre para representar tanto las curvas preprocesadas de F0, como de energía. Mediante este procedimiento se ajusta una función polinómica a las curvas originales, y los coeficientes de la función obtenida caracterizan sus morfologías.

Esta aproximación de funciones es válida para representar curvas que se encuentren en el intervalo $[-1, 1]$, por lo que antes de efectuar el ajuste se normalizaron los contornos de F0 y energía a nivel de oración, tal que resultasen curvas de media 0 y desvío estándar 1.

Además, se aplicó otra operación de preprocesamiento sobre las curvas de F0: se efectuó una interpolación a nivel de frases, empleando funciones splines para obtener funciones continuas, y evitar cambios abruptos en las transiciones sonoro-sordo y sordo-sonoro.
Esta operación de alguna manera está en consonancia con el suavizado perceptual del pitch, comentado en el Capítulo 3.

Con las curvas de energía y F0 así preprocesadas se efectuó un estudio sobre el grado de los polinomios de Legendre que pudiesen garantizar una buena aproximación a las señales representadas.
Se realizó una prueba exhaustiva analizando para el conjunto completo de frases segmentadas, el grado de correlación entre la curva original y su regresión polinómica, así como la raíz cuadrada del error cuadrático medio (RMSE) entre ellas, tanto para el caso de F0 como de energía, variando los órdenes polinómicos en el rango $[2 - 25]$.

En la figura 39 se muestra el resultado obtenido para el caso de los coeficientes de correlación promedio.
De acuerdo a los resultados de la figura 39, se decidió representar las curvas de F0 empleando polinomios de orden 10, con lo cual se garantiza una correlación en el orden del 95 %. Por su parte cada curva de energía se representó utilizando polinomios de orden 20, que muestra un valor de un 90 % en la correlación entre la aproximaciones y las curvas originales, en promedio para el conjunto de datos disponibles.

En la figura 40 se muestra la aproximación de la curva de F0 correspondiente a una frase entonativa, empleando el ajuste polinómico detallado.
El conjunto completo de parámetros con el cual se representó cada grupo entonativo quedó compuesto por:

- duración del grupo entonativo.
- 10 coeficientes polinómicos que representan la curva de F0.
- 20 coeficientes polinómicos que representan la curva de energía.

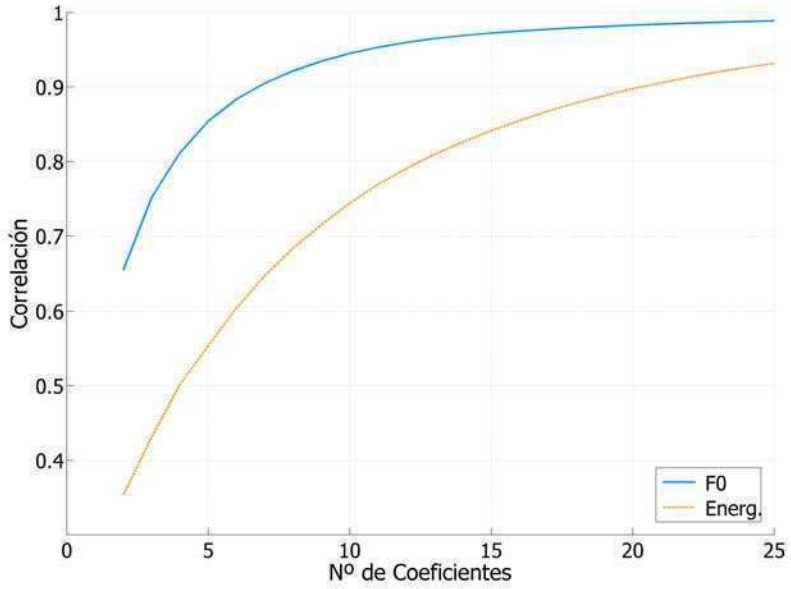

Figura 39: Valores promedio para el conjunto completo de frases, del coeficiente de correlación entre las señales originales de energía y F0 y sus aproximaciones polinómicas, empleando distintos órdenes polinómicos en su representación.

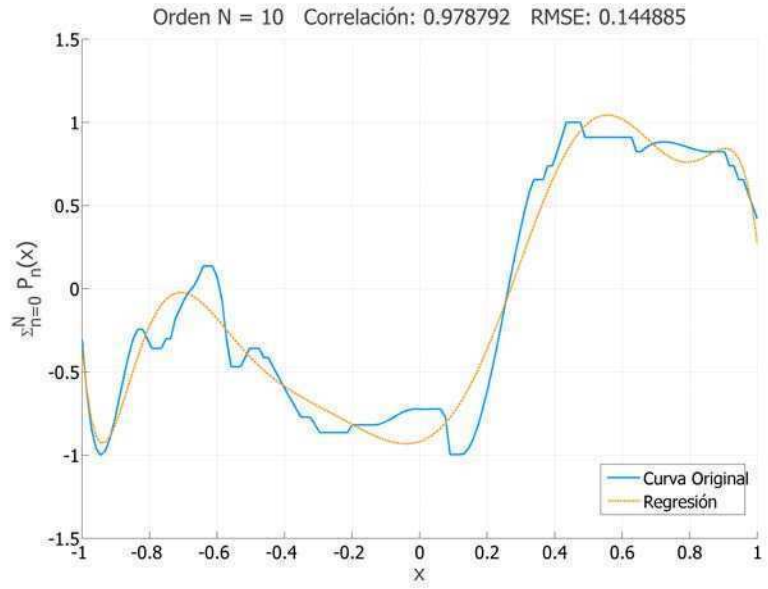

Figura 40: Curva de F0 preprocesada (sólida), y su aproximación con polinomios de Legendre de orden 10 (en línea de puntos), correspondiente a la frase: *"carga entre quince"*.

Una vez obtenida la representación paramétrica se procedió a realizar un agrupamiento no supervisado.

4.2.2 Agrupamiento de Frases Entonativas

El objetivo del agrupamiento es conocer, sin efectuar suposiciones a priori, si es posible encontrar conglomeraciones naturales en un conjunto de datos. Para resolver el problema de agrupamiento de las frases entonativas se decidió utilizar una de técnicas considerada entre las más eficientes para esta tarea: los mapas autoorganizados [84].

redes neuronales de Kohonen

Inspiradas en la organización neuronal de la corteza cerebral, los mapas autoorganizados **SOM** (del inglés: *Self-Organizing Map*), también denominados redes de Kohonen en reconocimiento a quien las propuso, son un tipo especial de redes neuronales artificiales. Se entrenan mediante aprendizaje no supervisado y competitivo, para producir una representación de baja dimensionalidad y discretizada del espacio de entradas presente en los datos de entrenamiento, a través de un conjunto de neuronas ordenadas en un arreglo geométrico o mapa.

Este tipo de redes garantizan que la vecindad en el espacio de entradas se preserve en su representación topológica [95].

Se evaluaron diferentes dimensiones de redes y tipos de vecindades. Los modelos adoptados finalmente quedaron determinados por mapas bidimensionales de 20 x 20. Tal determinación se basó en una relación de compromiso entre facilidad de distinción visual de los agrupamientos por una parte, y la necesidad de contar con suficiente cantidad de muestras para entrenar el conjunto de parámetros libres del modelo por otra.

Tras una inicialización aleatoria de los pesos correspondientes a cada unidad de los mapas, el entrenamiento de estas redes se llevó a cabo en dos fases: una de organización y otra de convergencia.

Para la fase de organización se utilizaron en promedio 5000 épocas de entrenamiento. Durante ese lapso la función de vecindad (de tipo gaussiana) se redujo de forma gradual desde un valor inicial de 10 a 1 neurona.

En la fase de entrenamiento se emplearon en promedio 50000 épocas. Se empleó un valor de radio inicial para la vecindad de 1, el cual al final del entrenamiento se hizo decrecer hasta 0,01.

La figura 41 muestra la Matriz-U correspondiente a un mapa obtenido a través del procedimiento detallado. Se puede ver dos regiones bastante definidas en esta representación. Una extensa, ubicada en la parte central e inferior del mapa, y otra menor en la parte superior izquierda.

Una de las ventajas del procedimiento de parametrización y agrupamiento detallado es que permite sintetizar la forma de onda correspondiente a cualquier neurona. Así, elegida una neurona específica en el mapa que se distinga como el centro de un grupo o cluster, sus pesos asociados tienen una interpretación directa. Esos pesos contienen el valor de los atributos de una curva prototipo asociada con el grupo. En nuestro caso el conjunto de atributos de energía, F0, a través de los coeficientes de funciones de Legendre, y de duración.

Por otro lado, gracias a las propiedades ya mencionadas de los polinomios de Legendre, es posible utilizar un conjunto de coeficientes para sintetizar la curva correspondiente.

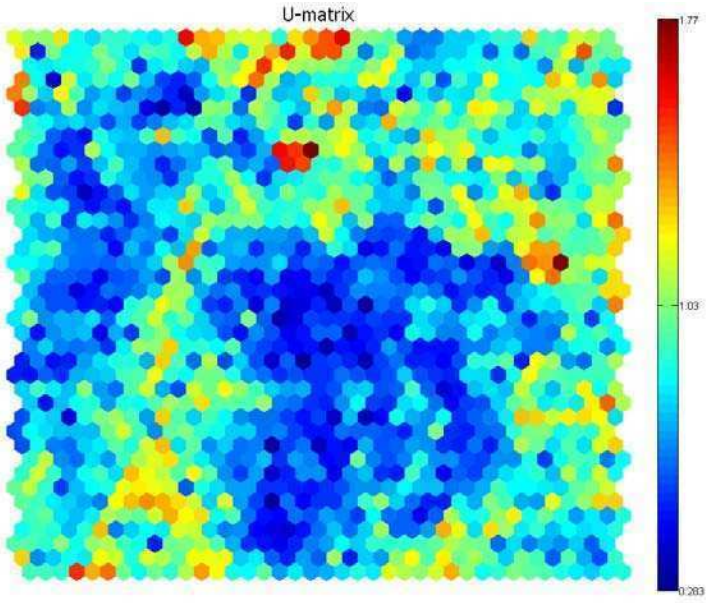

Figura 41: Matriz-U correspondiente a un Mapa autoorganizado obtenido luego del proceso de entrenamiento. En esta representación se muestran codificados mediante colores las distancias Euclídeas entre vectores de pesos vecinos.

Analizando de esa manera los *clusters* conformados se pudo distinguir algunos patrones vinculados a las formas globales de la tendencia de la frase (ascendente o descendiente), y de sus números de picos.

Ello motivó que se realizara un análisis a nivel de tipos de frases que permitieran explicar los conglomerados constituídos.

La caracterización de frases que finalmente se asociaron con los clusters observados fue la siguiente:

- *Clase* 1*:* grupo único de la oración.
- *Clase* 2*:* primer grupo de una oración con más de un grupo.
- *Clase* 3*:* grupo intermedio de una oración con al menos tres grupos.
- *Clase* 4*:* último grupo de una oración con al menos dos grupos.

Considerando tales clases, se identificó en el conjunto completo de datos el número de instancias de cada tipo. En la figura 42 se muestra un histograma que indica el conjunto de casos disponibles:

Finalmente se planteó la posibilidad de verificar la asociación entre esos grupos detectados y las clases que se definieron.

Para ello se empleó la red de Kohonen como un clasificador: una vez entrenada, se etiquetó cada neurona con una identidad de clase. Dicha asignación de clases a cada neurona se efectuó por votación simple, de acuerdo a la cantidad de instancias de entrenamiento de cada clase que se asociaran con la neurona.

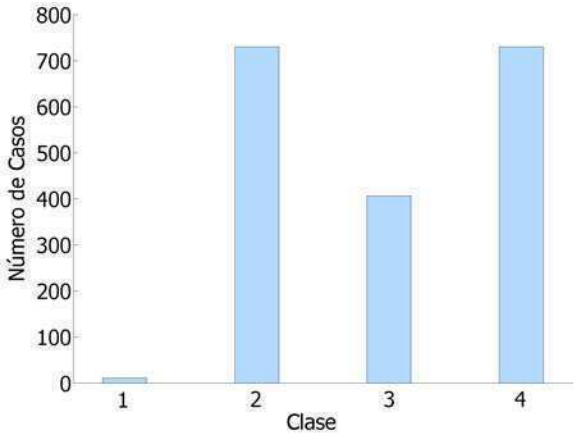

Figura 42: Histograma de frecuencia por clases de grupos entonativos.

Utilizando la red así etiquetada, se le presentó cada ejemplo de un conjunto de datos de evaluación, disjunto al empleado durante el entrenamiento. Finalmente se clasificó cada ejemplo de entrada de acuerdo a la clase (neurona) que activó, y se realizó un estudio sobre la bondad de esta red como clasificador.

La tabla 2 presenta los resultados de la validación cruzada empleando 6 particiones diferentes del conjunto de datos, y el procedimiento descripto. En todas se empleó un 80 % de los datos disponibles para entrenamiento y el 20 % restante para evaluación.

Prueba	Precisión de Asociación				
	Clase 1	Clase 2	Clase 3	Clase 4	**General**
1	13,00	78,63	49,75	89,86	76,29
2	17,37	80,14	37,44	84,66	72,35
3	7,27	78,90	52,71	91,37	77,78
4	9,09	75,34	54,93	90,82	76,56
5	14,28	75,75	50,98	90,55	75,87
6	10,27	76,99	50,00	92,33	76,82

Cuadro 2: Resultados de clasificación empleando SOM para 6 particiones diferentes del conjunto de datos disponibles. **Clase 1**:Frase única de la oración; **Clase 2**: Primera frase de la oración; **Clase 3**: Frase intermedia; **Clase 4**: Última frase de la oración.

Entre las conclusiones obtenidas de este análisis se pueden citar las siguientes:

Empleando los mapas entrenados y habiendo asignado una etiqueta de clase a cada nodo de la red, utilizando las categorías de frase mencionadas, se clasificaron datos no vistos en el entrenamiento. Los resultados de esta clasificación mostró una tasa de clasificación de $75-80\,\%$ para la clase 2: primera frase de la oración; de un 90 % para la clase 4:

frase final de la oración, y desempeños muy inferiores para las demás clases.

Esto va en línea con observaciones previas sobre la mayor homogeneidad de grupos entonativos finales de oraciones, para el corpus empleado. Tales similitudes provocan que se genere en las redes autoorganizadas una región correspondiente a instancias de frases de este tipo.

Algo similar ocurre con la clase 2 (primer grupo de oraciones multigrupo), mientras que existe más problemas al tratar de verificar clusters de grupos intermedios. Esto indica que los grupos entonativos iniciales y finales para este tipo de oraciones son más similares entre sí que los intermedios.

Los malos resultados obtenidos al considerar frases únicas de oración se explican por el escaso número de ejemplos disponibles para esa clase, y el procedimiento empleado para asignar una identidad de clase a cada neurona de la red.

4.3 CORRELACIÓN ENTRE PALABRAS DE CONTENIDO Y PICOS TONALES

Como se presentó en la Sección 3.2.6, una de las funciones del acento en el Español es la contrastiva, que le permite distinguir entre palabras acentuadas y no acentuadas. Las palabras acentuadas en el habla de nuestro idioma se pueden asociar con la clase de palabras de contenido, o clase de palabras abiertas, mientras que las inacentuadas tienen una correspondencia con las palabras de función, o de clase cerrada.

Poder establecer el número de palabras de contenido presentes en una frase entonativa, a través sus rasgos suprasegmentales significaría información útil para los sistemas de RAH. Se la podría utilizar por ejemplo para desambiguar hipótesis de reconocimiento.

Es así que se decidió estudiar la morfología de las curvas de F0 en relación al número de palabras de contenido de cada grupo.

La investigación desarrollada en esta Sección se apoya en los resultados que para el mismo cuerpo de datos que el utilizado en este Capítulo se presentan en [26]. En ese estudio, un análisis del conjunto de etiquetas fonológicas asociadas a los datos reveló que un 84 % de las palabras de contenido presentaron acento tonal de tipo H*, mientras que 10 % exhibieron un pico retardado del tipo L*+H. Con eso resultados los autores concluyeron que para el Español leído de Buenos Aires hay un acento tonal alto en casi todas las palabras de contenido, excepto para la de final de oración.

En primer lugar se determinó de manera automática el número de palabras de contenido a partir de las anotaciones de clases sintácticas, presentes en el corpus empleado. Se consideraron como palabras de contenido aquellas con etiquetas:

- **N**: número cardinal

- **D**: adjetivo
- **S**: sustantivo
- **M**: pronombre
- **V**: adverbio
- **I**: verbo infinitivo
- **P**: verbo participio/gerundio
- **O**: verbo conjugado
- **R**: verbo auxiliar

Las demás palabras se consideraron de función.

El proceso de exploración llevado a cabo en la sección anterior permitió detectar que el tipo de grupo entonativo (inicial, intermedio, final, único) imponía una morfología general a las curvas de entonación, sobre la que se superponía la información de los picos correspondientes a acentos tonales.

Debido a esos efectos, se decidió realizar las pruebas de esta sección discriminando los resultados a nivel de las cuatro clases de frases descriptas en la sección anterior. Esto permitiría desagregar los resultados obtenidos en la detección automática de picos de F0 para cada contexto.

4.3.1 Detección del Número de Picos

Con las curvas de F0 estimadas de la forma descripta en la sección anterior, se construyeron dos algoritmos para la detección de los picos presentes en las mismas.

El primer método hace uso de del algoritmo Momel (ver Sección 3.5.2) para estilizar las curvas antes de buscar sus picos. Debido a los efectos de la microprosodia, las curvas de F0 presentan diversos ascensos locales que no están vinculados con acentos tonales, sino con las características de las unidades a nivel segmental. El algoritmo de estilización busca filtrar esas variaciones locales, produciendo una curva más suave y con mayor similitud a la entonación percibida. Para nuestros fines, la aplicación del algoritmo eliminará picos que no se correspondan a los acentos tonales.

El segundo método aprovecha el procedimiento de aproximación polinómica de Legendre, detallada en la sección anterior, para suavizar las curvas de F0 antes de efectuar la búsqueda de picos.

Una vez preprocesadas las señales de las dos formas descriptas, ambos detectores de picos efectúan una búsqueda de máximos locales a través de las derivadas de primer orden de las funciones correspondientes a las curvas de F0. También se utiliza un umbral de distancias mínimas a las que se pueden ubicar dos picos, y umbrales de amplitud en relación a la del pico máximo encontrado.

Por otra parte, un aspecto particular que muestran las curvas de F0 viene determinado por el fenómeno de declinación, descripto en la Sección 3.2.5. Producto de este fenómeno, la tendencia descendiente de la curva de F0 puede provocar la subestimación de algunos picos.

Por ello se estudiaron alternativas para la compensación automática de este efecto. En la literatura se pueden encontrar diferentes procedimientos para determinar la declinación y emplear esa información en el proceso de compensación, entre ellas:

- Calcular el F0 de la primer y última sílabas acentuadas de la oración y unir ambos valores con una recta.
- Hacer una regresión lineal empleando solo los picos de la curva de F0.
- Hacer una regresión lineal empleando solo los valles de la curva de F0.
- Hacer una regresión lineal empleando todos los puntos de la curva de F0.

Finalmente, se usó un procedimiento similar al presentado en [106], que básicamente emplea la última de las alternativas mencionadas.

Específicamente se decidió realizar una regresión lineal sobre los segmentos de frases delimitados por silencios, ajustando sólo los puntos de la curva de F0 que de acuerdo al algoritmo RAPT, se encontrasen dentro de segmentos especificados como sonoros. Además se eliminaron en dicha regresión las porciones iniciales y finales de cada segmento, por ser regiones susceptibles de presentar artefactos tanto causados por la microprosodia, como por el algoritmo de estimación de F0.

Se efectuaron varias pruebas para seleccionar la mejor combinación de parámetros para los métodos de preprocesamiento descriptos.

En la tabla 3 se presentan los resultados de emplear los dos procedimientos de detección de picos descriptos, bajo dos condiciones distintas: usando directamente las curvas de F0, y realizando una compensación por declinación previa.

Estos resultados muestran el desempeño de los detectores de picos empleados como estimadores del número de palabras de contenido presentes en las frases evaluadas, y se reportan en términos de precisión de clasificación.

En todos los casos se utiliza la mejor combinación de parámetros que permita maximizar las tasas de clasificación.

En la tabla 3 el detector designado como **MOM** empleó como parámetros para el algoritmo MOMEL un ancho en la ventana de análisis $A = 150$ ms, un valor de umbral de distancias $D = 5\,\%$, una ventana de reducción $R = 200$ ms. En la Sección 3.5.2 se describen estos parámetros.

Para el detector **MOM-D**, la combinación de parámetros óptimos vino dado por: $A = 200$ ms, $D = 5\,\%$, $R = 200$ ms.

Por su parte el detector indicado como **Leg** utiliza una aproximación polinómica de orden 5, y el detector **Leg-D** emplea polinomios de orden de 7.

Prueba	Frase	Precisión por Nº de CW (%)					Prec. Frase (%)	Prec. Tot. (%)
		1	2	3	4	5		
MOM	1	0,0	0,0	22,2	50,0	0,0	0,0	43,9
	2	55,4	35,0	28,3	50,0	100,0	45,7	
	3	61,6	23,7	38,9	33,3	0,0	46,1	
	4	54,4	39,9	34,0	23,7	16,7	41,4	
MOM-D	1	53,2	29,2	23,3	0,0	0,0	41,8	44,6
	2	55,4	35,0	28,3	50,0	100,0	45,7	
	3	54,4	23,7	22,2	33,3	0,0	41,2	
	4	66,0	45,1	38,4	42,1	33,3	48,9	
Leg	1	0,0	0,0	0,0	0,0	0,0	0,0	43,5
	2	46,7	70,2	0,0	0,0	0,0	51,0	
	3	60,2	68,9	0,0	0,0	0,0	59,9	
	4	58,1	29,9	0,0	0,0	0,0	28,9	
Leg-D	1	0,0	50,0	20,0	0,0	0,0	16,7	46,8
	2	66,5	53,9	15,2	0,0	0,0	57,3	
	3	66,0	47,4	0,0	0,0	0,0	55,2	
	4	54,4	35,1	10,7	0,0	0,0	32,7	

Cuadro 3: Comparación de la correlación entre número de palabras de contenido, y de picos en la curva de F0 empleando distintos detectores. Se indican los valores de precisión al clasificar el número de palabras de contenido a través de la cantidad de picos detectados en las curvas de F0 correspondientes. **MOM**: detector de picos empleando el algoritmo Momel en el preprocesamiento de F0; **MOM-D**: similar al anterior pero compensando el efecto de la declinación en las curvas de F0; **Leg**: detección de picos sobre la aproximación polinómica de Legendre de las curvas de F0; **Leg-D**: similar al anterior pero compensando el efecto de la declinación. Los resultados se discriminan por número de palabras de contenido (CW) y tipos de frases, 1:Frase única de la oración; 2: Primera frase; 3: Frase intermedia; 4: Última frase de la oración.

Como se puede observar, los resultados brindados en la tabla 3 no permiten aceptar la hipótesis indagada en esta sección.

Es decir que no se encontraron relaciones directas entre picos de la curva de F0 y número de palabras de contenido.

Se pueden aducir varias razones para intentar explicar esta situación.

La primera podría estar relacionada con el estilo de habla de la informante. En la bibliografía se puede encontrar reportes donde se indica que el estilo de habla de los locutores profesionales suele contener acentos enfáticos. Aunque las palabras acentuadas en el Español contienen un solo acento enfático, en estos casos pueden aparecer más de dos sílabas acentuadas en una palabra, por ejemplo: *"es mi **RES**ponsabili**DAD**"*. El objetivo de este acento es poner de relieve una palabra determinada, poner énfasis a ciertas partes del enunciado, o distinguir dos enunciados que podrían resultar confusos.

Otra razón podría atribuirse a errores en la deteminación de las curvas de F0. Como se mencionó en el apartado 3.5.1, aunque parezca una tarea sencilla, la estimación automática del F0 está sujeta a una serie de errores, que los algoritmos actuales aún no son capaces de resolver completamente. Estos posibles errores podrían traer aparejados picos anómalos de F0, y confundir el proceso de sus asociaciones con acentos tonales.

También puede suceder que si bien cada palabra de contenido esté asociada con un acento tonal, el acento tonal no presente como correlato un pico de F0 sino por ejemplo un cambio en la tasa de ascenso o descenso; o incluso que ese acento se exprese utilizando los otros atributos suprasegmentales como duración e intensidad, en vez de hacerlo mediante el F0.

Finalmente, es probable que en el habla continua no todas las palabras de contenido se acentúen, y que por otra parte no todas las palabras de función se manifiesten inacentuadas. Por lo tanto deberían existir relaciones un poco más complejas entre las estructuras sintácticas y las posibilidades de encontrar prominencias en el perfil de F0.

4.4 CLASIFICACIÓN DE ACENTO LÉXICO EN PALABRAS FINALES DE FRASE

Como se mencionó en las Secciones 1.4.1 y 3.2.7, los rasgos prosódicos observados en una frase son el resultado de la superposición de múltiples factores. De esta forma, es posible encontrar diversas morfologías a nivel de atributos suprasegmentales correspondiente a una misma palabra.

Esto conspira con el objetivo de correlacionar patrones suprasegmentales específicos respecto a cierta información léxica, para aplicar esa información en el proceso de RAH.

Sin embargo es posible encontrar regiones específicas dentro de una elocución, donde algunas restricciones lingüísticas acotan los grados de libertad de las variaciones suprasegmentales. Una de esas regiones las constituyen los segmentos finales de frase.

El número de variantes posibles en los finales de frase es acotado, y además se han propuesto métodos automáticos para la clasificación relativamente eficiente de los mismos [148].

Partiendo de esas observaciones, en esta sección se evalúa la posibilidad de encontrar un comportamiento distintivo en los patrones suprasegmentales, que permita caracterizar palabras agudas, graves y esdrújulas de final de oración.

Como todas las oraciones disponibles en el corpus utilizado en estos experimentos son declarativas, pronunciadas por una misma locutora y aproximadamente con el mismo ritmo, al acotar el análisis a las palabras finales de oración es posible reducir la variabilidad que tienen los rasgos prosódicos de acuerdo a la ubicación del segmento considerado dentro de la frase.

A continuación se describirán las dos partes en que se puede agrupar los experimentos de este apartado.

4.4.1 Curvas de F0 Prototipo para Clases de Acento Léxico

Con el objetivo de saber si es posible encontrar un perfil de F0 prototípico que permita diferenciar las palabras finales de oraciones declarativas de acuerdo a sus tonicidades léxicas, se comenzó segmentando las curvas de F0 para las palabras finales de oración del corpus disponible, recurriendo para ello a los archivos de anotaciones grafémicas (*.gra), como el mostrado en la figura 38.

Por otra parte, se desarrolló un segmentador silábico, y un clasificador de acentos léxicos, ambos basados en reglas. Estos programas en conjunto permitieron determinar de manera automática el tipo de palabras finales de oración según sus acentos léxicos.

En la tabla 4 se presenta la frecuencia de ocurrencia de cada tipo de clase de palabra según sus tonicidades léxicas en posición final de oración para el conjunto de datos empleado:

Clase Acentual	Número de Casos
Monosílabas	11
Agudas	154
Graves	507
Esdrújulas	66
Sobreesdrújulas	3

Cuadro 4: Cantidad de oraciones para cada tipo de acento léxico en palabras finales de oración del corpus LIS-SECyT.

Además se encontró que 733 sobre un de 741 palabras en posición final de oración eran diferentes.

Para determinar las curvas de F0 prototípicas por clase acentual se promediaron los contornos de F0 correspondientes a las palabras con el mismo tipo de acento léxico. Debido al escaso número de ejemplares disponibles se dejaron fuera de este análisis a las palabras monosílabas y sobreesdrújulas.

Se estudiaron diferentes alternativas para llevar a cabo ese promedio. A continuación se detalla el procedimiento definitivo.

Como las duraciones de palabras pertenecientes a una misma clase acentual eran muy variables, se hacía imprescindible introducir algún tipo de manipulación temporal, que posibilitase generar un promedio que tuviera sentido.

Se descartó normalizar temporalmente todas las curvas, dado que de hacerlo se estarían comparando morfologías artificiales y perdiendo información.

Por otra parte, como se sabía que las oraciones estaban pronunciadas aproximadamente a la misma velocidad, con el mismo ritmo, y

presentaban un comportamiento global similar (por tratarse de oraciones declarativas), es decir que todas terminaban en un descenso del F0; se decidió promediar las curvas alineándolas desde sus extremos finales.

Antes de llevar a cabo el promedio se eliminaron los intervalos sordos iniciales y finales de cada palabra.

Los segmentos sordos en el interior de las palabras, pueden distorsionar la forma general del promedio resultante. Como en este caso se consideraron palabras individuales, por lo que estaba garantizado que no habían silencios interiores de palabra que correspondieran a pausas, se decidió interpolar las curvas de F0 correspondientes a cada palabra utilizando sólo los segmentos sonoros de cada curva. La interpolación se realizó usando funciones de tipo *spline*.

Esta operación también se puede justificar desde el punto de vista perceptual, ya que como fuera mencionado en el apartado 3.2.5, el cerebro integra perceptualmente los breves segmentos sordos como los mencionados, generando la sensación de continuidad en el contorno de F0.

A continuación se presentan las curvas promedio obtenidas para cada clase acentual.

La figura 43 muestra la curva prototípica obtenida promediando los perfiles de F0 correspondientes a palabras finales agudas.

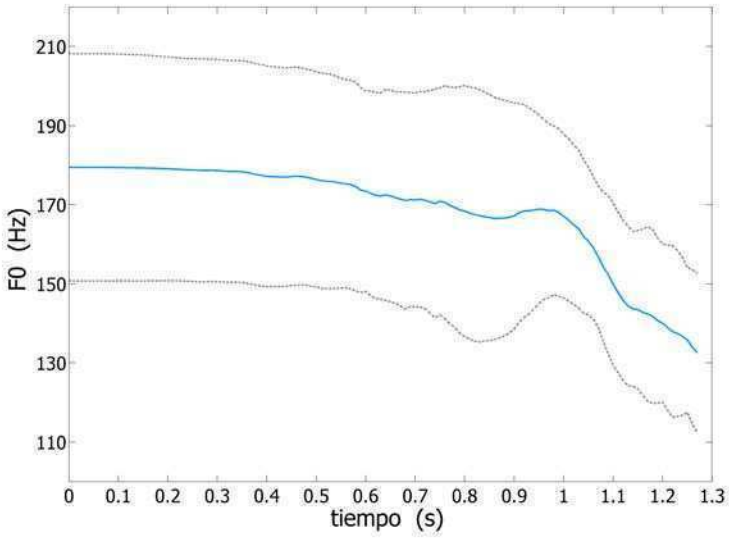

Figura 43: Curvas promedio del contorno de F0 para palabras finales agudas. En línea de puntos se representa el desvío estándar medido punto a punto al efectuar el promedio.

Para el caso de palabras finales agudas, la curva promedio presenta un pico distintivo que correspondería con la ubicación de la sílaba tónica. Ese pico tiene aproximadamente 150 ms de duración y comienza unos 300 ms antes de finalizar la palabra.

En la figura 44 se presenta la curva promedio para las palabras finales graves.

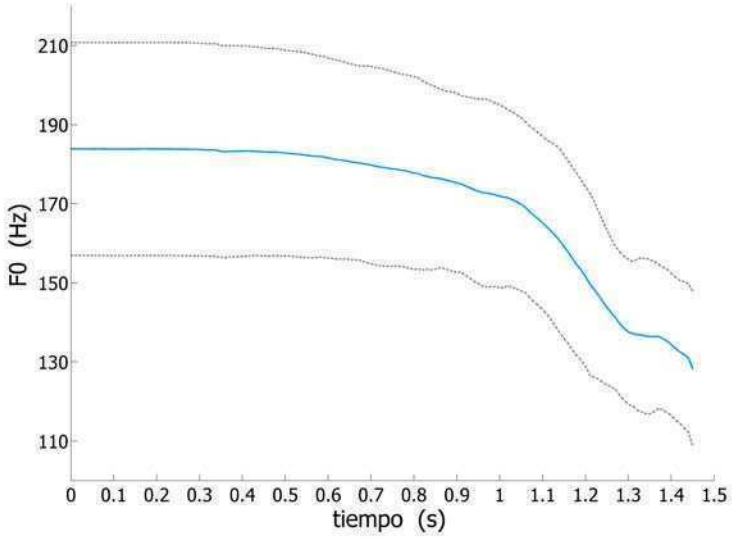

Figura 44: Curvas promedio del contorno de F0 para palabras finales graves. En línea de puntos se representa el desvío estándar medido punto a punto al efectuar el promedio.

En el caso de palabras graves, no se aprecia un pico que sobresalga como en el contorno correspondiente a palabras agudas. El pico que uno esperaría observar en la penúltima sílaba no aparece reflejado en la curva promedio.

Uno de los posibles argumentos para explicar esta situación podría ser el desbalance en el número de instancias, como se indica en la tabla 4. Al promediarse muchos mas ejemplos que en las otras clases, la curva resultante es más probable que resulte *"suavizada"*.

Por otro lado, en este caso aparece una pequeña inflexión final, que en el caso del prototipo correspondiente a palabras agudas es menos significativo. Finalmente en el caso de palabras graves, en un contexto de descenso global del F0, se distingue un aumento en la tasa de descenso a unos 400 ms previos al final de las palabras. Ese mismo comportamiento para el caso de las agudas se da a unos 250 ms del final.

La figura 45 refleja el comportamiento de la curva de F0 correspondiente al promedio de las palabras finales esdrújulas.

En el caso de la figura 45 se puede ver que para las palabras esdrújulas existe una protuberancia notable, como en el caso de las agudas, con una duración también equivalente, pero comenzando bastante antes, aproximadamente unos 600 ms previos al final de la oración.

Finalmente también se aprecia un último pico hacia el final del trazado, similar al encontrado en las palabras graves pero más pronunciado que en ese caso y con una duración aproximada de 100 ms.

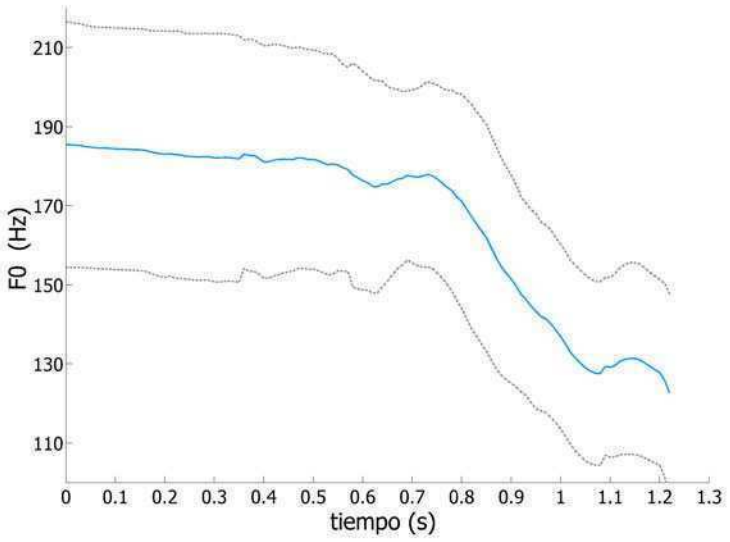

Figura 45: Curvas promedio del contorno de F0 para palabras finales esdrújulas. En línea de puntos se representa el desvío estándar medido punto a punto al efectuar el promedio.

En la figura 46 se resume el comportamiento de las curvas de F0 correspondientes a los tres tipos de palabras considerados.

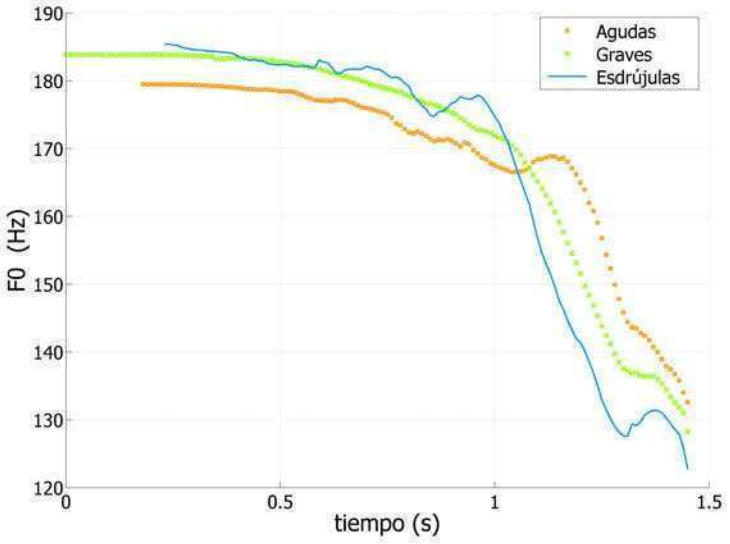

Figura 46: Comparación de los contorno de F0 prototípicos para palabras finales de oración agudas, graves y esdrújulas. En la gráfica las tres curvas se muestran alineadas a partir de sus extremos finales.

Observando las tres curvas de la figura 46, si bien todas presentan un decaimiento hacia el final, por corresponder a oraciones declarativas, se puede distinguir una morfología diferente para cada clase acentual.

De esta forma, en principio se podría considerar viable el empleo de la forma de las curvas de F0 de oraciones declarativas para conocer el tipo de acento de la palabra final de la oración.

Considerando válida la observación anterior, en una serie de estudios posteriores exploramos la asociación entre los acentos tonales y los parámetros de frecuencia fundamental obtenidos a partir del modelo de Fujisaki (ver Sección 3.4.2).

El método de asociaciones presentado en esos trabajos permite llevar a cabo la clasificación automática de clases acentuales de palabras, ubicadas en sectores específicos de una oración, y de esa forma llevar a la práctica los resultados de este apartado.

En esos estudios se emplearon oraciones más controladas que las de la base LIS-Secyt, diseñadas específicamente para el estudio de las manifestaciones prosódicas en estructuras sintácticas bien definidas. Este grupo de datos forman parte del proyecto AMPER (**A**tlas **M**ultimedia de la **P**rosodia del **E**spacio **R**ománico) [43].

Por las características del diseño experimental, en el caso de AMPER-Argentina no existe el problema del desbalance en la cantidad de ejemplos por clases de palabras, que se comentó respecto al corpus empleado en este apartado.

En [67] estudiamos frases declarativas sin expansión, que presentan la estructura mostrada en la figura 47.

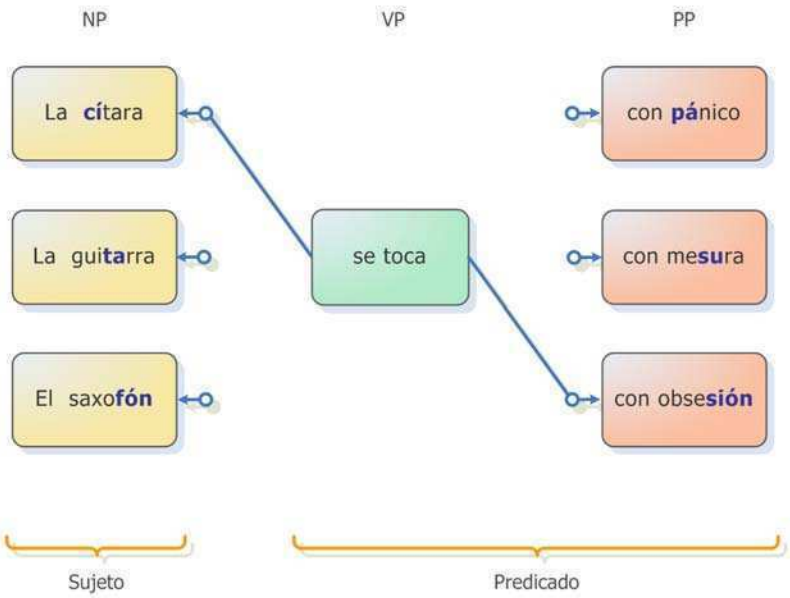

Figura 47: Gramática utilizada para la generación de las oraciones AMPER-Argentina. NP: frase nominal; VP: frase verbal; PP: frase preposicional.

El conjunto de oraciones estudiadas resultó de la combinación de palabras mostradas en la figura 46. La combinación dada por esa gramática hace que las oraciones presenten los tres tipos de acentos léxi-

cos tanto en el final de la frase fonológica (ver 3.3.4) como al final de la frase entonativa (3.3.5).

El material acústico de AMPER-Argentina se obtuvo por medio de la técnica de entrevista dirigida (conocida como *elicitation task* en Inglés): los informantes responden de forma natural a distintas interrogaciones que se le formulan, es decir que no se trata de habla leída como en el caso de LIS-Secyt.

Con este conjunto de oraciones se analizaron los comandos de acento correspondientes al modelo de Fujisaki asociado al acento tonal de la última palabra de sujeto y predicado, y su asociación con las sílabas tónicas de esas palabras.

Mediante ese procedimiento se pudo determinar diferencias en los alineamientos de los picos de F0 en relación al tipo acentual, y describir sus comportamientos en función al tipo de frase considerada.

En la tabla 5 se presentan los resultados obtenidos para frases declarativas.

Tipo Palabras	Tipo Acento	μ(T1) (ms)	σ(T1) (ms)	μ(T2) (ms)	σ(T2) (ms)
Final	Aguda	−33	89	260	111
	Grave	−45	100	268	141
	Esdrújula	−62	87	235	136
No Final	Aguda	−423	208	−137	120
	Grave	−364	250	−102	146
	Esdrújula	−372	219	−60	109

Cuadro 5: Valor medio (μ) y desvíos estándar (σ) para los parámetros correspondientes a comandos de acento T1 y T2, para cada tipo de sílaba acentuadas en palabras acentuadas finales y no finales.

En la tabla 5 se puede ver que el valor del parámetro T1 se separa progresivamente del ataque vocálico correspondiente al núcleo de la sílaba acentuada.

Por otro lado, el valor de T2 se aproxima progresivamente al ataque vocálico para palabras acentuadas en posición final.

Estos ordenamientos pueden indicar una aparente correspondencia uno a uno entre T1 y T2 asociado al tipo de sílaba.

Para verificar esa relación, se efectuó un análisis de varianza de T1 y T2 para las tres posiciones silábicas. Para el caso de T2 en palabras acentuadas de final de oración, este análisis mostró que los alineamientos difieren significativamente en función del tipo de sílaba ($df = 2$, $F = 6{,}38$, $p = 0{,}002$), pero no así para T1 en posición de palabra no final.

Por otro lado en [66] efectuamos un análisis equivalente pero empleando frases interrogativas.

Esta serie de experimentos sobre el corpus AMPER corroboraron la idea propuesta en este apartado, y se pueden considerar como un ejemplo de su implementación.

4.4.2 Análisis Estadístico de Rasgos Prosódicos en Función de Acentos Léxicos

En este apartado se buscó clasificar la tonicidad de las palabras finales de oración siguiendo una metodología diferente a la detallada anteriormente.

Los estudios presentados intentan definir si es posible determinar la tipología acentual de las palabras finales de oración analizando el valor de los atributos prosódicos de las tres últimas sílabas de estas palabras.

Para realizar el estudio se emplearon los archivos de anotaciones silábicas ∗.or2 del corpus LIS-Secyt. En la figura 48 se muestra un ejemplo de estas anotaciones, correspondiente al mismo caso mostrado de la figura 38.

Secyt-m1-1.or2	
0,677868	al
0,924597	Ba
1,077550	Res
1,390222	se
1,622957	a
1,724597	Bi
1,958972	a
2,046472	a
2,146472	ni
2,302722	ma
2,522731	Do
2,736183	a
2,822590	kon
3,059878	tar
3,295754	le
3,452963	un
3,678301	Hih
3,942276	te

Figura 48: Archivos de anotaciones silábicas correspondientes a la oración *"Álvarez, se había animado a contarle un chiste"*. La primera columna indica el tiempo de inicio de la anotación (en segundos), y la segunda la identidad silábica en notación fonética Sampa.

A diferencia de los archivos de anotaciones de la figura 38, en este caso los archivos de anotaciones silábicas no fueron realizadas de forma manual, sino que se emplearon reglas de silabificación para el español y los archivos de anotaciones fonéticas correspondientes.

Con los archivos ∗.or2 se determinaron los instantes de inicio y fin para cada una de las tres sílabas finales de cada oración.

Sobre los segmentos así determinados se efectuó un análisis estadístico de duraciones, valor máximo de F0 y de energía para las tres últimas sílabas de cada oración en relación al tipo de acento de cada palabra final.

En los análisis efectuados se consideraron para todas las clases de palabras los valores de rasgos suprasegmentales para las tres últimas sílabas, eso hace que para algunas palabras (por ejemplo en el caso de monosílabas, ciertas graves y agudas), algunas de las sílabas consideradas no pertenezcan a la palabra considerada.

En la figura 49 se muestran los boxplots para el máximo de F0 de las tres sílabas finales, ordenadas por clase de acento léxico.

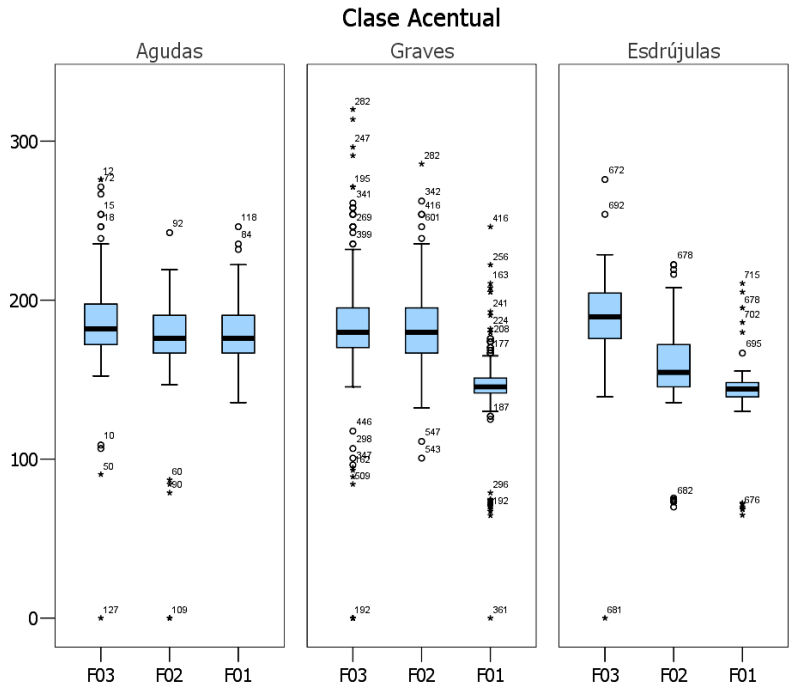

Figura 49: Boxplots para las tres últimas sílabas para la variable máximo de Fo. En cada caso se emplea la siguiente notación, F03: máximo de F0 en la ante-penúltima sílaba de la oración correspondiente; F02: máximo de F0 en la penúltima sílaba de la oración; F01: máximo de F0 en la última sílaba de la oración.

En la figura 49 se puede ver un patrón coherente en todas las clases acentuales: un descenso abrupto en el valor de F0 después de la sílaba con acento léxico. Para el caso de las palabras agudas, donde se observa un valor muy equilibrado para las tres sílabas, se puede suponer válida la afirmación anterior, pensando que falta la sílaba a la derecha de la tónica para que se exprese el contraste mencionado.

En la figura 50 se presenta una gráfica equivalente para la variable máximo de energía.

En términos generales se puede observar un comportamiento para el máximo de energía similar al descripto para el máximo de F0: disminución abrupta en la magnitud de la variable en la sílaba posterior a la que presenta el acento léxico. Además se puede ver que en las agudas el pico que uno esperaría observar en la sílaba tónica no sobresale

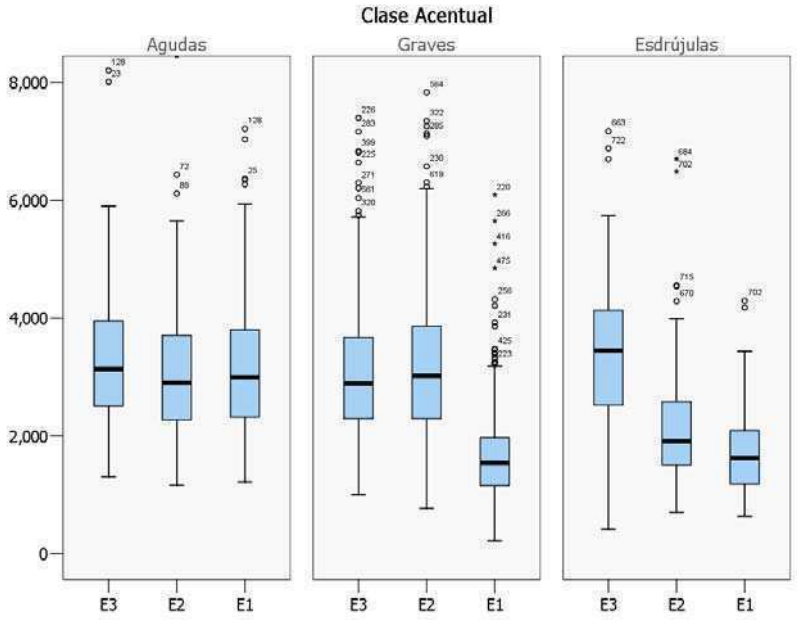

Figura 50: Boxplots para las tres últimas sílabas de cada oración, correspondiente al máximo de Energía para las tres últimas sílabas de cada oración, discriminando por clases acentuales. E3: máximo de Energía en la ante-penúltima sílaba; E2: máximo de Energía en la penúltima sílaba; E1: máximo de Energía en la última sílaba de la oración.

en el valor medio respecto a las demás sílabas, es probable que esté compensado por el perfil descendiente de las oraciones declarativas.

Para las palabras graves la penúltima sílaba manifiesta una mayor energía máxima, coincidiendo con la ubicación de la sílaba tónica. Lo mismo sucede para palabras esdrújulas.

Finalmente en la figura 51 se presenta la gráfica para la variable duración silábica.

Analizando la variable duración se observa lo anticipado por la bibliografía: un alargamiento en la última sílaba en posición final de frase, que se impone sobre el efecto de acento. Sin embargo ese alargamiento es mayor en las agudas porque se suman los factores de frase y acento sobre la misma sílaba.

En las graves la duración de la última sílaba es la mayor (preponderancia del aporte de frase sobre el de acento) aunque la diferencia de duración con la penúltima es menor. También se observa que la sílaba acentuada tiene mayor duración que la precedente.

Para el caso de esdrújulas, la duración de la sílaba acentuada (antepenúltima) es mayor que la penúltima. Se podría concluir que a medida que nos alejamos del final de la frase, el efecto de frase pierde peso, y el alargamiento atribuido al acento se manifiesta con mayor claridad.

En resumen, analizando estos atributos suprasegmentales en relación al tipo de palabras, se puede afirmar que se observan correlatos

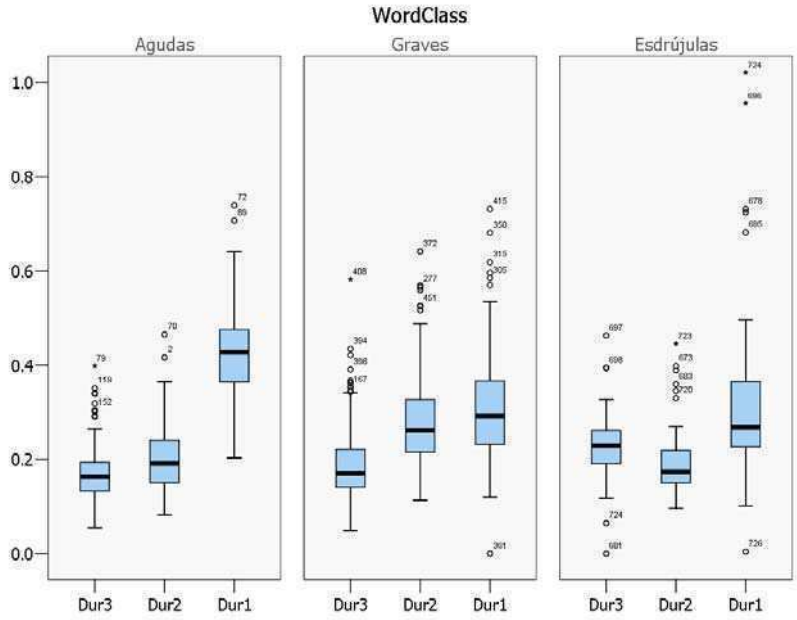

Figura 51: Boxplots correspondiente a las duraciones de las tres últimas sílabas de cada oración, discriminando por clase acentual. Dur3: duración de la ante-penúltima sílaba; Dur2: duración de la penúltima sílaba; Dur1: duración de la última sílaba de la oración.

acústicos característicos para cada clase de acento léxico de final de oración.

A partir de estos resultados se decidió construir un clasificador simple, que permitiese analizar el aporte de las variables individuales al discriminar los tipos de palabras finales de oración en relación de sus acentos léxicos. Para ello se decidió efectuar un análisis discriminante.

De esta forma, empleando como rasgos observados los valores de duración silábica, máximos de energía y de F0 para las tres últimas sílabas de cada palabra final, se implementó un clasificador compuesto por funciones discriminantes.

El análisis discriminante es una técnica estadística que ayuda a identificar las características (variables) que diferencian a dos o más grupos, cuántas de ellas son necesarias para lograr la mejor clasificación posible y crear una función capaz de distinguir a los miembros de cada grupo.

El fin último de este análisis es hallar esa función discriminante, dada por una combinación lineal de las variables independientes que permiten diferenciar mejor a los integrantes de cada grupo.

Una vez hallada, esa función puede ser empleada para clasificar nuevos casos.

En primer lugar se realizó un análisis de ANOVA univariado para determinar si existe alguna variable no significativa, el resultado de dicha prueba se presenta en la tabla 6.

Variable	Est. de Levene	Sig.	F	Sig.
Dur3	3,083	0,046	17,821	0,000
F03	0,149	0,862	0,855	0,426
E3	0,953	0,386	7,067	0,001
Dur2	6,544	0,002	68,747	0,000
F02	2,708	0,067	30,459	0,000
E2	1,876	0,154	19,343	0,000
Dur1	10,898	0,000	64,128	0,000
F01	8,346	0,000	172,659	0,000
E1	22,108	0,000	154,561	0,000

Cuadro 6: Análisis de igualdad de medias (Anova) para las variables duración, máximos de F0 y Energía de las tres últimas sílabas de cada oración.

El estadístico de Levene solamente permite aceptar la hipótesis nula de igualdad de varianzas para el caso de F03, mientras que se debe rechazar esa hipótesis para las demás variables. La igualdad de varianzas entre variables analizadas es un requisito para aplicar Anova. Sin embargo el análisis de Anova muchas veces se puede considerar robusta aunque se viole esa primer restricción.

La prueba de Anova brinda estadísticos denominados (**F**) que permiten contrastar la hipótesis nula de igualdad de medias entre los grupos en cada variable independiente. Se puede ver que si se adopta el umbral de 0,05 como limite para aceptar dicha hipótesis, en ningún caso se debería aceptar la hipótesis de medias iguales.

Uno de los supuestos del análisis discriminante es que todos los grupos proceden de la misma población y que concretamente, las matrices de varianzas-covarianzas poblacionales de cada grupo son iguales entre sí.

Para contrastar la hipótesis nula de igualdad de las matrices de varianzas-covarianzas poblacionales se aplico el análisis descriptivo *M de Box*, cuyo valor fue de 458,86, con un valor de significancia de 0,00, por lo que se debe rechazar la la hipótesis de igualdad de matrices varianza-covarianza entre las clases, y asumir que alguna de las clases varía más que las otras.

Función	Autovalor	Varianza Acumulada (%)	Correlación
1	0,926	79,7	0,693
2	0,235	100,0	0,437

Cuadro 7: Autovalores de las dos funciones discriminantes halladas para clasificar las palabras finales de oración de acuerdo a sus acentos léxicos empleando atributos suprasegmentales.

Dado el resultado anterior se corrió un nuevo análisis que permitiera verificar si usando matrices de covarianza separadas por grupos resultaba en cambios de la clasificación. En esas nuevas pruebas los

cambios en la clasificación no resultaron significativos, por lo que no se usaron matrices de covarianzas separadas por casos.

La tabla 7 presenta los autovalores de las funciones discriminantes halladas. La misma permite comparar de manera global la capacidad discriminante de cada función. En este caso la función 1 explica el 79,7 % de las diferencias existentes entre los casos de cada grupo, mientras que la segunda función logra explicar el restante 20,3 % de dichas diferencias.

En la tabla 8 se brinda el detalle de los coeficientes estandarizados de la función canónica discriminante.

Variable	Función	
	1	2
Dur3	−0,046	−0,402
F03	0,035	−0,075
E3	0,054	−0,467
Dur2	−0,257	0,347
F02	−0,204	0,272
E2	−0,375	0,574
Dur1	0,288	−0,087
F01	0,475	0,443
E1	0,579	−0,031

Cuadro 8: Análisis de igualdad de medias (Anova) para las variables duración, máximos de F0 y Energía de las tres últimas sílabas de cada oración.

Los valores de la tabla 8 se pueden interpretar como el grado de relevancia de cada variable en la construcción de cada una de las funciones discriminantes, a la hora de predecir el grupo de pertenencia de las clases de acentos. Se puede notar por ejemplo que la variable E1 y F01 son las más importantes para la primera función discriminante, mientras que las variables E2, E3, F01, y Dur3 se encuentran entre las más importantes respecto a la segunda función discriminante.

Variable	Clase Acentual		
	Agudas	Graves	Esdrúj.
Dur3	33,940	36,298	46,037
F03	0,100	0,097	0,102
E3	0,002	0,001	0,002
Dur2	33,249	40,487	30,503
F02	0,172	0,190	0,165
E2	−0,001	−0,001	−0,002
Dur1	16,446	10,459	13,904
F01	0,341	0,284	0,268
E1	0,000	−0,001	−0,001
Constante	−66,644	−58,647	−54,386

Cuadro 9: Coeficientes lineales de Fisher para la función de clasificación.

En la tabla 9 se presentan los coeficientes de clasificación de Fisher que permiten obtener la función de clasificación para cada grupo. Para aplicar estos coeficientes en la clasificación de una nueva instancia, se calcula cada una de las funciones de acuerdo al valor de atributos de dicha instancia y se lo clasifica en el grupo para el cual la función muestra el mayor valor.

Finalmente, en la tabla 10 se presentan los resultados de la clasificación obtenidos mediante esta técnica y validación cruzada utilizando la opción 1 de N.

		Grupo de pertenencia pronosticado			Total
	Clase	Agudas	Graves	Esdrúj.	
Recuento	Agudas	132	10	12	154
	Graves	25	408	74	507
	Esdrúj.	6	12	48	66
%	Agudas	85,7	6,5	7,8	100
	Graves	4,9	80,5	14,6	100
	Esdrúj.	9,1	18,2	72,7	100

Cuadro 10: Matriz de confusión resultante de la validación cruzada para el clasificador implementado.

De los resultados observados en la tabla 10 se puede apreciar que tanto para palabras graves como agudas las tasas de reconocimiento superan el 80 %, mientras que para las esdrújulas es de aproximadamente un 73 %.

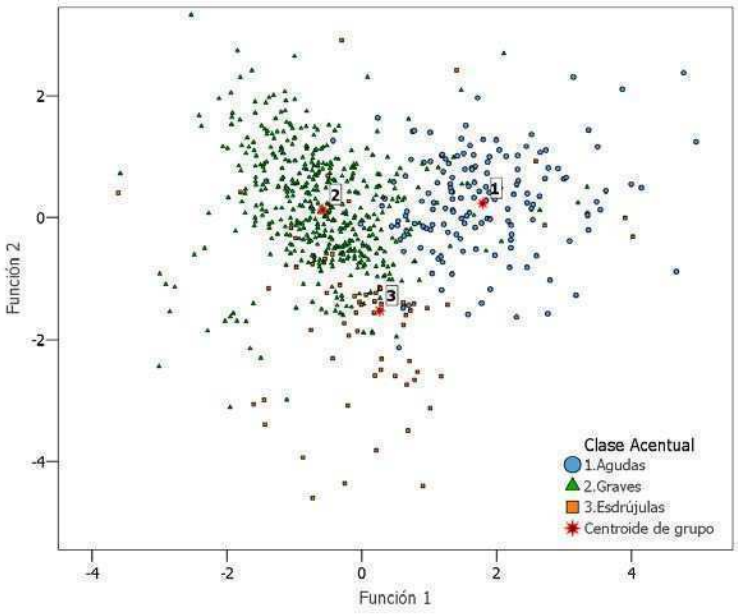

Figura 52: Representación de la clasificación de los ejemplos disponibles empleando las funciones discriminantes canónicas.

En la figura 52 se presenta la clasificación del conjunto de casos disponibles en el corpus LIS-Secyt empleando la función discriminante obtenida en este apartado. Se puede ver una posible separación de instancias, a pesar de un claro solapamiento en algunas regiones.

4.5 CONCLUSIONES

En este capítulo se realizó un análisis exploratorio sobre una serie de hipótesis que vinculan algún tipo de información léxica con atributos suprasegmentales observados.

La primera hipótesis establecía que se podrían establecer agrupamientos de las frases entonativas considerando sus curvas de F0 y energía. Esto se pudo verificar en la primer sección, además se implementó un procedimiento para la clasificación de las frases en esas categorías.

En la segunda parte del capítulo se estudia la posibilidad de correlacionar el número de picos observados en las curvas de F0 con la cantidad de palabras de contenido presentes en las frases.

Los resultados obtenidos en este caso sugieren que no es posible establecer una relación tan directa entre estos dos elementos. Además indican sería conveniente efectuar una detección de prominencias, para realizar la asociación con el número de palabras de contenido presentes en una frase, y no basta con vincular una prominencia acentual con picos de F0.

Finalmente la tercera sección trata la posibilidad de clasificar palabras finales de oración de acuerdo a sus acentos léxicos, utilizando como observaciones sus rasgos suprasegmentales.

Se pudo demostrar esta hipótesis mediante dos experimentos: obteniendo las curvas prototípicas de F0 correspondientes a cada tipo de acento de palabra y mostrando que eran diferentes, y mediante un análisis estadístico a nivel de atributos suprasegmentales sobre las últimas tres sílabas de cada oración.

Los experimentos de este capítulo permitieron caracterizar algunos comportamiento prosódicos presentes en oraciones declarativas para el Español de Argentina. Además abren posibles líneas de investigación tanto para comprender mejor cómo se manifiestan los fenómenos prosódicos en nuestro idioma, como para su aplicación específica dentro de sistemas de síntesis y reconocimiento del habla.

5 INFORMACIÓN ACENTUAL EN EL MODELADO ACÚSTICO

ÍNDICE

Este capítulo presenta una alternativa para la utilización de información acentual dentro de los modelos acústicos de un sistema de RAH.

En primer lugar se hace una introducción al tema y se presenta la estrategia propuesta.

Posteriormente se detallan algunos antecedentes vinculados al empleo de información de acentos léxicos en el proceso de RAH.

En la tercera sección se presentan los materiales y la metodología que se emplearon para llevar a cabo el estudio.

La cuarta sección muestra los resultados obtenidos, y finalmente el capítulo termina con una conclusión sobre la estrategia y resultados alcanzados.

5.1 INTRODUCCIÓN

Como se comentó en el Capítulo 1, es posible emplear información suprasegmental en todas las fases del proceso de RAH. Se la puede aprovechar por ejemplo durante el preprocesamiento a través de la segmentación de fragmentos del habla continua, o en el modelo acústico como un rasgo adicional a los coeficientes espectrales de tiempo corto, en el modelo de lenguaje, por ejemplo considerando modelos variantes en el tiempo y dependientes del contexto prosódico, durante el procedimiento de búsqueda del decodificador, determinando el tamaño del haz de decodificación, o en la fase de posprocesamiento ponderando hipótesis candidatas mediante evidencias suprasegmentales.

En este capítulo se propone y evalúa la utilización de información sobre acentos léxicos en los modelos acústicos, creando modelos separados para vocales acentuadas e inacentuadas lexicalmente. La evaluación se lleva a cabo comparando el desempeño del sistema propuesto con respecto a un sistema de referencia estándar.

Como se destacó en el Capítulo 3, el acento léxico se refiere al acento de palabra que está determinado por las reglas de la ortografía y es un rasgo abstracto y descontextualizado, que obedece a la fonología de la lengua. En el Español la mayoría de las palabras presentan un acento léxico que corresponde a la vocal de la sílaba acentuada, y en muchos casos ese acento léxico permite diferenciar palabras, por ejemplo en el par mínimo *papá-papa*. Al menos en el caso de pronunciación de palabras aisladas, las realizaciones acústicas de las vocales con acento léxico tienen correlatos físicos determinados por la mayor duración, mayor energía, mayor frecuencia glótica o fundamental y una estructura espectral mejor delineada [101, 18].

los acentos léxicos son reservorios potenciales del acento tonal

También se expuso previamente que en el Español los acentos tonales se ubican generalmente donde existe un acento léxico [98]. Es decir, los acentos léxicos son reservorios potenciales del acento tonal .

Sin embargo existen casos en los que no se respeta esa regla, y la sincronía entre acento léxico y tonal no es exacta. Es posible encontrar situaciones en que se producen defasajes entre el pico de la frecuencia fundamental y las fronteras de la sílaba acentuada, e incluso ocurrir que una sílaba con acento léxico reciba un acento tonal pero que la frecuencia fundamental no tenga un valor alto o aumentado (denominado H*). Además en las frases de habla continua es común que el contorno de entonación de la frase predomine sobre los acentos tonales, generando por ejemplo un descenso de la frecuencia fundamental hacia finales de una frase afirmativa. En esta situación, el acento léxico se manifiesta a través de un aumento de los parámetros restantes (duración y energía), y el acento tonal recibe la categoría de tono bajo (L*) [26].

La idea subyacente en esta propuesta es que el atributo suprasegmental designado como acentuación (ver el apartado dedicado a este tema en la Sección 3.2.6), condiciona la forma en que se articula el habla, y determina rasgos acústicos diferentes respecto a segmentos no acentuados.

Además ese atributo prosódico está correlacionado con el acento léxico, y de esa forma brinda una pista adicional que permite distinguir palabras. Por lo tanto si se representa de manera indistinta unidades acentuadas e inacentuadas se pierde esa información, lo que justificaría realizar una representación diferenciada para ambos casos.

Sin embargo como en el habla continua, acento prosódico y léxico pueden aparecer desfasados, alterados por el componente entonativo de frase, o incluso encontrarse palabras que si se pronunciaran de manera aislada resultarían acentuadas, al articularse en este continuo no manifiestan acentos prosódicos, la aproximación seguida en este estudio se debe pensar como una simplificación de lo observado en la realidad.

En otras palabras, durante este estudio se adopta la simplificación de considerar que para todas las palabras de contenido, la vocal sobre la

que recae el acento léxico presenta una manifestación física distintiva, que se representa modelando esta vocal como acentuada en contraposición a las vocales de sílabas átonas lexicalmente y a las vocales de palabras de función.

Este trabajo fue presentado en [41], mientras que el desarrollo del sistema de referencia se presentó en [180].

5.2 ANTECEDENTES

Al igual que para el caso más general del empleo de información suprasegmental en el proceso de reconocimiento, se pueden encontrar antecedentes respecto al empleo de información de acentos léxicos en las distintas fases del proceso de RAH: en la fase de preprocesado, durante el posprocesamiento de las hipótesis de reconocimiento, y aquellos que intentan introducir la información directamente en el proceso de búsqueda de los sistemas basados en HMM.

En [189] se estudian los parámetros acústicos que tienen mayor correlación con el acento léxico, y se los agrega como atributos adicionales al vector de características de un sistema de reconocimiento del habla telefónica. Los autores obtienen una reducción relativa de 5,3 % de la tasa de error a nivel palabra respecto a los sistemas de referencia.

En [3] se trata de explotar la relación entre acentos tonales y unidades léxicas para mejorar el desempeño de un sistema de RAH. La estrategia propuesta utiliza: 1) un modelo acústico prosódico basado en redes neuronales para estimar la presencia de acentos tonales a partir de rasgos acústicos, 2) un modelo probabilístico de secuencias de etiquetas de acentos tonales y 3) un modelo probabilístico de secuencias de etiquetas tonales dada una secuencia de palabras y acentos lexicales.

Los modelos se utilizan para calcular la lista de N-mejores hipótesis de reconocimiento. Empleando esta estrategia los autores reportan una reducción del error de reconocimiento a nivel de palabras de 1,3 % respecto a un sistema de referencia.

Para el Español, un aporte pionero en la utilización de acentos léxicos en el reconocimiento del habla fue presentado en [120]. En ese trabajo la información de los correlatos acústicos del acento lexical: frecuencia fundamental, energía, duración y espectro, se utilizaron en el modelo lingüístico del reconocedor. Específicamente se propuso emplear información sobre secuencias de acentos léxicos para modificar las probabilidades de transición entre palabras en el marco de un modelo de lenguaje variante en el tiempo. Se reportaron disminuciones relativas del error del 28,91 % respecto al sistema de referencia empleando como material de evaluación habla continua para el Español peninsular. En ese mismo trabajo se estudia la correlación entre tonicidad léxica y manifestaciones suprasegmentales para ese cuerpo de datos.

Como el estudio presentado en este capítulo evalúa la utilización de información acentual en los modelos acústicos, en cierta medida com-

plementa el realizado en [120] enfocado en la utilización de acentos léxicos en el modelado lingüístico.

El antecedente más próximo, desde el punto de vista metodológico a la propuesta presentada en este capítulo probablemente sea el que se describe en [182]. Allí se propone utilizar modelos acústicos diferentes para las vocales acentuadas de las inacentuadas lexicalmente en un sistema de reconocimiento del habla para el Holandés. Sin embargo los resultados reportados no mostraron mejoras de las tasas de reconocimiento.

A pesar de los resultados poco promisorios obtenido en ese caso, al efectuar este estudio se consideró que la diferencia en la manifestación del acento léxico entre el Holandés y el Español podría determinar resultados diferentes. En este sentido se parte de la hipótesis que el efecto del acento léxico puede ser más evidente en el Español, dado que es una lengua de ritmo silábico para el habla leída [171], donde existe mayor consistencia en la diferenciación entre vocales acentuadas e inacentuadas.

5.3 MATERIALES Y MÉTODOS

La metodología empleada para evaluar la propuesta presentada de este capítulo consistió en desarrollar por un lado un sistema de RAH estándar, que fue empleado como referencia, y por el otro sistemas entrenados con los modelos acústicos en que se diferencian vocales acentuadas e inacentuadas lexicalmente, para posteriormente evaluar el desempeño de los mismos.

5.3.1 Corpus de Datos Empleados

La base de datos empleada, forma parte del proyecto SALA I (SpeechDat Across Latin America) [125]. El subconjunto correspondiente al Español de Argentina [64] está constituido por cinco regiones distribuidas en todo el país. El estilo de habla corresponde a párrafos leídos, extraídos de diarios y libros de la Argentina, palabras aisladas de comando y control, secuencias de números, nombre de empresas o apellidos, algunas frases cortas espontáneas y oraciones elaborados por lingüistas.

Las grabaciones se realizaron a través de la red de telefonía fija por medio de una computadora equipada con una placa de adquisición AVM-ISDN-A1 y una interfaz de acceso básico a ISDN (BRI). La frecuencia de muestreo empleada fue de 8 kHz a 16 bits por muestra.

Para este trabajo se utilizaron frases de habla continua de la base SALA Argentina, región SUR. Esta región comprende las provincias de Buenos Aires, Santa Fe, Entre Ríos, La Pampa, Neuquén, Río Negro, Chubut, Santa Cruz y Tierra del Fuego. La región SUR es la más populosa de Argentina con un número aproximado de 21 millones de habitantes (corresponde al 65 % del total del país) y forma parte de una de las divisiones dialectales propuestas en [186].

Como criterio de selección se eligieron todas las frases de habla continua de esta región que no presentaban anotaciones de ruidos externos, ruidos del locutor, o errores de pronunciación.

De esta forma el corpus quedó delimitado a 1301 frases, con un total de 9948 palabras, correspondientes a un vocabulario de 2722 palabras distintas, emitidas por 138 hablantes (48 hombres y 90 mujeres) correspondientes a 99 minutos de grabación.

5.3.2 Sistema de RAH de Referencia

Se desarrollaron tres sistemas de reconocimiento del habla basados en HMM, uno con modelos acústicos conformados por monofonos y los siguientes por distintas variantes de trifonos.

A continuación se detallarán las características de estos reconocedores.

Preprocesamiento

Esta etapa es común para los tres sistemas de reconocimiento de referencia como del sistema propuesto.

La parametrización de la señal acústica se realizó empleando frecuencia de muestreo de 8 KHz, 16 bits de resolución y sustracción de la media temporal, de manera de eliminar cualquier nivel de continua proveniente de la etapa de adquisición.

Se utilizaron ventanas de análisis del tipo Hamming de 25 ms de duración y 10 ms de avance, filtro de preénfasis de primer orden (con coeficiente de valor 0,97), y normalización de la energía a nivel de frase.

Se codificó cada ventana de la señal empleando 12 coeficientes cepstrales en escala Mel a los cuales se les adicionó los coeficientes delta y aceleración (derivadas temporales de primer y segundo orden), conformando un total de 39 parámetros para cada vector de observaciones.

Diccionario de Pronunciaciones y Modelo de Lenguaje

El diccionario de pronunciaciones de los sistemas de referencia se construyó empleando el alfabeto fonético para fines tecnológicos, denominado SAMPA (Speech Assesment Methods: Phonetic Alphabet), adaptado para el Español de Argentina [61], que utiliza un total de 30 unidades fonéticas. Para la generación de este diccionario se construyó un transcriptor de grafema a fonema basado en reglas, que dada una palabra devuelve su contraparte fonética.

Como modelo de lenguaje se emplearon bigramas, estimados en base a las transcripciones de las frases de la base de datos. En la Tabla 11 se presentan las características generales de la gramática generada.

Modelos Acústicos

Las unidades empleadas en los modelos acústicos de los sistemas de referencia fueron: monofonos, trifonos dependientes del contexto inte-

Atributo	Valor
Vocabulario (palabras)	2722
Número de nodos	2723
Entropía	5,4
Perplejidad	42,5
Longitud de frase promedio (palabras)	9,2
Longitud de frase mínima (palabras)	2
Longitud de frase máxima (palabras)	40

Cuadro 11: Características del modelo de lenguaje utilizado.

rior de las palabras (**TdCIP**) y trifonos dependientes del contexto entre palabras (**TdCEP**).

El conjunto de monofonos estándar quedó conformado por las 30 unidades fonéticas correspondientes al alfabeto SAMPA para Argentina, a las cuales se les agregó un modelo de "silencio" y otro "pausa corta" completando un total de 32 unidades.

Los modelos empleados en el caso de los TdCIP y TdCEP fueron generados a partir de la expansión de los modelos de monofonos estándar en trifonos, considerando el contexto de cada fonema. Posteriormente se los agrupó en clases acústicas, mediante agrupamientos basados en reglas fonéticas. Este agrupamiento permitió reducir la cantidad de modelos a emplear dada la limitada cantidad de datos de entrenamiento para estimar cada modelo. El número de unidades definitivas, luego de la fase de agrupamiento fue de 849 para TdCPI, y de 1314 para TdCEP.

El número promedio de mezclas de Gausianas tanto para monofonos como para trifonos fue de 144, a saber de 256, 128, 128 y 64 para los subconjuntos 1, 2, 3 y 4 de los vectores de características respectivamente.

5.3.3 Sistema de RAH Propuesto

Como se mencionó, el sistema de RAH propuesto realiza el mismo preprocesamiento sobre la señal de habla que los sistemas de referencia. A continuación se describirá los módulos restantes de este sistema.

Diccionario de Pronunciaciones y Modelo de Lenguaje

Para el caso del sistema propuesto se modificó el conversor de grafemas a fonemas desarrollado para los sistemas de referencia con el fin de distinguir las vocales acentuadas de las inacentuadas lexicalmente. Para ello se utilizó un conjunto de reglas ortográficas presentadas en [83].

En experimentos exploratorios previos no se logró una mejoría en el reconocimiento de palabras empleando las monosílabas acentuadas. Por ello las vocales correspondientes a palabras monosilábicas fueron consideradas no-acentuadas. Esto puede justificarse por el hecho que

aproximadamente el 90 % de las palabras átonas son monosílábicas [141].

Al igual que en el caso de los sistemas de referencia, se emplearon bigramas como modelo de lenguaje.

Modelos Acústicos

La clase de monofonos acentuados fue generada sumando al conjunto de monofonos estándar los modelos correspondientes a las cinco vocales acentuadas, quedando constituido por 37 unidades.

5.3.4 Entrenamiento y Evaluación de los Reconocedores

A continuación se presenta la metodología empleada para la construcción de los sistemas de reconocimiento, es decir para el entrenamiento de los modelos los acústicos, así como el proceso seguido para la evaluación de estos sistemas.

El entrenamiento de los modelos acústicos siguió la metodología propuesta por [194], que consistente en la siguiente secuencia de pasos:

Etapa de Entrenamiento

1. Creación de un HMM simple de 3 estados de izquierda a derecha para cada uno de los fonemas, exceptuando la pausa corta, que es asociada al estado central del modelo de silencio.

2. Generación de nuevos modelos a partir de los modelos ya entrenados. Los nuevos modelos comparten el mismo conjunto de funciones de densidad de probabilidad, variando únicamente los pesos de ponderación aplicados a cada una de ellos (HMM semicontinuos).

3. Para los monofonos y monofonos acentuados el re-entrenamiento de los modelos semi-continuos hasta obtener los HMM definitivos que se emplean en la etapa de reconocimiento.

4. Para el caso de TdCIP y TdCEP se realizó la expansión automática de los modelos de monofonos estándar a trifonos. En el caso de los TdCIP la expansión abarca los difonos y trifonos presentes dentro de cada palabra, mientras que para el caso de los TdCEP se consideran también trifonos entre palabras.

5. Re-entrenamiento de los modelos para ambos tipos de trifonos.

6. Agrupamiento de los modelos en clases acústicas similares.

7. Re-entrenamiento de los grupos de modelos generados, empleando enlazado de parámetros.

8. Re-entrenamiento final hasta lograr los HMM definitivos a ser empleados en la etapa de reconocimiento para ambos tipos de trifonos.

Etapa de Reconocimiento

Teniendo en cuenta el modelo acústico, el modelo de lenguaje y la secuencia de observaciones acústicas correspondientes a la frase a reconocer, se utiliza el algoritmo de Viterbi para buscar la secuencia de transiciones de estados de máxima verosimilitud. Esta secuencia es la que mejor explica la señal acústica recibida, considerando los modelos acústicos entrenados y el modelo de lenguaje.

El reconocedor posee diversos parámetros configurables, para los cuales se emplearon los siguientes valores: ancho del haz de decodificación 120, factor de preponderancia del modelo de lenguaje sobre el acústico 5, factor de penalización de palabras insertadas 0.

ancho del haz de decodificación

El **"ancho del haz de decodificación"** restringe el crecimiento de la red de reconocimiento a aquellos HMM cuyas probabilidades de verosimilitud se encuentran dentro de un ancho de haz determinado en relación al modelo más probable. De esta manera disminuyendo el ancho de haz se procesan menos modelos y se reduce el tiempo de decodificación, aunque puede reducirse también el porcentaje de reconocimiento debido a que se ignoran posibles soluciones.

"factor de lenguaje"

Por su parte el **"factor de lenguaje"** post-multiplica la verosimilitud de la red de palabras de forma de incrementar la importancia del modelo de lenguaje con respecto al modelo acústico. En el caso de un factor de lenguaje nulo sólo se consideraría el modelo acústico en el reconocimiento.

Finalmente el **"factor de penalización de palabras insertadas" factor de penalización de palabras insertadas** permite controlar la probabilidad de inserción de palabras durante el reconocimiento. Al aumentar este factor se hacen más probables hipótesis de secuencias con mayor número de palabras, aumentando también el riesgo de inserciones de palabras erróneas.

5.4 RESULTADOS

Como en estos estudios se evaluó la velocidad de respuesta de los sistemas propuestos y de referencia es pertinente comentar las características del equipo empleado para correr dichas pruebas, dado que los recursos hardware disponibles condicionan los resultados obtenidos. Dicho equipo consistió en una computadora con procesador AMD Athlon XP-M 2200+ con 512 MB de memoria RAM.

Para la comparación del desempeño de los reconocedores empleando las distintas unidades acústicas se utilizó la metodología de validación cruzada, generando 10 particiones distintas del conjunto de datos disponibles. Para cada partición se separaron 20 % de los casos para evaluación.

En la Tabla 12 se presenta el desempeño de los reconocedores evaluados en términos de sus tasas de reconocimiento de palabras (R), y precisiones de reconocimiento (P), según se definió en las ecuaciones 2.86 y 2.87 de la Sección 2.3.5, para cada partición del proceso de validación cruzada.

También se puede encontrar los valores de R y P promedios y sus desvíos estándar para cada sistema de reconocimiento correspondiente a los diferentes modelos acústicos evaluados.

Partición	Mono.		MonoAc.		TdCIP		TdCEP	
	R	P	R	P	R	P	R	P
1	85,5	83,5	85,1	82,9	84,8	83,1	84,3	80,9
2	88,8	87,3	88,0	86,1	86,5	89,9	89,4	86,5
3	88,4	87,0	88,5	86,8	88,9	87,2	90,4	85,5
4	92,5	91,7	92,9	92,4	90,5	88,3	93,4	89,8
5	86,6	86,2	85,5	85,1	85,2	84,4	87,6	85,6
6	91,0	90,0	93,0	91,5	91,9	90,4	93,3	89,8
7	83,8	82,6	86,0	84,7	85,0	82,1	86,5	80,4
8	81,1	79,6	81,4	80,3	80,6	79,1	82,9	79,0
9	78,1	76,9	80,1	79,2	76,0	73,9	81,2	77,9
10	85,7	84,1	88,1	87,0	83,7	81,7	86,4	82,7
Promedio	86,1	84,9	86,8	85,6	85,3	83,5	87,6	83,8
Desvío Estd.	4,4	4,5	4,2	4,2	4,7	4,8	4,1	4,3

Cuadro 12: Resultados de reconocimiento en términos de porcentajes de reconocimiento de palabras (R), y de precisión (P) empleando diferentes unidades acústicas y validación cruzada de 10 conjuntos. **Mono.**: monofonos convencionales; **MonoAc.**: modelo propuesto de monofonos acentuados; **TdCIP**: trifonos dependientes del contexto interior de las palabras; **TdCEP**: trifonos dependientes del contexto entre palabras.

En la Tabla 12 se puede observar que en término de precisión de reconocimiento, los mejores resultados se obtienen para monofonos acentuados. Aún cuando los monofonos estándar tienen la ventaja de disponer de más muestras para estimar cada modelo durante el entrenamiento, el desempeño fue menor, confirmándose la hipótesis de existencia de diferencias entre los atributos físicos de vocales acentuadas e inacentuadas, que pueden aprovecharse para mejorar el reconocimiento del habla.

Comparando los resultados del modelo de **MonoAc** respecto a los de **TdCEP** que son los más complejos y que mejor modelan detalles contextuales, se puede ver que el modelo propuesto aumenta la precisión (o reduce el WER de acuerdo a la ecuación 2.88) en un 1,8 % en términos absolutos, o en un 11,11 % en términos relativos.

Realizando una prueba T para muestras relacionadas, se encontró que tal diferencia de desempeño es significativa. Como parámetros en la prueba T para muestras relacionadas se utilizó un intervalo de confianza de 95 %, y se obtuvo un valor t de 3,461, con nivel crítico bilateral de 0,007.

Los resultados obtenidos para trifonos son inferiores a los obtenidos para monofonos. En este caso el número de muestras disponibles de entrenamiento es aún menor. Es de esperar que al aumentar el número de ejemplos de entrenamiento, mejore el desempeño empleando estas

unidades, ya que existe una relación de compromiso entre la calidad de representación del contexto y el número de ejemplos disponibles para estimar adecuadamente los modelos para cada uno de esos contextos.

La Figura 53 permite comparar las ventajas y desventajas de cada conjunto de unidades acústicas, de acuerdo a la metodología propuesta en [146].

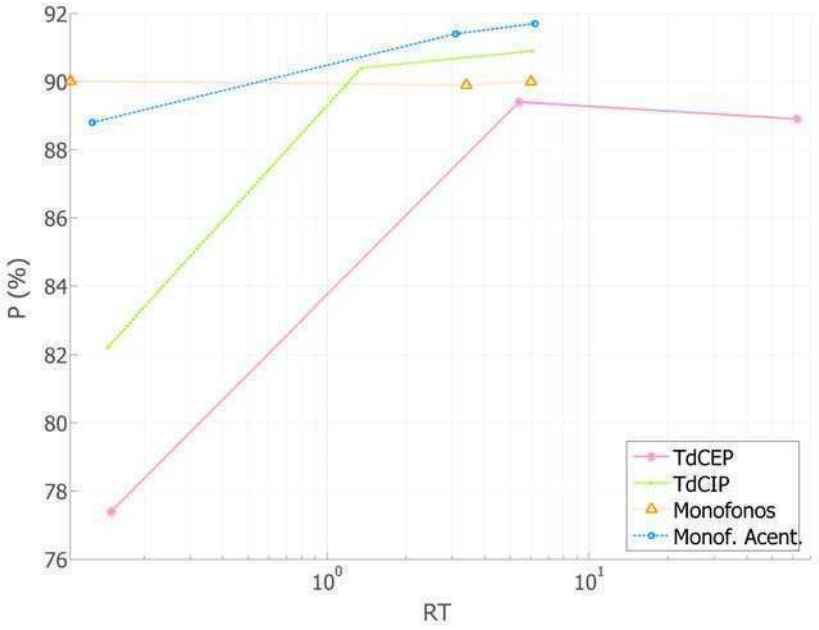

Figura 53: Tasas de Precisión (**P**) en función del factor de tiempo real (**RT**) para **Monofonos**: monofonos convencionales; **Monof. Acent.**: modelo propuesto de monofonos acentuados; **TdCIP**: trifonos dependientes del contexto interior de las palabras; **TdCEP**: trifonos dependientes del contexto entre palabras

En esta figura se pueden ver los distintos valores del factor de tiempo real y las correspondientes precisiones para cada modelo acústico. Estos datos se obtuvieron realizando la evaluación de los distintos sistemas de reconocimiento con diferentes valores del parámetro que controla el haz de decodificación dentro del algoritmo de Viterbi. Cuando se brinda al reconocedor una mayor exploración, la precisión de reconocimiento mejora pero también aumenta el tiempo de procesamiento requerido. Se puede ver que a partir de un factor de tiempo real de 0,8 aproximadamente, el mejor desempeño se logra con las unidades propuestas. Además los trifonos requieren mayor tiempo para mostrar un desempeño equivalente.

Considerando una precisión de al menos 89 %, en la Tabla 13 se presenta la información vinculada al factor de tiempo real, y al uso de memoria para cada modelo.

En la Tabla 13 se puede observar que los requerimientos de memoria son consistentes con la cantidad de unidades de los modelos propuestos. A pesar de requerir mayor espacio en memoria, los modelos de

Modelo	TR	Memoria (KB)
Monofonos	0,10	142
Monofonos acentuados	0,13	170
TdCIP	0,95	940
TdCEP	3,89	986

Cuadro 13: Resultados comparativos de precisión de reconocimiento, tasa de tiempo real (**TR**) y memoria de almacenamiento de los HMM. **TdCIP**: trifonos dependientes del contexto interior de las palabras; **TdCEP**: trifonos dependientes del contexto entre palabras.

monofonos acentuados presentan un factor de tiempo real cercano al de los monofonos estándar. Los modelos de trifonos TdCEP mostraron la mayor relación de tiempo real en concordancia con el mayor número de unidades.

5.5 CONCLUSIONES

En este trabajo se construyeron modelos acústicos basados en la información de acento lexical. Se evaluó el desempeño del modelo propuesto con respecto al modelado acústico estándar empleando como material de evaluación habla telefónica leída para el reconocimiento automático del habla del Español de Argentina.

Para el conjunto de datos disponibles el empleo de modelos semicontinuos independientes del contexto (monofonos y monofonos acentuados) permitió obtener porcentajes de reconocimiento similares a los obtenidos empleando modelos dependientes del contexto (trifonos).

El empleo de los monofonos acentuados permitió reducir de manera significativa el WER en un 1,8 % en términos absolutos con respecto a los trifonos dependientes del contexto entre palabras.

Comparando los sistemas de reconocimiento evaluados en términos de factor de tiempo real, se observó que para valores de este factor superores a 0,8 el sistema propuesto presenta la mejor relación precisión en función del tiempo de procesamiento.

Si se requiere implementar un sistema de reconocimiento en tiempo real es necesario imponer restricciones al haz de decodificación en el algoritmo de Viterbi para lograr tasas de tiempo real menores a un RT de 1. Si bien esta restricción provoca una disminución en las tasas de reconocimiento, la misma no es muy significativa.

Como en este trabajo se distinguen las vocales acentuadas de las inacentuadas sólo a partir del texto, deja abierta la posibilidad de experimentar realizando esa distinción utilizando información acústica.

6 EMPLEO DE INFORMACIÓN ENTONATIVA EN EL RAH

ÍNDICE

Este capítulo investiga la utilización de información suprasegmental en la fase de posprocesamiento del RAH.

Específicamente se propone utilizar un índice de compatibilidad entonativa para reordenar las mejores hipótesis de reconocimiento. Tal índice se obtiene contrastando el valor de F0 que presenta el segmento de habla a reconocer, con el valor de F0 de las hipótesis de reconocimiento, obtenidos a través de un proceso de predicción.

En la primera sección se presentan las motivaciones de la propuesta, y se describen de forma sintética las hipótesis de trabajo.

La segunda sección brinda algunos antecedentes vinculados con la utilización de información entonativa en el RAH.

La siguiente sección describe el sistema de reconocimiento propuesto y expone en detalle cada componente del mismo.

La cuarta sección describe los materiales y métodos empleados durante el desarrollo de los experimentos.

La siguiente sección contiene los resultados obtenidos al aplicar el sistema propuesto en tareas de reconocimiento de habla continua, y su relación con el desempeño de sistemas de RAH convencionales.

Finalmente se brindan las conclusiones del capítulo.

6.1 INTRODUCCIÓN

En la serie de experimentos presentados en este Capítulo se evaluó la posibilidad de utilizar información entonativa para mejorar el ordena-

miento de las oraciones candidatas generadas por un sistema estándar de RAH.

El trabajo efectuado surge a partir de dos hipótesis. La primera se puede definir de la siguiente manera:

"La realización acústica de las distintas hipótesis de reconocimiento deberían manifestar diferencias a nivel suprasegmental. Comparando los posibles rasgos suprasegmentales de las hipótesis, con respecto a los observados en la señal de habla a reconocer, sería posible establecer una medida de ***compatibilidad suprasegmental****. Finalmente esa medida sería útil para reordenar la lista de hipótesis y consecuentemente disminuir los errores de reconocimiento".*

Esta primera hipótesis asume que si se analiza un conjunto de frases pertenecientes a un mismo locutor, pronunciadas a una velocidad similar, entonces la estructura sintáctica y el contenido léxico de las frases determina que haya algunos patrones prosódicos más susceptibles de ocurrir que otros.

De esta forma sería factible explotar la correlación entre la estructura de las oraciones y la entonación como una evidencia adicional durante la desambiguación de hipótesis en el RAH.

En la sección 1.2.5 se presentó la dicotomía entre procesamiento top-down y bottom-up durante el reconocimiento humano del habla. Nuestra propuesta tiene un enfoque bottom-up, dado que hace uso de la información suprasegmental una vez que se obtienen las hipótesis de reconocimiento.

Los sistemas de RAH basados en HMM pueden considerarse como un híbrido: si bien se parte del análisis acústico-fonético por un lado, procesamiento bottom-up, durante el proceso de búsqueda mediante Viterbi, se optimiza a nivel de la frase completa y se tiene en cuenta información lingüística de alto nivel representada por el modelo de lenguaje.

En la estrategia presentada en este capítulo, en caso de emplear modelo de lenguaje, el sistema presenta la dualidad bottom-up top-down y luego se agrega información suprasegmental en una etapa final de procesamiento.

Por otra parte, en caso de no utilizar modelo de lenguaje, el reconocimiento se lleva a cabo haciendo uso del modelo acústico y el diccionario de pronunciaciones, en una forma que se podría considerar bottom-up, para finalmente utilizar la información entonativa.

La segunda hipótesis de trabajo se puede expresar de la siguiente forma: *"En la literatura se pueden encontrar reportes en los que se muestra que la información suprasegmental es más robusta que la información espectral de tiempo corto frente a degradaciones en la señal de habla [21]. De esta forma, si se demuestra la primera hipótesis, el beneficio de emplear información entonativa en el reconocimiento, debería ser mayor para señales de habla con menores niveles de relación señal-ruido".*

El proceso para el reordenamiento de hipótesis que se presenta en este Capítulo consiste en la combinación de la verosimilitud original de cada hipótesis, con el índice de compatibilidad entonativa propuesto, buscando que el nuevo ranking de hipótesis muestre menores errores de reconocimiento.

Los resultados preliminares de este Capítulo fueron presentados en [40].

6.2 ANTECEDENTES

Si bien se pueden encontrar muchas propuestas para el empleo de información suprasegmental en la fase de reestimación de hipótesis, son escasas las que hacen uso de la información entonativa para esta tarea.

En [184] se evalúa tres alternativas diferentes para el empleo de información prosódica en la reestimación de hipótesis de reconocimiento. En primer lugar se propone utilizar un modelo de duraciones de palabras dado por un vector de duraciones de los fonos que la conforman.

En segundo lugar se estudia la utilización de un modelo predictivo de pausas a partir del contexto entre palabras y emplear esa información considerando las pausas detectadas en la señal de entrada.

Finalmente se experimenta con un modelo predictivo de eventos ocultos (como límites segmentales y disfluencias) a partir de parámetros suprasegmentales. Se reporta una reducción en las tasas de error absoluto por palabra que van del 0.2 % al 0.9 %.

En [185] se utiliza información suprasegmental para detectar límites de palabras en lenguajes de acento fijo. Las hipótesis de inicio de palabras se integran en la reestimación de las hipótesis de reconocimiento.

Los autores proponen un sistema basado en HMM para la segmentación prosódica del habla en frases fonológicas para lenguajes de acento fijo. A partir de los límites prosódicos se derivan los límites de palabra, que se integran en el proceso de reordenamiento de hipótesis de reconocimiento de un sistema de RAH.

Usando esta estrategia los autores reportan una mejora relativa de un 3.82 % para tareas de reconocimiento sobre un corpus médico en idioma Húngaro.

En [154] se propone integrar en el proceso de reestimaciones un modelo de duraciones por palabra empleando una función de densidad de probabilidades. Experimentos con el modelo propuesto para el Inglés y Español mostraron reducciones pequeñas pero consistentes en las tasas de error. Las evaluaciones realizadas mostraron una reducción del WER de 0.2 % para el Inglés y de 0.1 % para el Español.

A partir de la observación que para un mismo locutor, una misma frase y velocidad de habla, cierto tipo de patrón entonativo es más frecuente de encontrar que otros, en este capítulo se estudia la posibilidad de utilizar tales patrones entonativos para mejorar el desempeño de un sistema de RAH estándar.

6.3 SISTEMA PROPUESTO

En la figura 54 se presenta el esquema del modelo propuesto.

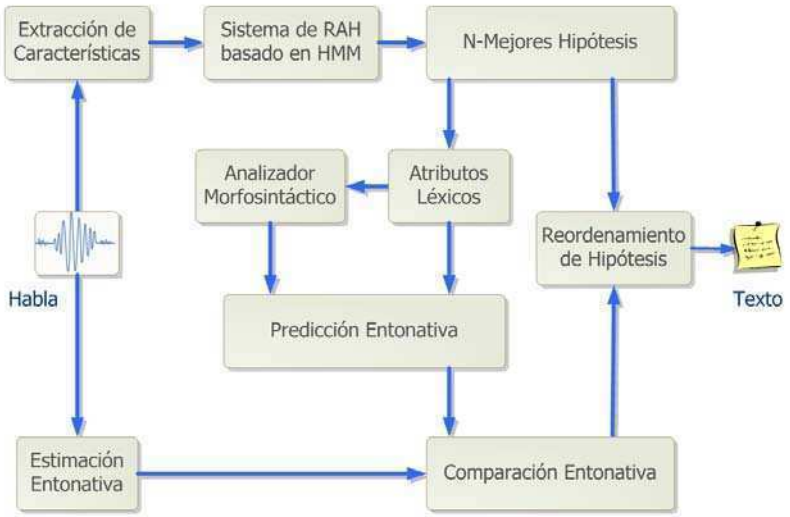

Figura 54: Modelo propuesto para el empleo de información entonativa en la reestimación de las mejores hipótesis de reconocimiento.

La señal de habla a reconocer es recibida por el módulo de extracción de características el cual deriva un conjunto de parámetros acústicos a un sistema de convencional de RAH basdo en HMM. Ese sistema determina una lista de N-mejores hipótesis de reconocimiento. Una vez que el conjunto de oraciones candidatas está disponible, se lleva a cabo un análisis de atributos léxicos y gramáticos: mediante un procedimiento de alineamiento forzado se obtiene la segmentación de las oraciones tanto a nivel fonético como de palabras, y, a través de un analizador morfo-sintáctico [7], se extraen las etiquetas léxicas de cada palabra.

El módulo de extracción de atributos léxicos proporciona la información necesaria para que el módulo de predicción entonativa genere un contorno estimado de F0 por hipótesis de reconocimiento.

Por otro lado, en el módulo de estimación de entonación, también se analiza la señal de habla a reconocer para extraer el contorno de F0 de referencia, empleando el algoritmo RAPT (ver Sección 3.5.1).

Finalmente, el módulo de comparación entonativa contrasta la curva de F0 de referencia contra la curva de F0 predicha para cada hipótesis, y genera un índice de compatibilidad entonativa.

Existen varias alternativas para medir el índice de compatibilidad entre pares de curvas. Para el caso que debe resolver este módulo, como tanto la curva de referencia como las curvas predichas poseen la misma duración, es posible aplicar una base de análisis punto a punto, no siendo necesario recurrir a técnicas de deformación temporal (warping), u otro tipo de adaptación durante la comparación. Se evaluó tres medidas alternativas para cuantificar las diferencias entre los contornos: error cuadrático medio, error global en Hz, y error global en escala ERB.

A continuación se describe en detalle los componentes del sistema propuesto. En primer lugar se presenta el módulo de predicción de contornos de F0, y el módulo para la extracción de parámetros empleados en la predicción a partir de cada hipótesis de reconocimiento.

En segundo lugar se presenta la metodología empleada para comparación de contornos entonativos.

6.3.1 Predicción Entonativa

Este módulo fue desarrollado en [174], aquí se presentarán los detalles más importantes. El módulo es similar a los empleados durante la generación de la información entonativa en los sistemas convencionales para la síntesis concatenativa del habla. Está compuesto por dos submódulos: el predictor de comandos de Fujisaki, y el sintetizador de F0 propiamente dicho.

El empleo de una representación intermedia, en este caso empleando el modelo de entonativo de Fujisaki, permite simplificar el proceso de predicción de la curva de F0. Empleando esa representación paramétrica, el problema de predicción de F0 consiste en hallar los valores de un grupo de parámetros en vez de la forma de onda completa.

En [49] se evaluó el modelo de Fujisaki para el Español de Argentina y se demostró que es una representación adecuada para modelar los contornos entonativos encontrados en nuestra variante del Español. Por otra parte, en [62] se puede encontrar una comparación entre esta representación y la aproximación fonológica ToBI para la misma lengua.

Debido a la compleja formulación del modelo de Fujisaki, el análisis y la extracción automática de sus parámetros es laboriosa y necesita revisiones manuales. Uno de los métodos estándar para determinar de manera automática esos parámetros es el basado en análisis por síntesis [122]. Este método realiza una búsqueda completa dentro de un rango razonable para cada parámetro, y su proceso de naturaleza iterativo continua hasta que la mejor combinación de parámetros ajuste el contorno especificado.

Para la predicción de los comandos de Fujisaki se emplearon árboles de clasificación y regresión (CARTS) siguiendo el algoritmo propuesto en [19].

Durante la fase de entrenamiento se brindó como entradas a los árboles el siguiente conjunto de datos:

- Ubicación y longitud de la frase entonativa.
- Identidad, localización y longitud de la vocal acentuada.
- Identidad de los fonemas contextuales, dos previos y dos posteriores al fonema correspondiente a la vocal acentuada.
- Etiqueta de parte del habla (POS), en un contexto de dos palabras previas y dos palabras siguientes.
- Distancia a la vocal acentuada anterior y siguiente que tengan un comando de acento asociado.

- Valor de los parámetros del comando de acento o frase previos.

Para los atributos vinculados con la ubicación, distancia y longitud, se evaluaron distintas unidades: duración en segundos, número de fonemas, número de sílabas, número de palabras y de frases entonativas. Todas estas medidas se evaluaron tanto en forma absoluta como relativa.

El rol de este módulo es predecir los comandos de Fujisaki descriptos en la Sección 3.4.2:

- Af: amplitud del comando de frase.
- Tor: ubicación del comando de frase relativa al comienzo de la frase entonativa.
- Aa: amplitud del comando de acento.
- T1r: localización del comando de acenteo relativa a la vocal acentuada de la palabra de contenido, excepto por la última palabra de contenido que se asume que no posee comando de acento asociado.
- T2-T1: duración del comando de acento.

Considerando el alto número de grados de libertad que existen para combinar la cantidad y los valores de los parámetros de Fujisaki tal que permitan ajustar una curva de F0 dada, durante la construcción de los árboles se emplearon las siguientes simplificaciones: se admite solamente un comando de frase por oración, y un comando de acento por palabra de contenido excepto por la última palabra de contenido de la oración que se considera no estar asociada con ningún comando de acento.

El módulo de predicción entonativa contiene un árbol de predicción para cada parámetro de Fujisaki. Los experimentos de construcción de tales predictores se llevaron a cabo sobre cinco particiones del conjunto de datos disponibles, y los resultados se promediaron. Para cada conjunto de datos se llevó a cabo una serie de experimentos con el fin de encontrar los parámetros de los árboles que minimicen los errores de clasificación.

Para evaluar el desempeño de cada predictor, se utilizó validación cruzada dividiendo aleatoriamente el 80 % de los datos para entrenamiento (usados para estimar los CARTs), y el 20 % restantes para evaluación.

El mejor balance entre precisión de predicción y generalización se obtuvo con árboles de profundidad 3. La tabla 14 muestra los porcentajes promedio de la raíz cuadrada de errores cuadráticos medios (RMSE), durante la estimación de los parámetros del modelo de Fujisaki sobre todas las particiones y empleando la profundidad óptima.

Los errores RMSE son relativos a los valores de los parámetros estimados empleando un proceso de búsqueda basado en algoritmos genéticos, como se describe en [176].

Parámetro	RMSE (%)
Af	8
Aa	18
Tor	31
T1r	25
T2-T1	29

Cuadro 14: Porcentajes de error RMSE promedio en la predicción de los parámetros del modelo de Fujisaki, asumiendo como estándar de referencia los valores determinados a partir de una búsqueda exhaustiva empleando algoritmos genéticos.

Durante la construcción de los CARTs, el método selecciona automáticamente las entradas que a brindan un mayor nivel de información. A continuación se detallan los atributos que finalmente se utilizan para la predicción de cada parámetro:

- Af: identidad de los fonemas anterior y siguiente, distancia desde el centro de la vocal acentuada de la primera palabra de contenido al final de la frase en segundos, duración del comando de acento previo en segundos.
- Tor: localización de la vocal acentuada a la primer palabra de contenido en segundos, longitud de frases en segundos, duración del comando de acento previo en segundos, amplitud del comando de frase previo.
- Aa: identidad de los fonemas anterior y siguiente, modo de articulación del siguiente fonema, etiqueta de parte de habla (POS) para la palabra de contenido actual, las dos palabras de contenido previas y las dos palabras de contenido siguientes, número de frases entonativas en la oración y longitud de la frase en segundos.
- T1r: identidad de los fonemas anterior y siguiente, distancia desde la vocal acentuada de la palabra de contenido actual hasta el final de la frase en segundos.
- T2-T1: identitdad del fonema siguiente, longitud de la frase en segundos, distancia desde la vocal acentuada de la palabra de contenido actual al comienzo de la frase en segundos, distancia de la vocal acentuada de la actual palabra de contenido al final de la frase en segundos, duración de los comandos de acento previos en segundos.

Vale la pena mencionar que para los parámetros de comando de frase, cada uno de los atributos detallados son relativos a la localización de la vocal acentuada de la primera palabra de contenido. Mientras que para los comandos de acento, las medidas de los atributos se calculan sucesivamente con respecto a la localización de la vocal acentuada de la palabra de contenido actual. Como se mencionó anteriormente,

se asocia un comando de acento por cada palabra de contenido, excepto la última.

Una información que resulta interesante conocer para usar el módulo de predicción entonativa dentro de la estrategia de reordenamiento es cuán bien se ajusta el F0 predicho al observado. En el cuadro 15, se presentan los resultados de evaluaciones objetivas, utilizando distintas medidas de errores, entre el valor de F0 real y el predicho usando el procedimiento descripto.

RMSE (Hz)	RMSE (ST)	RMSE (Ln)	RMSE (ERB)	MSE (Ln)
64	4,7	0,27	1,12	0,078

Cuadro 15: Cuantificación en distintas escalas de los errores promedio en la predicción de F0 para todos los conjuntos de validación cruzada, empleando el modelo entonativo propuesto en [176]. Hz: escala lineal medida en Hertzios; ST: semitonos; Ln: escala logarítmica; ERB: escala perceptual.

Donde el valor RMSE se mide entre el F0 original y el predicho considerando solamente regiones sonoras, sin interpolar sobre segmentos sordos. Los valores MSE son calculados con F0 en el dominio logarítmico para hacerlo comparable con las medidas presentadas en [149]. Estas diferentes escalas se utilizan más adelante para comparar el ajuste de las hipótesis de reconocimiento.

Una vez que se predicen los parámetros de Fujisaki, el sintetizador de F0 utiliza las ecuaciones 3.6, 3.6, y 3.7 para generar los contornos de F0.

6.3.2 Comparación Entonativa

Como se mencionó, el módulo de comparación de contornos entonativos es el responsable de determinar el grado de similitud o distancia entre la curva de F0 real y las predichas. Se utilizaron dos estrategias para realizar las comparaciones entre las curvas de F0.

En primer lugar se efectuó la comparación de perfiles entonativos considerando frases completas. En este caso, el mecanismo de funcionamiento del módulo de predicción entonativa hace que todas las curvas predichas tengan igual duración y resulten alineadas con la curva de F0 calculada a partir de la frase a reconocer.

Por lo tanto es posible realizar la comparación entre la curva original y las correspondientes a las hipótesis midiendo sus distancias punto a punto.

La segunda estrategia para contrastar las curvas de entonación surgió una vez implementado y evaluado el sistema siguiendo la estrategia de comparación a nivel de frases completas.

Considere el ejemplo de la figura 55, donde se muestran las 5 mejores hipótesis de reconocimiento correspondientes a la frase *"el diurético le causó pérdida de peso"*.

Hipótesis h1

48	67	el	−1237
67	136	directivo	−4464
136	146	de	−555
146	207	paso	−3664
207	223	al	−896
223	303	equivalentes	−4751

Hipótesis h2

48	67	el	−1237
67	136	directivo	−4464
136	146	de	−555
146	207	causO	−3659
207	223	al	−902
223	303	equivalentes	−4751

Hipótesis h3

48	67	el	−1237
67	136	directivo	−4464
136	146	de	−555
146	207	regreso	−3672
207	223	al	−893
223	303	equivalentes	−4751

Hipótesis h4

48	67	el	−1237
67	136	directivo	−4464
136	207	descanso	−4213
207	223	al	−908
223	303	equivalentes	−4751

Hipótesis h5

48	67	el	−1237
67	136	diurEtico	−4468
136	207	descanso	−4209
207	223	al	−908
223	303	equivalentes	−4751

Figura 55: Mejores 5 hipótesis de reconocimiento empleando el sistema de RAH estándar para la frase de entrada: *"el diurético le causó pérdida de peso"*. Para cada hipótesis se muestra el tiempo de inicio y fin de la palabra reconocida, su identidad y el valor de verosimilitud. Las unidades temporales están expresadas en múltiplos de 10 ms.

De manera similar a lo que se puede observar en el ejemplo presentado, en muchas instancias de reconocimiento se detectaron casos en los que ninguna de las hipótesis obtenidas era completamente correcta, aunque se podían encontrar eventualmente algunos fragmentos correctos en el interior de las hipótesis.

A partir de esa observación se decidió proponer una estrategia de comparación entonativa que intentase *rescatar* esas porciones correctas para distintos fragmentos de la frase completa, y componer una hipótesis alternativa mediante la concatenación de las mejores hipótesis para cada uno de los segmentos de una frase.

Para llevar esta idea a la práctica se propuso realizar la estimación de compatibilidad entonativa a nivel de segmentos de frases, empleando para la determinación de esos segmentos los instantes temporales en que las hipótesis de reconocimiento muestran similitudes y discrepancias, aplicando el siguiente procedimiento:

- Considerar como tiempo inicial del primer segmento, el menor tiempo de todas las hipótesis.

- Asumir como tiempo final del último segmento, el instante correspondiente al final de palabra de mayor valor temporal entre todas las hipótesis.
- Adoptar como puntos intermedios de segmentación, aquellos donde coincidan los tiempos de inicio de todas las hipótesis.

Ilustremos cómo se aplican estas reglas para el ejemplo de la figura 55.

Siguiendo el conjunto de reglas detalladas, los intervalos de comparación quedarán conformados de la siguiente manera:

Segmento	Inicio (s)	Final (s)
S1	0,48	0,67
S2	0,67	1,36
S3	1,36	2,07
S4	2,07	2,23
S5	2,23	3,03

Cuadro 16: Segmentación de la frase *"el diurético le causó pérdida de peso"* empleando la segunda estrategia para la comparación de contornos entonativos.

Partiendo de la segmentación de la tabla 16, uno se podría preguntar cuál sería la secuencia de hipótesis que minimiza el error de reconocimiento por palabras. Tal secuencia de *sub-hipótesis* óptimas en el sentido de maximización las tasas de reconocimiento, seria la que se muestra en la tabla 17:

Segmento	Elección Óptima
S1	(h1 | h2 | h3 | h4 | h5)
S2	h5
S3	h2
S4	(h1 | h2 | h3 | h4 | h5)
S5	(h1 | h2 | h3 | h4 | h5)

Cuadro 17: Fragmentación de la frase *"el diurético le causó pérdida de peso"* empleando la segunda estrategia para la comparación de contornos entonativos.

Ahora bien, lo que se debe determinar son las compatibilidades a nivel de similitudes entonativas por sub-segmentos. En la figura 56 se muestran las curvas de F0 original y predichas para cada hipótesis, correspondientes al segundo segmento del ejemplo de la figura 55.

Comparando a nivel de los segmentos presentados en la tabla 16 el contornos de F0 correspondiente a la señal original respecto al de cada hipótesis, se obtienen los siguientes resultados:

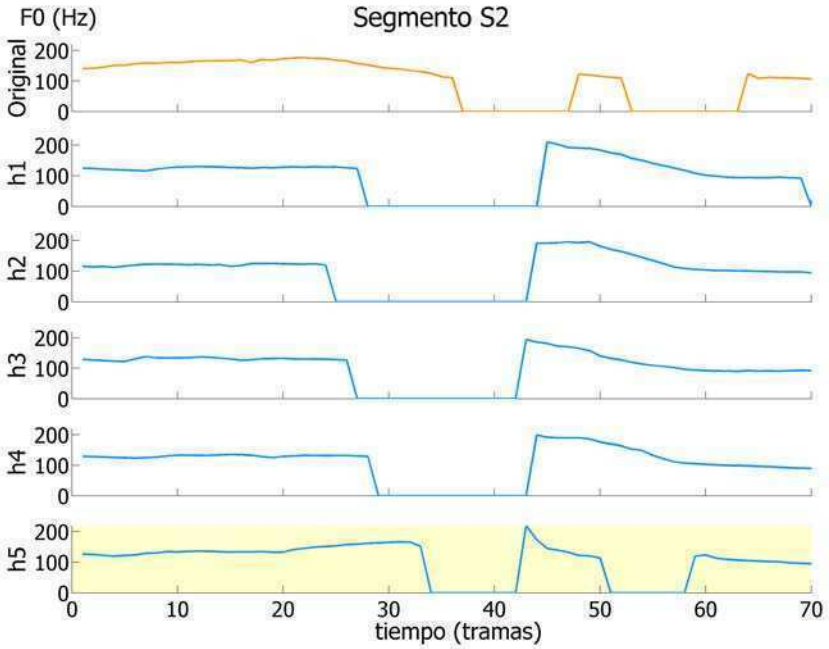

Figura 56: Comparación de contornos de F0 original (arriba), y generados asumiendo un predictor de F0 ideal, tal como se describe en la Sección 6.4, para el segundo segmento del ejemplo presentado en la figura 55. El segmento original, así como el correspondiente a la hipótesis **h5** corresponden a la palabra: *diurético*, mientras que las demás hipótesis corresponden a la palabra *directivo*. La curva de F0 más compatible con la entonación de la frase original se muestra sombreada y corresponde a la hipótesis **h5**.

Segmento	MSE (Hz)				
	h1	h2	h3	h4	h5
S1	**304,3**	492,7	494,7	455,3	644,4
S2	4663,6	5169,9	4478,9	4420,1	**2879,5**
S3	2578,8	**2011,0**	2519,8	2815,7	3622,3
S4	771,1	**490,3**	750,4	773,8	1089,0
S5	5925,7	5014,3	4923,0	**4721,9**	5762,9

Cuadro 18: Medida de distancia entre el valor real de la curva de F0 y el de cada hipótesis de reconocimiento según los segmentos definidos en la tabla 16.

Analizando la tabla 18, se puede ver que la comparación objetiva de perfiles de F0, empleando la estrategia de segmentación propuesta, resulta en una elección óptima de sub-segmentos de acuerdo a lo especificado en la tabla 17.

De esta forma la nueva hipótesis de reconocimiento, sintetizada a partir de la estrategia descripta sería la presentada en la figura 57.

Mejor Hipótesis Original				Hipótesis Sintetizada			
48	67	el	−1237	48	67	el	−1237
67	136	directivo	−4464	67	136	diurEtico	−4468
136	146	de	−555	136	146	de	−555
146	207	paso	−3664	146	207	causO	−3659
207	223	al	−896	207	223	al	−908
223	303	equivalentes	−4751	223	303	equivalentes	−4751

Figura 57: Mejor hipótesis para el sistema de RAH convencional (izquierda) y empleando la compatibilidad de F0 a nivel de sub-segmentos para la frase de entrada: *"el diurético le causó pérdida de peso"*. Para cada hipótesis se muestra el tiempo de inicio y fin de la palabra reconocida, su identidad y el valor de verosimilitud. Las unidades temporales están expresadas en múltiplos de 10 ms.

Como se puede apreciar en la figura 57, para el caso analizado la segunda estrategia presenta dos errores de reconocimiento menos que la hipótesis original. Mientras el error WER para la hipótesis original es de 85,72 %, para la hipótesis sintetizada es de 57,14 %, lo que significa una reducción relativa en la tasa de error por palabra de 33,34 %.

En todos los estudios presentados en este Capítulo, para ambas estrategias de comparación entonativa las distancias entre los perfiles de F0 se evaluaron empleando tres escalas diferentes: lineal medida en Hz, diferencias en términos de MSE, y escala perceptual medida en ERBs.

6.4 COMPATIBILIDAD ENTONATIVA EMPLEANDO UN PREDICTOR IDEAL

Uno de los argumentos en contra de la estrategia de predicción entonativa que se emplea en el sistema propuesto, viene dado por las características que presentan las hipótesis generadas por el reconocedor.

Como se comentó, este módulo de predicción entonativa se desarrolló pensando en generar contornos de F0 adecuados para sistemas de síntesis de habla. En esos sistemas por lo general se parte de oraciones bien formadas sintácticamente, y es fundamentalmente esa información sintáctica la que se explota durante la predicción.

Por otro lado, como se puede observar en el ejemplo presentado en la figura 55, las hipótesis emitidas por el reconocedor, por lo general carecen de una estructura sintáctica aceptable, lo que siembra dudas respecto a los resultados del proceso de predicción descripto.

Si bien se pudo determinar que el modelo de predicción ofrece un desempeño aceptable comparando los errores entre las curvas de F0 reales del corpus y las predichas para esas mismas oraciones, como se detalló en la Sección 6.3.1, ello no garantiza que esa bondad de predicción se extrapolara a las frases candidatas resultantes del proceso de

reconocimiento, con estructuras sintácticas tan diferentes a las vistas durante el entrenamiento del modelo.

De esa manera se decidió evaluar adicionalmente la viabilidad de la estrategia de reestimación de hipótesis propuesta, pero partiendo de una estimación *correcta* de las curvas de F0 para cada hipótesis de reconocimiento, es decir, asumiendo que se contaba con un predictor ideal de contornos entonativos.

Para hacerlo se reemplazó el módulo de predicción entonativa por la lectura por parte de un humano de las oraciones candidatas y la posterior extracción de sus curvas de F0 correspondientes.

6.5 MATERIALES Y MÉTODOS

Para evaluar la viabilidad del sistema propuesto se planteó contrastar el desempeño de un sistema de RAH estándar, en el que se utiliza como resultado final de reconocimiento la hipótesis de máxima verosimilitud, contra un sistema en que el resultado de reconocimiento sea la mejor hipótesis de la lista generada por el sistema estándar, reordenada de acuerdo a la compatibilidad suprasegmental. Como medida comparativa del desempeño se estableció la tasa de errores de reconocimiento por palabra.

6.5.1 Corpus de Datos Empleados

El cuerpo de datos empleados como caso de estudio para el sistema propuesto fue el subconjunto del corpus LIS-SECYT [64] correspondiente a la locutora femenina, que se detalló en la Sección 4.1.

La razón de utilizar como material de estudio el correspondiente a una locutora femenina es que los atributos entonativos estarían expresados de manera más notable que los encontrados en un locutor masculino. La locutora fue instruida para leer las oraciones con la mayor riqueza de variaciones tonales posibles. Para cada archivo correspondiente a una oración se etiquetó manualmente y de manera redundante por dos terapeutas diferentes, con instrucción musical, quienes distinguieron las ocurrencias prosódicas como grupos entonativos y acentos tonales.

Además se generaron las versiones del mismo conjunto de datos pero contaminadas con ruido tipo babble con SNR de 0 dB, 5 dB, 10 dB, 15 dB y 20 dB. Como fuente de ruido babble se emplearon grabaciones de la base de datos NOISEX [183].

Se efectuaron diez particiones diferentes del conjunto de datos, eligiendo en cada una de manera aleatoria un 80 % de los datos para entrenamiento y 20 % de los datos para evaluación. Con estos conjuntos se realizaron todas las pruebas empleando validación cruzada.

6.5.2 Sistema de RAH de Referencia

Como sistema de RAH de referencia se construyó un reconocedor basado en modelos ocultos de Markov (HMMs) dependiente del locutor. Se emplearon 32 monofonos como modelos acústicos, incluyendo un modelo para silencio y otro para pausa corta. Salvo los modelos de silencio, la estructura de los demás está conformada por monofonos de 3 estados de izquierda a derecha sin saltos.

El modelo de silencio presenta 3 estados pero con posibilidad de transición hacia atrás. El modelo de pausa corta es de un estado, enlazado al estado central del modelo de silencio. Estos HMM son semicontinuos que comparten 248 mezclas de funciones gausianas.

Se utilizaron monofonos en lugar de unidades más complejas debido a la escasa cantidad de datos disponibles para el entrenamiento de los modelos.

La parametrización de las señales acústicas fue la misma que la detallada en la sección 5.3.2: se empleó una codificación de 12 MFCC, más energía, delta y aceleración, conformando vectores de 39 elementos. Se utilizó ventanas de análisis de Hamming de 25 ms y paso de avance de 10 ms; normalización de energía a nivel de frase, sustracción de la componente de continua de la señal acústica y coeficiente del filtro de preénfasis de 0,97.

El modelo de lenguajes se creó emplando un corpus textual obtenido de periódicos argentinos, conteniendo alrededor de 17 millones de palabras.

Ese cuerpo de datos se generó a partir un proceso automático de captura, análisis del contenido textual de diarios publicados on-line, y su posterior procesamiento: expansión de números, fechas, acrónimos y símbolos especiales.

Este modelo está conformado por bigramas, que fueron estimadas empleando el programa SRILM [161]. En su estimación se aplicó la técnica de suavizado Kneser-Ney [94].

Las características generales del modelo de lenguaje se presentan en la tabla 19.

Atributo	Valor
Vocabulario (palabras)	54277
Número de nodos	54278
Entropía	8,328
Perplejidad	321,192

Cuadro 19: Características del modelo de lenguaje utilizado.

En las pruebas se utilizó la condición sin gramáticas, para observar el aporte de la información entonativa aislada del modelo de lenguaje, y un valor en el factor de ponderación del modelo de lenguaje de 5. Es decir, el valor de ponderación usado para el modelo de lenguaje fue nulo.

Para la construcción de los sistemas de reconocimiento se utilizó el paquete Hidden Markov Model Toolkit (HTK) [194], siguiendo el mismo procedimiento que el detallado en la Sección 5.3.4.

Para cada partición de datos se entrenó un conjunto diferente de modelos acústicos con el subconjunto de entrenamiento, y se realizó la verificación del desempeño con el conjunto de evaluación, generando como salida las mejores 5 hipótesis de reconocimiento.

Cabe aclarar que para la evaluación de sistemas empleando señales con ruido, se emplearon los modelos de HMM entrenados con el subconjunto de datos correspondiente, es decir con la misma SNR. Esto significa que si por ejemplo el conjunto de evaluación correspondía a 5 dB de SNR, el conjunto con el que se entrenaban los modelos acústicos era el de entrenamiento correspondiente presentando 5 dB de SNR.

Los experimentos se llevaron a cabo empleando la metodología de validación cruzada sobre 10 conjuntos, usando 80 % de los datos para entrenamiento y el 20 % restante para evaluación.

En una primera fase del trabajo se empleó un modelo para la predicción de contornos entonativos a partir de información léxica. Con estos modelos se determinaron las curvas de F0 para cada hipótesis de reconocimiento y se realizó una comparación respecto a la curva de F0 observada en la señal de audio a reconocer. De acuerdo al grado de similitud de dicha comparación se reordenaron las hipótesis y se estimaron las nuevas tasas de reconocimiento.

Se pueden distinguir dos partes dentro del modelo de predicción entonativa: por un lado un árbol de decisión y clasificación emplea información léxica y gramatical para estimar parámetros del modelo superposicional de Fujisaki, y por otro lado se convierten los parámetros predichos en una curva de F0.

La información de entrada a los árboles de clasificación implica obtener la segmentación fonética y las etiquetas de clase de palabra para cada hipótesis. La segmentación fonética se realiza empleando alineamiento forzado, y las etiquetas gramaticales son obtenidas empleando el POS tagger de Freeling [7].

Por otro lado, la comparación de contornos de entonación se realizó considerando frases completas, y debido a que todas tenían la misma longitud se determinó las diferencias comparando las curvas punto a punto. En la comparación se emplearon como escalas la diferencia en Hz, en MSE y en ERB. En la siguiente tabla se resumen los resultados obtenidos.

6.6 RESULTADOS

Una vez obtenida una curva de F0 para cada hipótesis, contrastando el ajuste de cada una de ellas con respecto al F0 real, se obtiene la nueva lista de N-mejores hipótesis de reconocimiento.

La tabla 20, muestra la comparación de desempeño entre el sistema de referencia y el propuesto en términos de porcentajes de WER promedio para los 10 conjuntos correspondientes al proceso de validación cruzada, y bajo las distintas condiciones de relación señal ruido.

Sistema	Comp. F0	SNR (dB)	Peso LM	μ(R) (%)	σ(R) (%)
Referencia	—	∞	0	44,10	1,44
Propuesto	ERBs			43,28	1,74
	Hz			47,46	1,87
	MSE			41,59	1,35
Referencia	—		5	93,63	0,59
Propuesto	ERBs			89,95	0,75
	Hz			87,28	1,13
	MSE			86,78	1,11
Referencia	—	20	0	41,65	1,38
Propuesto	ERBs			40,98	1,11
	Hz			44,77	1,33
	MSE			39,35	0,98
Referencia	—		5	93,48	0,60
Propuesto	ERBs			90,80	0,82
	Hz			87,00	1,07
	MSE			86,39	0,85
Referencia	—	15	0	36,03	1,25
Propuesto	ERBs			35,90	1,13
	Hz			39,00	1,09
	MSE			34,30	1,11
Referencia	—		5	92,70	0,62
Propuesto	ERBs			88,94	0,76
	Hz			86,74	0,68
	MSE			85,76	0,86
Referencia	—	10	0	27,47	0,73
Propuesto	ERBs			27,05	1,69
	Hz			29,58	1,05
	MSE			25,96	0,62
Referencia	—		5	89,79	0,66
Propuesto	ERBs			86,84	1,16
	Hz			84,27	0,64
	MSE			84,10	0,73
Referencia	—	5	0	15,08	1,07
Propuesto	ERBs			14,95	0,92
	Hz			16,79	1,10
	MSE			14,19	1,00
Referencia	—		5	77,81	1,80
Propuesto	ERBs			75,10	1,19
	Hz			73,21	1,77
	MSE			72,04	1,66

Cuadro 20: Comparación del desempeño entre los sistemas de RAH de referencia y el propuesto para distintas condiciones acústicas.

Los resultados de la tabla 20 incluyen el desempeño observado utilizando diferentes escalas para comparar las curvas de F0. En todos los casos reportados en este capítulo se evaluó esta estrategia empleando reestimación sobre las 5 mejores hipótesis de reconocimiento.

Como puede observarse en la tabla 20, las tasas de reconocimiento empleando el modelo propuesto solamente mejora el desempeño del sistema de referencia cuando no se utiliza el modelo de lenguaje. Por otro lado, cuando se emplea modelo de lenguaje con peso 5 (valor que puede considerarse estándar), el reordenamiento propuesto empeora la calidad del ordenamiento original.

En la figura 58 se muestra un gráfico de cajas comparando el desempeño del sistema original respecto al propuesto, adoptando un factor de ponderación de 5 para el modelo de lenguaje estándar.

Dado que para el conjunto completo de testeo la escala de similitudes que mostró el menor WER resultante fue el RMSE medido en Hz, en los análisis siguientes se comparó solamente ese caso con respecto al sistema de referencia.

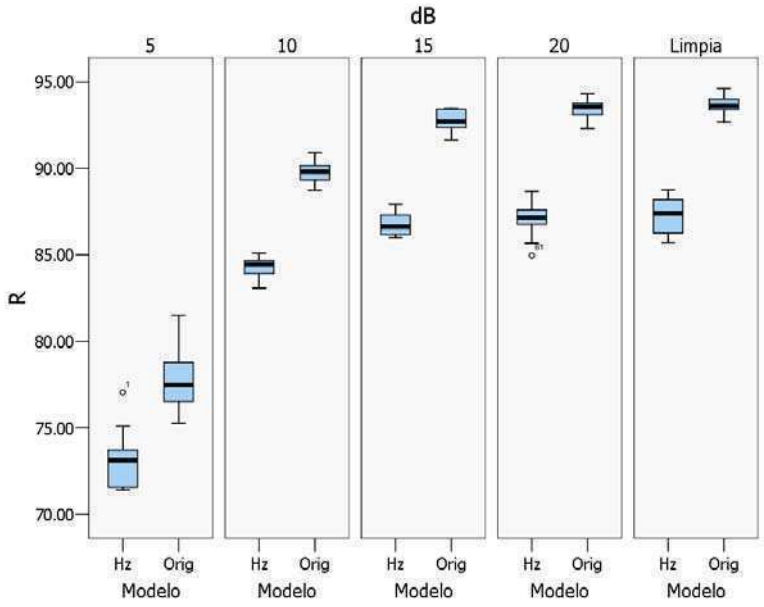

Figura 58: Gráfico de cajas correspondiente a las tasas de reconocimiento del sistema original y propuesto (empleando la escala Hz) ante diferentes grados de degradación de la señal a reconocer, empleando un factor de ponderación de 5 para el modelo de lenguaje.

En la figura 59 se muestra un gráfico de cajas comparando el desempeño del sistema original respecto al propuesto empleando la escala Hz, sin considerar el aporte del modelo de lenguaje.

Para descartar que las diferencias en el desempeño observado entre el sistema propuesto y el de referencia de la tabla 20 se deba al azar, se efectuó un estudio estadístico de igualdad de medias entre estos casos.

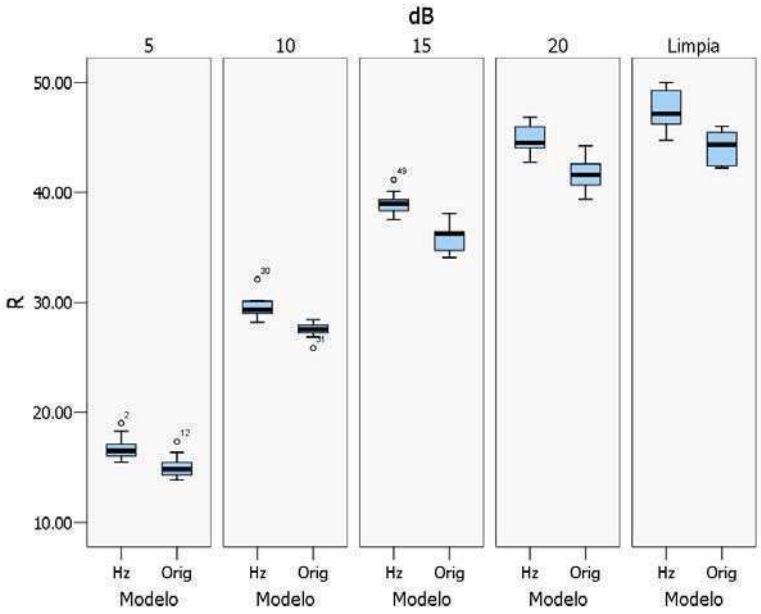

Figura 59: Gráfico de cajas correspondiente a las tasas de reconocimiento del sistema original y propuesto (empleando la escala Hz) ante distintas condiciones de relación señal ruido, para el caso de un factor nulo en el peso del modelo de lenguaje.

En la tabla 21 se presentan los resultados de la prueba Anova, contrastando los valores de tasas de reconocimiento entre el sistema original y el propuesto adoptando la escala de Hz en el módulo de comparación entonativa.

Peso LM	SNR (dB)	Est. de Levene	Sig.	F	Sig.
0	5	0,012	0,914	12,386	0,02
0	10	0,606	0,447	27,285	0,00
0	15	0,113	0,740	32,142	0,00
0	20	0,000	0,983	26,425	0,00
0	∞	1,084	0,312	20,356	0,00
5	5	0,400	0,844	33,1730	0,00
5	10	0,015	0,905	357,670	0,00
5	15	0,160	0,694	422,524	0,00
5	20	1,649	0,215	281,823	0,00
5	∞	7,203	0,015	248,342	0,00

Cuadro 21: Análisis de igualdad de medias (Anova) para las tasas de reconocimiento comparando el sistema original y el propuesto empleando la escala de Hz.

Analizando los resultados del estadístico de Levene, y asumiendo 0,05 como valor mínimo, para todos los casos de la tabla 21, salvo el

último, se puede aceptar la hipótesis nula de varianzas similares entre los dos grupos.

Dicha igualdad de varianzas es un requisito para la validez de la prueba Anova, sin embargo como para las dos clases consideradas el número de instancias es similar, sería válido considerar que la prueba de Anova es válida también para el último caso de la tabla.

Con respecto a los estadísticos correspondientes a la prueba Anova (las dos últimas columnas de la tabla 21), se ve que para todos los casos el desempeño del sistema propuesto, empleando una escala lineal para comparar las curvas de F0, difiere significativamente respecto al desempeño del sistema de referencia (nuevamente usando un $p = 0{,}05$).

Posteriormente se efectuaron las mismas pruebas pero empleando la segunda estrategia de comparación entonativa; la comparación sobre la base de subsegmentos de oraciones. La tabla 22 resume los resultados obtenidos para este caso.

Los resultados muestran nuevamente un comportamiento que en líneas generales es similar al descripto para la primera estrategia de contraste entonativo. El modelo propuesto es mejor consistentemente respecto al de referencia en caso de no emplear modelo de lenguaje, pero inferior en caso contrario.

Además, la escala que mejores resultados exhibe para el caso en que se emplea un coeficiente de ponderación nulo en el modelo de lenguaje, es la lineal, medida en Hz.

Respecto a las otras formas de cuantificar las similitudes entre curvas de F0 que se muestran en la tabla 22, se puede ver que el método de correlación empleando interpolación en segmentos sordos de la curva de referencia (R_2^2) produce mejores resultados que el caso en que se emplea la curva sin interpolar (R^2).

En este caso también se encontraron que las diferencias en el desempeño entre el método propuesto y el de referencia fueron significativos a nivel 0,05, tanto para el caso de utilizar o no modelo de lenguaje.

Finalmente se evaluó la propuesta detallada en la Sección 6.4.

Empleando un subconjunto de 50 oraciones originales, se grabó la versión leída para cada una de sus 5 mejores hipótesis de reconocimiento, así como de la oración correcta.

La operación de comparación de perfiles de F0 en el caso del predictor entonativo era sencilla, ya que todas las curvas se encontraban alineadas, respetando las segmentaciones obtenidas por el sistema de RAH.

En el caso del predictor *ideal* la misma operación es más compleja dado que tanto los instantes de inicio como las duraciones de cada palabra, que fueron grabadas de manera independiente no mostraban tales alineamientos.

Sistema	Comp. F0	SNR (dB)	Peso LM	$\mu(R)$ (%)	$\sigma(R)$ (%)
Referencia	–	∞	0	44,226	1,22
Propuesto	Hz			46,71	1,32
	R^2			45,189	1,13
	R_2^2			45,87	1,26
Referencia	–		5	93,58	0,30
Propuesto	Hz			88,74	0,86
	R^2			88,73	0,79
	R_2^2			89,98	0,99
Referencia	–	20	0	41,93	1,39
Propuesto	Hz			44,90	1,21
	R^2			42,90	1,25
	R_2^2			43,52	1,36
Referencia	–		5	93,54	0,54
Propuesto	Hz			86,92	1,22
	R^2			87,39	0,83
	R_2^2			88,85	0,96
Referencia	–	15	0	36,35	1,35
Propuesto	Hz			37,61	1,16
	R^2			38,16	1,23
	R_2^2			39,38	1,27
Referencia	–		5	92,75	0,63
Propuesto	Hz			86,23	1,07
	R^2			88,34	0,61
	R_2^2			86,54	0,85
Referencia	–	10	0	27,74	0,97
Propuesto	Hz			30,10	1,06
	R^2			28,72	0,93
	R_2^2			28,68	1,01
Referencia	–		5	89,88	0,65
Propuesto	Hz			84,11	0,96
	R^2			84,08	0,96
	R_2^2			85,99	0,57
Referencia	–	5	0	15,10	1,04
Propuesto	Hz			16,70	1,02
	R^2			15,75	1,05
	R_2^2			15,65	1,08
Referencia	–		5	77,89	1,83
Propuesto	Hz			72,49	1,53
	R^2			72,75	1,82
	R_2^2			74,33	1,68

Cuadro 22: Desempeño del sistema de referencia y el propuesto para distintas condiciones acústicas empleando comparación por subsegmentos

Por lo tanto en este caso se debía recurrir a un procedimiento de alineación entre los segmentos de las hipótesis alternativas.

Se decidió llevar a cabo tal alineamiento adaptando las duraciones de las nuevas grabaciones a las de las encontradas para esos segmentos en la salida del reconocedor.
El procedimiento completo de adaptación de instantes de inicio y duraciones de las hipótesis consistió en los siguientes pasos.

- En primer lugar se efectuó la segmentación de las nuevas grabaciones a nivel de palabras.
- Tomando como base los tiempos de inicio y fin de cada palabra en las salidas del reconocedor, se modificaron las duraciones de las nuevas grabaciones empleando las segmentaciones grafémicas y el algoritmo *PSola* de **Praat** [17].
- Una vez que todas las grabaciones fueron adaptadas a las duraciones de las hipótesis originales, se utilizó el algoritmo RAPT para estimar las curvas de F0 de cada una.

En la figura 60 se muestran las curvas de F0 generadas a partir del procedimiento detallado en este apartado y correspondientes al segundo segmento del ejemplo de la figura 55.

En estas pruebas se empleó la versión de habla limpia del subconjunto mencionado y adoptando un peso en el modelo de lenguajes de 5. Los resultados obtenidos para el conjunto de estas oraciones efectuando las comparaciones de contornos entonativos empleando la escala medida en Hz, mostraron un aumento en términos absolutos de 1,52 % respecto a las tasas de reconocimiento de referencia: 91,89 % vs. 93,41 %.
Las diferencias más distintivas entre las curvas de F0 así obtenidas, y las determinadas a partir del predictor entonativo estuvieron dadas en las regiones sordas. Mientras en este caso las regiones sonoras y sordas correspondientes a cada hipótesis se preservaron, para el caso del módulo de predicción las regiones sordas vienen interpoladas respecto a las regiones sonoras circundantes.
Esto último supone pérdida de información potencialmente útil en la comparación de perfiles de F0 para desambiguar hipótesis.

6.7 CONCLUSIONES

Este trabajo muestra una metodología nueva para hacer uso del conocimiento entonativo durante el reordenamiento de las hipótesis de reconocimiento.

En primer lugar se demostró las diferencias existentes entre los contornos de F0 de las hipótesis candidatas, y la viabilidad de la propuesta, simulando un predictor entonativo ideal. En este estudio preliminar se pudo verificar que si se dispone de una buena aproximación de las curvas de F0 correspondientes a cada hipótesis, es posible en muchos

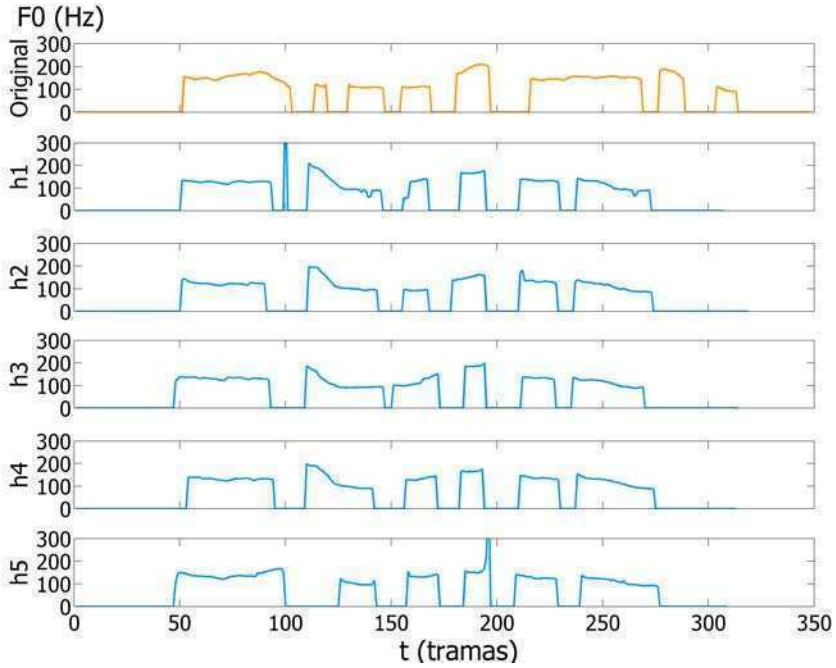

Figura 60: Comparación de contornos de F0 original (arriba), y generados asumiendo un predictor de F0 ideal, para el ejemplo presentado en la figura 55.

casos distinguir la mejor secuencia candidata a partir de su similitud con el perfil de F0 del fragmento de habla a reconocer.

Por otra parte empleando un predictor entonativo del campo de síntesis del habla para estimar el contorno de F0 de las hipótesis, y como material de estudio habla continua correspondiente al Español de Buenos Aires, el método propuesto demostró ser útil para mejorar el desempeño de un reconocedor del habla convencional, en caso de no utilizar modelos de lenguaje. La mejora del reconocedor empleando el método propuesto para estos casos es de un 3 % en términos absolutos, sobre la tasas de reconocimiento.

Esta mejora indica que existe información entonativa que se puede aprovechar a la hora de reestimar las hipótesis reconocidas.

Sin embargo, la misma metodología demostró ser contraproducente en caso de utilizar modelo de lenguaje durante la decodificación. Esto hace necesario replantear la forma de ponderación de las evidencias entonativas respecto a las verosimilitudes exhibidas por cada hipótesis.

Resulta lógico pensar que los contornos de entonación para una determinada oración sean dependientes del locutor.

De esta forma se debe notar que el modelo propuesto utilizó un módulo de predicción ajustado al locutor a reconocer. Esto puede significar la necesidad de adaptar este módulo a las características entonativas de nuevos locutores.

También se debe notar que la misma técnica presentada puede extenderse para usar otros atributos prosódicos durante el proceso de desambiguación.

Los resultados obtenidos en la serie de estudios presentados indican que no se puede aceptar la segunda hipótesis de trabajo: el método de reestimación de hipótesis empleando información entonativa no resultó más útil ante mayores niveles de degradación en la señal a reconocer.

Esto se puede justificar a partir del siguiente razonamiento. El método de reestimación propuesto resulta útil siempre y cuando dentro de las hipótesis de reconocimiento se encuentren las palabras correctas. A medida que la señal a reconocer se ve más degradada, es menos probable que lleguen a la fase final de reconocimiento fragmentos de hipótesis correctos y por lo tanto pierde efectividad la propuesta de reestimación.

Es posible que el aporte de la información suprasegmental a la robustez del reconocimiento sea viable solamente en las fases de preprocesamiento y durante el proceso de búsqueda, no así durante el posprocesamiento.

Finalmente, estudiando el desempeño del sistema propuesto a través de un predictor entonativo ideal se reveló la conveniencia que los contornos de F0 para las hipótesis conserven la información respecto a regiones sonoras y sordas, para mejorar el proceso de desambiguación de segmentos candidatos. El módulo de predicción entonativa integrado en el modelo propuesto pierde esta información potencialmente valiosa.

Como trabajos futuros, resta evaluar la carga computacional adicional que implica el método propuesto sobre el sistema de RAH convencional, proponer una nueva alternativa para pesar las evidencias de similitud entonativa respecto a la verosimilitud de cada hipótesis en caso de emplear modelos de lenguaje, y el comportamiento bajo un escenario de múltiples locutores.

En [10] se discute que el mismo texto se puede pronunciar con diferentes curvas de entonación sin perder naturalidad, teniendo en cuenta esta situación, se planea extender el método propuesto al uso de un módulo de predicción entonativa que ofrezca un conjunto de curvas candidatas, en lugar de solamente una. Esto representa una visión más realista respecto de la relación entre atributos léxico-sintácticos y entonación, ya que se puede observar que a partir de un texto determinado los locutores eligen uno de un conjunto de contornos de F0 posibles.

7 CONCLUSIONES

A continuación se presentan las principales conclusiones de esta Tesis.

En el **Capítulo** 1 se presentaron una serie de argumentos que motivan la investigación del uso de información suprasegmental en el RAH.

Se describieron aspectos teóricos sobre producción, percepción y reconocimiento del habla en el ser humano y en máquinas, así como los aspectos fundamentales de la prosodia, lo que constituye el marco teórico de la investigación. Además se detallaron antecedentes en la integración de información suprasegmental y RAH.

Los **Capítulos** 2 y 3 presentan en detalle los temas del Reconocimiento Automático del Habla e Información Suprasegmental, en este último caso haciendo especial hincapié en las particularidades del idioma Español.

En el **Capítulo** 4 se estudian algunos correlatos acústicos distintivos de la información suprasegmental en su relación a información léxica para el Español de Argentina. Se evalúa la posibilidad de distinguir agrupamientos a nivel de frases entonativas observando en la duración y curvas de F0 y energía de las frases, se estudia la relación entre el número de picos encontrados en los perfiles de F0 y la cantidad de palabras de contenido presentes en una frase, y finalmente se experimenta la posibilidad de establecer a través de rasgos suprasegmentales la clase de acento léxico de las palabras finales de oración.

Se pudo demostrar la primera hipótesis, detectando a través de un proceso de autoorganización la presencia de agrupamiento naturales en las frases entonativas. Éstas se pudieron considerar separadas en: frases iniciales de oración, frases intermedias y frases finales, constituyendo estas últimas el conglomerado más homogéneo. A partir de ese agrupamiento se propuso un mecanismo de clasificación automática de frases entonativas, que mostró un desempeño global en torno al 75 %.

Respecto a la posibilidad de correlacionar el número de picos observados en las curvas de F0 con la cantidad de palabras de contenido presentes en las frases, se propusieron dos algoritmos para la determinación del número de picos en las curvas de F0, y se evaluó el empleo de estos detectores de picos como clasificadores del número de palabras de contenido. Los resultados obtenidos mostraron tasas de clasificación en torno al 44 %, sugiriendo que no es posible establecer una relación tan directa entre estos dos elementos.

Como alternativa a esta hipótesis se cree que sería conveniente efectuar una detección de prominencias, para realizar la asociación con el

número de palabras de contenido presentes en una frase, en vez de recurrir directamente a un solo atributo prosódico.

Finalmente en relación a la tercera hipótesis, se pudo determinar la viabilidad de clasificar palabras finales de oración de acuerdo a sus acentos léxicos. Se expuso esa viabilidad especificando patrones comunes en la morfología de las curvas de F0 correspondientes a finales de oración, como mediante análisis estadísticos a nivel de atributos suprasegmentales sobre las últimas tres sílabas de cada oración.

Los resultados de la serie de experimentos de este Capítulo, abren posibles líneas de investigación tanto para comprender mejor cómo se manifiestan los fenómenos prosódicos en nuestro idioma, como para su aplicación específica dentro de sistemas de síntesis y reconocimiento del habla para el Español de Argentina.

En el **Capítulo** 5 se presenta una alternativa para la utilización de información acentual dentro de los modelos acústicos de un sistema de RAH. Se propone y evalúa la utilización de acentos léxicos en los modelos acústicos, creando modelos separados para vocales acentuadas e inacentuadas lexicalmente. La evaluación se lleva a cabo comparando el desempeño del sistema propuesto con respecto a un sistema de referencia estándar.

Empleando como corpus de evaluación habla telefónica leída correspondiente al Español de Argentina, el empleo de los monofonos acentuados propuestos permitió mejorar el porcentaje de reconocimiento en un 1,78 % con respecto a los trifonos dependientes del contexto entre palabras, con una reducción del 89 % en el tiempo de procesamiento.

Como en esos estudios distinguen las vocales acentuadas de las inacentuadas sólo a partir del texto, deja abierta la posibilidad de experimentar realizando esa distinción a partir de información acústica.

El **Capítulo** 6 propone una nueva metodología para hacer uso del conocimiento entonativo durante el reordenamiento de las hipótesis de reconocimiento.

El método propuesto demostró ser útil para mejorar el desempeño de un reconocedor del habla de referencia para el Español de Buenos Aires, dependiente del locutor, en caso de no utilizar modelos de lenguaje.

También se demostró que el aporte en esas circunstancias no aumenta para menores relaciones señal-ruido. Lo que se puede justificar considerando que el método de reestimación propuesto resulta útil siempre y cuando dentro de las hipótesis de reconocimiento se encuentren las palabras correctas. A medida que la señal a reconocer se ve más degradada, es menos probable que lleguen a la fase final de reconocimiento fragmentos de hipótesis correctos y por lo tanto pierde efectividad la propuesta de reestimación.

Es posible que el aporte de la información suprasegmental a la robustez del reconocimiento sea viable solamente en las fases de preprocesamiento y durante el proceso de búsqueda, no así durante el posprocesamiento.

Resulta lógico pensar que los contornos de entonación para una determinada oración sean dependientes del locutor.

De esta forma se debe notar que el modelo propuesto utilizó un módulo de predicción ajustado al locutor a reconocer. Esto puede significar la necesidad de adaptar este módulo a las características entonativas de nuevos locutores.

Como trabajo futuro se planea extender esta idea al uso de un módulo de predicción entonativa que ofrezca un conjunto de curvas entonacionales candidatas, en lugar de solamente una, ya que esto sería más ajustado a la realidad.

En resumen, a través de esta Tesis se ha estudiado diferentes alternativas para el empleo de información suprasegmental en el reconocimiento automático del habla.

Se describieron y estudiaron algunas fuentes de información potencialmente valiosa para el RAH, asociada con atributos suprasegmentales.

Además se propuso una alternativa para emplear esta fuente de información dentro de los modelos acústicos de un reconocedor.

También se presentaron algunas ideas novedosas respecto a cómo se podría emplear información entonativa en la reestimación de hipótesis, durante el posprocesamiento de los sistemas de RAH.

Finalmente, esta Tesis es una de las primeras investigaciones realizada en pos de emplear información prosódica para el RAH del Español hablado en Argentina, y constituye así un aporte a este amplio área de estudio.

BIBLIOGRAFÍA

[1] Abercrombie, D. (1967), *Elements of General Phonetics*, Edinburgh University Press, Edinburgo, Inglaterra. (Citado en la página 109.)

[2] Adami, A. G. (2007), «Modeling Prosodic Differences for Speaker Recognition», *Speech Communication*, vol. 49, pp. 277–291. (Citado en la página 48.)

[3] Ananthakrishnan, S. y Narayanan, S. (2007), «Improved Speech Recognition Using Acoustic and Lexical Correlates of Pitch Accent in a N-Best Rescoring Framework», in «Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)», pp. 873–876, Honolulu, Hawai. (Citado en la página 169.)

[4] Andruski, J. E. y Costello, J. (2004), «Using Polynomial Equations to Model Pitch Contour Shape in Lexical Tones: an Example from Green Mong», *Journal of the International Phonetic Association*, vol. 34 (2), pp. 125–140. (Citado en la página 141.)

[5] Arias, J. P., Becerra-Yoma, N. y Vivanco, H. (2010), «Automatic Intonation Assesment for Computer Aided Language Learning», *Speech Communcation*, vol. 52 (3), pp. 254–267. (Citado en la página 48.)

[6] Atal, B. S. y Hanauer, S. L. (1971), «Speech Analysis and Synthesis by Linear Prediction», *Journal of the Acoustical Society of America*, vol. 50 (2), pp. 637–655. (Citado en la página 30.)

[7] Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., y Padró, M. (2006), «FreeLing 1.3: Syntactic and Semantic Services in an Open-Source NLP Library», in «Proceedings of the 5th International Conference on Language Resources and Evaluation», pp. 48–55, Génova, Italia. (Citado en las páginas 182 y 193.)

[8] Aubert, X. y Dugast, C. (1995), «Improved Acoustic-Phonetic Modeling in Philips' Dictation System by Handling Liaisons and Multiple Pronunciations», in «Proceedings of the 4th European Conference on Speech Communication and Technology (Eurospeech)», pp. 767–770, ISCA, Madrid, España. (Citado en la página 66.)

[9] Aull, A. M. (1985), «Lexical Stress and its Application to Large Vocabulary Speech Recognition», in «Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)», pp. 1449–1552, Tampa, Estados Unidos. (Citado en la página 45.)

[10] Bassi, A., Becerra-Yoma, N. y Loncomilla, P. (2006), «Estimating Tonal Prosodic Discontinuities in Spanish using HMM», *Speech Communication*, vol. 48, pp. 1112–1125. (Citado en la página 201.)

[11] Bates, R. A. (2003), *Speaker Dynamics as a Source of Pronunciation Variability for Continuous Speech Recognition Models*, Phd, University of Washington, Washington, Estados Unidos. (Citado en la página 50.)

[12] Beckman, M., Díaz-Campos, M., McGory, J. y Morgan, T. (2002), «Intonation Across Spanish, in the Tones and Break Indices Framework», *Probus. Special Issue on Intonation in Romance*, vol. 14, pp. 9–36. (Citado en la página 117.)

[13] Beckman, M. E., Hirschberg, J. y Shattuck-Hufnagel, S. (2005), «The Original ToBI System and the Evolution of the ToBI Framework», in Jun, S.-A., editor, «Prosodic Typology: The Phonology of Intonation and Phrasing», cap. 1, pp. 9–54, Oxford University Press, Oxford, Inglaterra. (Citado en la página 127.)

[14] Bellman, R. (1957), *Dynamic Programming*, Princeton University Press, Nueva Jersey, Estados Unidos. (Citado en la página 83.)

[15] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V. y C., W. (2007), «Automatic Speech Recognition and Speech Variability: A Review», *Speech Communication*, vol. 49 (10-11), pp. 763–786. (Citado en las páginas 3 y 27.)

[16] Black, A., Lenzo, K. y Pagel, V. (1998), «Issues in Building General Letter to Sound Rules», in «Proceedings of the ESCA/COCOSDA Workshop on Speech Synthesis», pp. 767–770, Jenolan Caves, Australia. (Citado en la página 65.)

[17] Boersma, P. (2001), «Praat, a System for Doing Phonetics by Computer», *Glot International*, vol. 9-10 (5), pp. 341–345. (Citado en la página 199.)

[18] Borzone, A., Signorini, A. y Massone, M. (1982), «Rasgos Prosódicos: el Acento», *Fonoaudiológica*, vol. 28, pp. 19–36. (Citado en la página 168.)

[19] Breiman, L., Friedman, J. y Stone, C. (1984), *Clasification and Regression Trees*, Chapman and Hall, Nueva York, Estados Unidos. (Citado en la página 183.)

[20] Campbell, N. (1990), «Evidence for a Syllable-based Model of Speech Timing», in «Proceedings of the International Conference on Spoken Language Processing», pp. 9–12, Kobe, Japón. (Citado en la página 99.)

[21] Chan, O. (2008), *Prosodic Features for a Maximum Entropy Language Model*, Phd, School of Electrical, Electronic and Computer Engineering, The University of Western Australia. (Citado en la página 180.)

[22] Chen, S. F., Kingsbury, B., Mangu, L., Povey, D., Saon, G., Soltau, H. y Geoffrey, Z. (2006), «Advances in Speech Transcription at IBM Under the DARPA EARS Program», *IEEE Transactions On Audio, Speech, And Language Processing*, vol. 14 (5), pp. 1596–1608. (Citado en la página 3.)

[23] Chomsky, N. (1995), *The Minimalist Program. Current Studies in Linguistics*, The MIT PRess, Cambridge, Estados Unidos. (Citado en la página 111.)

[24] Chomsky, N. y Halle, M. (1968), *The Sound Patterns of English*, The MIT Press, Cambridge, Estados Unidos. (Citado en la página 117.)

[25] Chung, G. y Seneff, S. (1999), «A Hierarchical Duration Model for Speech Recognition based on the ANGIE Framework», *Speech Communication*, vol. 27 (2), pp. 113–134. (Citado en la página 50.)

[26] Colantoni, L. y Gurlekian, J. (2004), «Convergence and Intonation: Historical Evidence from Buenos Aires Spanish», *Bilingualism: Language and Cognition*, vol. 7 (2), pp. 107–119. (Citado en las páginas 52, 108, 147 y 168.)

[27] Cooley, J. W. y Tukey, J. W. (1965), «An Algorithm for the Machine Calculation of Complex Fourier Series», *Mathematics of Computation*, vol. 19 (90), pp. 297–301. (Citado en la página 30.)

[28] Crystal, T. H. y House, A. S. (1990), «Articulation Rate and the Duration of Syllables and Stress Groups in Connected Speech», *Journal of the Acoustic Society of America*, vol. 88 (1), pp. 101–112. (Citado en la página 99.)

[29] Cutler, A., Dahan, D. y van Donselaar, W. (1997), «Prosody in the Comprehension of Spoken Language: a Literature Review», *Language and Speech*, vol. 40 (2), pp. 141–201. (Citado en la página 120.)

[30] D' Imperio, M., Elordieta, G., Frota, S., Prieto, P. y Vigário, M. (2005), «Intonational Phrasing in Romance: The Role of Syntactic and Prosodic Structure», in Frota, S., Vigário, M. y Freitas, M., editores, «Prosodies: with Special Reference to Iberian Languages», pp. 59–98, Mouton de Gruyter, Berlín, Alemania. (Citado en las páginas 114 y 119.)

[31] Davis, M. y Johnsrude, I. (2007), «Hearing Speech Sounds: Top-down Influences on the Interface between Audition and Speech Perception», *Hearing Research*, vol. 229 (1-2), pp. 132–147. (Citado en la página 26.)

[32] Davis, S. y Mermelstein, P. (1980), «Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences», *IEEE Transactions On Audio, Speech, And Language Processing*, vol. 28 (4), pp. 357–366. (Citado en la página 57.)

[33] De Pijper, J. R. (1983), *Modelling British English Intonation*, Foris, Dordrecht, Holanda. (Citado en la página 102.)

[34] Denes, P. B. y Pinson, E. N. (1993), *The Speech Chain: the Physics and Biology of Spoken Language*, W.H. Freeman. (Citado en la página 4.)

[35] Downey, S. y Wiseman, R. (1997), «Dynamic and Static Improvements to Lexical Baseforms», in «Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech)», pp. 1027–1030, ISCA, Rodas, Grecia. (Citado en la página 66.)

[36] Dudley, H. (1939), «The Vocoder», *Bell Labs. Record*, vol. 18 (4), pp. 122–126. (Citado en la página 11.)

[37] Duifhuis, H., Willems, L. F. y Sluyter, R. J. (1982), «Measurement of Pitch in Speech: An Implementation of Goldstein's Theory of Pitch Perception», *Journal of the Acoustical Society of America*, vol. 71 (6), pp. 1568–1580. (Citado en la página 131.)

[38] Elordieta, G., Frota, S., Prieto, P. y Vigário, M. (2003), «Effects of Constituent Length and Syntactic Branching on Intonational Phrasing in Ibero-Romance», in «Proceedings of the 15th International Congress of Phonetic Sciences», pp. 487–490, Barcelona, España. (Citado en la página 119.)

[39] Eluned, S. y Carey, M. (1996), «Language Independent Gender Identification», in «Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)», pp. 685–688, Atlanta, Estados Unidos. (Citado en la página 48.)

[40] Evin, D., Gurlekian, J. y Torres, H. (2010), «N-Best Rescoring Based on Intontational Prediction for a Spanish ASR System», in «21 Konferenz Elektronische Sprachsignalverarbeitung (ESSV)», pp. 234–242, Berlín, Alemania. (Citado en la página 181.)

[41] Evin, D., Univaso, P., Clausse, A., Milone, D. y Gurlekian, J. (2009), «Reconocimiento Automático de Habla Empleando Información de Acento Lexical en los Modelos Acústicos», in «Anales de las 38º Jornadas Argentinas de Informática, Simposio Argentino de Tecnología», pp. 189–200, Mar del Plata, Argentina. (Citado en la página 169.)

[42] Face, T. (2003), «Intonation in Spanish Declaratives: Differences between Lab Speech and Espontaneous Speech», *Catalan Journal of Linguistics*, vol. 2, pp. 115–131. (Citado en la página 116.)

[43] Fernández Planas, A. (2005), «Aspectos Generales Acerca del Proyecto Internacional AMPER en España», *Estudios de Fonética Experimental*, vol. 14, pp. 13–27. (Citado en la página 156.)

[44] Fosler-Lussier, E. (2003), «A Tutorial on Pronunciation Modeling for Large Vocabulary Speech Recognition», in Renals, S. y Grefenstette, G., editores, «Text and Speech-Triggered Information Access, LNAI 2705», cap. 32, pp. 38–77, Springer-Verlag, Berlín, Alemania. (Citado en las páginas 50 y 65.)

[45] Fox, A. (2000), *Prosodic Features and Prosodic Structure. The Phonology of Suprasegmentals*, Oxford University Press, Oxford, Inglaterra. (Citado en la página 97.)

[46] Frota, S. (2000), *Prosody and Focus in European Portuguese: Phonological Phrasing and Intonation*, Garland, Nueva York, Estados Unidos. (Citado en la página 113.)

[47] Fujisaki, H. (1992), «The Role of Quantitative Modeling in the Study of Intonation», in «Proceedings of the International Symposium on Japanese Prosody», pp. 163–174, Nara, Japón. (Citado en la página 42.)

[48] Fujisaki, H. y Hirose, K. (1982), «Modeling the Dynamic Characteristics of Voice Fundamental Frequency with Applications to Analysis and Synthesis of Intonation», in «Proceedings of the 13th International Congress of Linguists», pp. 121–130, Tokio, Japón. (Citado en la página 123.)

[49] Fujisaki, H., Ohono, S., Ichi Nakamura, K., Guirao, M. y Gurlekian, J. (1994), «Analysis of Accent and Intonation in Spanish Based on a Quantitative Model», in «Proceedings of the International Conference on Spoken Language Processing (ICSLP)», pp. 355–358, Yokohama, Japón. (Citado en las páginas 139 y 183.)

[50] Fukada, T., Yoshimura, T. y Sagisaka, Y. (1999), «Automatic Generation of Multiple Pronunciations Based on Neural Networks», *Speech Communication*, vol. 27 (1), pp. 63–73. (Citado en la página 65.)

[51] Gales, M. y Young, S. (2007), «The Application of Hidden Markov Models in Speech Recognition», *Foundations and Trends in Signal Processing*, vol. 1 (3), pp. 195–304. (Citado en la página 54.)

[52] Garrido, J., Listerri, J., de la Mota, C. y Ríos, A. (1993), «Prosodic Differences in Reading Style: Isolated vs. Contexualized Sentences», in «Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech)», pp. 573–576, Berlín, Alemania. (Citado en la página 116.)

[53] Garrido, J. M. (1996), *Modelling Spanish Intonation for Text-to-Speech Applications*, Phd, Departament de Filologia Espanyola, Universitat Autonoma de Barcelona, Barcelona, España. (Citado en las páginas 116 y 126.)

[54] Gaskell, G. M. y Marslen-Wilson, W. D. (1997), «Integrating Form and Meaning: A Distributed Model of Speech Perception», *Language and Cognitive Processes*, vol. 12 (5-6), pp. 613–656. (Citado en la página 24.)

[55] Goldinger, S. D. (1998), «Echoes of Echoes? An Episodic Theory of Lexical Access», *Psychological Review*, vol. 105 (2), pp. 251–279. (Citado en la página 24.)

[56] Goldstein, J. L. (1973), «An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones», *Journal of the Acoustical Society of America*, vol. 54 (1), pp. 1496–1516. (Citado en la página 131.)

[57] Goldwater, S., Jurafsky, D. y Manning, C. D. (2010), «Which Words are Hard to Recognize? Prosodic, Lexical and Disfluency Factors that Increase Speech Recognition Error Rates», *Speech Communication*, vol. 52 (3), pp. 181–200. (Citado en la página 50.)

[58] Grabe, E., Kochanski, G. y Coleman, J. (2007), «Connecting Intonation Labels to Mathematical Descriptions of Fundamental Frequency», *Language And Speech*, vol. 50 (3), pp. 281–310. (Citado en la página 141.)

[59] Greenberg, S. y Chang, S. (2000), «Linguistic Dissection of Switchboard Corpus Automatic Speech Recognition Systems», in «Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millenium», pp. 195–202, París, Francia. (Citado en la página 28.)

[60] Greenberg, S. y Fosler-Lussier, E. (2000), «The Uninvited Guest: Information's Role in Guiding the Production of Spontaneous Speech», in «Proceedings of the Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling», Kloster Seeon, Alemania. (Citado en la página 28.)

[61] Gurlekian, J., Colantoni, L. y Torres, H. (2001), «El Alfabeto Fonético SAMPA y el Diseño de Córpora Fonéticamente Balanceados», *Fonoaudiológica*, vol. 47 (3), pp. 58–70. (Citado en la página 171.)

[62] Gurlekian, J., Torres, H. y Colantoni, L. (2003), «Evaluación de las Descripciones Analítica y Perceptual de la Entonación de una Base de Datos de Oraciones Declarativas de Foco Amplio para el Español Hablado en Buenos Aires», *Estudios de Fonética Experimental*, vol. 13, pp. 275–302. (Citado en la página 183.)

[63] Gurlekian, J., Torres, H. y Colantoni, L. (2004), «Modelos de Entonación Analítico y Fonético-Fonológico Aplicados a una Base de Datos del Español de Buenos Aires», *Estudios de Fonética Experimental*, vol. 3, pp. 275–302. (Citado en la página 51.)

[64] Gurlekian, J., Colantoni, L., Torres, H., Rincón, A., Moreno, A. y Mariño, J. (2001), «Database for an Automatic Speech Recognition System for Argentine Spanish», in «Proceedings of the IRCS Workshop on Linguistic Databases», pp. 92–98, Filadelfia, Estados Unidos. (Citado en las páginas 170 y 191.)

[65] Gurlekian, J., Rodriguez, H., Colantoni, L. y Torres, H. (2001), «Development of a Prosodic Database for an Argentine Spanish Text to Speech System», in Bird, B. y Liberman, editores, «Proceedings of the IRCS Workshop on Linguistic Databases», pp. 99–104, SIAM, University of Pennsylvania, Philadelphia, USA. (Citado en las páginas 51 y 139.)

[66] Gurlekian, J., Evin, D., Mixdorff, H., Torres, H. y Pfitzinger, H. (2010), «Accent Command Model Parameter Alignment in Argentine Spanish Absolute Interrogatives», in «21 Konferenz Elektronische Sprachsignalverarbeitung (ESSV)», pp. 77–93, Berlín, Alemania. (Citado en la página 157.)

[67] Gurlekian, J., Mixdorff, H., Evin, D., Torres, H. y Pfitzinger, H. (2010), «Alignment of Fo Model Parameters with Final and Non-Final Accents», in «Proceedings of the International Conference on Speech Prosody», pp. 1–4, Chicago, Estados Unidos. (Citado en la página 156.)

[68] Gussenhoven, C. (2004), *The Phonology of Intonation*, Cambridge University Press, Cambridge, Estados Unidos. (Citado en la página 113.)

[69] Hain, T. (2001), *Hidden Model Sequence in Automatic Speech Recognition*, Phd, Cambridge University, Engineering Departament, Cambridge, USA. (Citado en la página 65.)

[70] Hansen, J., Koeppen, B. y Netter, F. (2002), *Netter's Atlas of Human Physiology*, Icon Learning Systems, Nueva Jersey, Estados Unidos. (Citado en la página 15.)

[71] Harley, T. (2001), *Psychology of Language: From Data to Theory, 2º Ed.*, Psychology Press, Chichester, Inglaterra. (Citado en la página 24.)

[72] Hermansky, H. (1990), «Perceptual Linear Predictive (PLP) Analysis of Speech», *Foundations and Trends in Signal Processing*, vol. 87 (4), pp. 1738–1752. (Citado en la página 57.)

[73] Hermansky, H. y Morgan, N. (1994), «RASTA Processing of Speech», *IEEE Transactions on Spech and Audio Processing*, vol. 2 (4), pp. 578–589. (Citado en la página 58.)

[74] Hermes, D. J. (1988), «Measurement of Pitch by Subharmonic Summation», *Journal of the Acoustical Society of America*, vol. 83 (1), pp. 257–264. (Citado en la página 131.)

[75] Hermes, D. J. (1993), «Pitch Analysis», in Cooke, M., Beet, S. y Crawford, M., editores, «Visual Representation of Speech Signals», cap. 1, pp. 3–25, John Wiley and Sons, Chichester, Inglaterra. (Citado en la página 131.)

[76] Hirose, K. y Minematsu, N. (2004), «Use of Prosodic Features for Speech Recognition», in «Proceedings of the INTERSPEECH», pp. 1445–1448, Isla de Jeju, Corea. (Citado en las páginas 49 y 50.)

[77] Hirst, D., Di Cristo, A. y Espesser, R. (2000), «Levels of Representation and Levels of Analysis for the Description of Intonation Systems», in Gösta, B. y Horne, M., editores, «Prosody : Theory and Experiment», cap. 4, pp. 51–87, Kluwer Academic Press, Dordrecht, Holanda. (Citado en las páginas 121 y 126.)

[78] Hockett, C. (1963), «The Problem of Universals in Language», in Greenberg, J. H., editor, «Universals of Language», cap. 000, pp. 000–000, The MIT Press, Cambridge, Estados Unidos. (Citado en la página 120.)

[79] Holmes, W. y Holmes, J. (2001), *Speech Synthesis and Recognition, 2º Ed.*, Taylor and Francis, Londres, Inglaterra. (Citado en las páginas 59 y 60.)

[80] Holter, T. y Svendsen, T. (1999), «Maximum Likelihood Modelling of Pronunciation Variation», *Speech Communication*, vol. 29 (2-4), pp. 177–191. (Citado en la página 65.)

[81] Hualde, J. I. (2003), «El Modelo Métrico y Autosegmental», in Prieto, P., editor, «Teorías de la Entonación», pp. 155–184, Editorial Ariel, Barcelona, España. (Citado en las páginas 117 y 119.)

[82] Hualde, J. I. (2007), «Stress Removal and Stress Addition in Spanish», *Journal of Portuguese Linguistics*, vol. 2 (1), pp. 59–89. (Citado en la página 117.)

[83] Hualde, J. I., Olarrea, A., Escobar, A. M. y Travis, C. E. (2010), *Introduccion a la Linguistica Hispanica, 2º Ed.*, Cambridge University Press, Cambridge, Inglaterra. (Citado en la página 172.)

[84] Jain, A. K., Murty, N. M. y Flynn, P. J. (1999), «Data Clustering: A Review», *ACM Computing Surveys*, vol. 31 (3), pp. 264–323. (Citado en la página 144.)

[85] Jelinek, M. y Mercer, R. L. (1980), «Interpolated Estimation of Markov Source Parameters From Sparse Data», in Gelsema, E. y L.N., K., editores, «Pattern Recognition in Practice», pp. 381–397, North Holland Pub. Co., Amsterdam, Holanda. (Citado en la página 68.)

[86] Johnson, M. (2000), *Incorporating Prosodic Information and Language Structure into Speech Recognition System*, Phd, Purdue University. (Citado en las páginas 47 y 51.)

[87] Juang, B. H., Levinson, S. E. y Sondhi, M. M. (1986), «Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains», *IEEE Transactions on Information Theory*, vol. 32 (2), pp. 307–309. (Citado en la página 89.)

[88] Jun, S.-A. (2005), «Prosodic Typology», in Jun, S.-A., editor, «Prosodic Typology: The Phonology of Intonation and Phrasing», pp. 430–458, Oxford University Press, Oxford, Inglaterra. (Citado en la página 117.)

[89] Junqua, J. C. (1993), «The Lombard Reflex and its Role on Human Listeners and Automatic Speech Recognizers», *Journal of the Acoustic Society of America*, vol. 93 (1), pp. 510–524. (Citado en la página 28.)

[90] Kessens, J. M., Cucchiarini, C. y Strik, H. (2003), «A Data-Driven Method for Modeling Pronunciation Variation», *Speech Communication*, vol. 40 (4), pp. 517–534. (Citado en la página 65.)

[91] King, S. A. (1998), *Using Information Above the World Level for Automatic Speech Recognition*, Phd, University of Edinburgh, Inglaterra. (Citado en la página 140.)

[92] Klatt, D. H. (1979), «Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access», *Journal of Phonetics*, vol. 77 (2), pp. 279–312. (Citado en la página 24.)

[93] Klatt, D. H. (1989), «Lexical Representation and Process», in W.D.Marslen-Wilson, editor, «Review of Selected Models of Speech Perception», pp. 169–226, MIT Press, Cambridge, Estados Unidos. (Citado en la página 24.)

[94] Kneser, R. y Ney, H. (1995), «Improved Backing-off for M-Gram Language Modeling», in «Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing», pp. 181–184, Detroit, Estados Unidos. (Citado en la página 192.)

[95] Kohonen, T. (2001), *Self-Organizing Maps*, Springer, Berlín Alemania. (Citado en las páginas 140 y 144.)

[96] Kompe, R. (1997), *Prosody in Speech Understanding Systems*, Springer-Verlag, Berlín, Alemania. (Citado en la página 97.)

[97] Ladd, R. D. (1986), «Intonational Phrasing: The Case for Recursive Prosodic Structure», *Phonology Yearbook*, vol. 3 (2), pp. 311–340. (Citado en la página 113.)

[98] Ladd, R. D. (1996), *Intonational Phonology*, Cambridge University Press, Cambridge, Inglaterra. (Citado en las páginas 105, 107, 117, 127 y 168.)

[99] Ladefoged, P. (2001), *A Course in Phonetics, 4º Ed.*, Heinle and Heinle, Boston, Estados Unidos. (Citado en las páginas 28 y 109.)

[100] Lea, W. (1980), «Prosodic Aids to Speech Recognition», in Lea, W., editor, «Trends in Speech Recognition», cap. 8, pp. 166–205, Prentice Hall, Nueva Jersey, Estados Unidos. (Citado en la página 47.)

[101] Lehiste, I. (1970), *Suprasegmentals*, The MIT Press, Cambridge, Estados Unidos. (Citado en la página 168.)

[102] Levelt, W. (1989), *Speaking: From Intentions to Articulation*, The MIT Press, Cambridge, Estados Unidos. (Citado en la página 115.)

[103] Levitt, H. y Rabiner, L. R. (1971), «Analysis of Fundamental Frequency Contour in Speech», *Journal of the Acoustical Society of America*, vol. 49 (2), pp. 569–582. (Citado en la página 141.)

[104] Lieberman, M. y Prince, A. (1977), «On Stress and Linguistic Rhythm», *Linguistic Inquiry*, vol. 8 (2), pp. 249–336. (Citado en la página 112.)

[105] Lieberman, P. (1960), «Some Acoustic Correlates of Word Stress in American English», *Journal of the Acoustical Society of America*, vol. 32 (4), pp. 451–454. (Citado en la página 48.)

[106] Lieberman, P., Katz, W., Jongman, A., Zimmerman, R. y Miller, M. (1985), «Measures of the Sentence Intonation of Read and Spontaneous Speech in American English», *Journal of the Acoustical Society of America*, vol. 77 (2), pp. 649–657. (Citado en la página 149.)

[107] Lippmann, R. P. (1997), «Speech Recognition by Machines and Humans», *Speech Communication*, vol. 22 (1), pp. 1–15. (Citado en las páginas 2, 3 y 51.)

[108] Llorach, E. (1999), *Gramática de la Lengua Española. Real Academia Española, Colección Nebrija y Bello*, Editorial Espasa Calpe, Madrid, España. (Citado en la página 51.)

[109] Luce, P. A., Goldinger, S. D., Auer, E. T. y Vitevitch, M. S. (2000), «Phonetic Priming, Neighborhood Activation, and PARSYN», *Perception and Psychophysics*, vol. 62 (3), pp. 615–625. (Citado en la página 24.)

[110] Marslen-Wilson, W. D. (1987), «Functional Parallelism in Spoken Word-Recognition», *Cognition*, vol. 25 (1-2), pp. 71–102. (Citado en la página 26.)

[111] Marslen-Wilson, W. D. y Welsh, A. (1978), «Processing Interactions and Lexical Access During Word Recognition in Continuous Speech», *Cognitive Psychology*, vol. 10 (1), pp. 29–63. (Citado en la página 24.)

[112] Martin, P. (1981), «Extraction de la Fréquence Fondamentale par Intercorrélation avec une Fonction Peigne», in «Actes des 12e Journées d'Etudes sur la Parole», pp. 221–232, Montreal, Canadá. (Citado en la página 134.)

[113] Martin, P. (1982), «Comparison of Pitch by Cepstrum and Spectral Comb Analysis», in «Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)», pp. 180–183, París, Francia. (Citado en la página 131.)

[114] McClelland, J. L. y Elman, J. L. (1986), «The TRACE Model of Speech Perception», *Cognitive Psychology*, vol. 18 (1), pp. 1–86. (Citado en las páginas 24 y 25.)

[115] McNeilage, P. y Davis, B. (2001), «Motor Mechanism in Speech Ontogeny: Phylogenetic, Neurobiological and Linguistic Implications», *Current Opinions in Neurobiology*, vol. 11 (000), pp. 696–700. (Citado en la página 115.)

[116] McQueen, J. M., Cutler, A., Briscoe, T. y Norris, D. G. (1995), «Models of Continuous Speech Recognition and the Contents of the Vocabulary», *Language and Cognitive Processes*, vol. 10 (3-4), pp. 309–331. (Citado en la página 27.)

[117] Mehler, J. (1981), «The Role of Syllables in Speech Processing: Infant and Adult Data», *Philosophical Transactions of the Royal Society*, vol. 295 (000), pp. 333–352. (Citado en la página 115.)

[118] Mercer, R. y Cohen, P. (1987), «A Method for Eficient Storage and Rapid Application of Context-Sensitive Phonological Rules for Automatic Speech Recognition», *IBM J. Res. Develop.*, vol. 31 (1), pp. 81–90. (Citado en la página 66.)

[119] Milone, D. (2003), *Información Acentual para el Reconocimiento Automático del Habla*, Phd, Universidad de Granada, España. (Citado en la página 51.)

[120] Milone, D. y Rubio, A. (2003), «Prosodic and Accentual Information for Automatic Speech Recognition», *IEEE Transactions on Speech and Audio Processing*, vol. 11 (4), pp. 321–333. (Citado en las páginas 51, 169 y 170.)

[121] Minematsu, N. y Nakagawa, S. (1998), «Modeling of Variations in Cepstral Coefficients Caused by Fo Changes and its Application to Speech Processing», in «Proceedings of the International Conference on Spoken Language Processing (ICSLP)», vol. 3, pp. 1063–1066, Sidney, Australia. (Citado en la página 48.)

[122] Mixdorff, H. (2000), «A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters», in «Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)», vol. 3, pp. 1281–1284, Estambul, Turquía. (Citado en la página 183.)

[123] Mixdorff, H. (2002), *An Integrated Approach to Modeling German Prosody*, Phd, Fakultät Elektrotechnik und Informationstechnik, Technische Universität Dresden. (Citado en la página 96.)

[124] Moore, R. y Cutler, A. (2001), «Constraints on Theories of Human vs. Machine Recognition of Speech», in Smits, R., Kingston, J., T.M., N. y Zondervan, R., editores, «Proceedings of the Workshop on Speech Recognition as Pattern Classification», pp. 145–150, Nijmegen, Holanda. (Citado en la página 3.)

[125] Moreno, A., Höge, H., Köhler, J. y Mariño, J. B. (1998), «SpeechDat Across Latin America Project SALA», in «Proceedings of the First International Conference on Language Resources and Evaluation (LREC)», pp. 367–370, Granada, España. (Citado en la página 170.)

[126] Navarro Tomás, T. (1964), «La Medida de la Intensidad», Boletín del Instituto de Filología de la Universidad de Chile. (Citado en la página 116.)

[127] Nespor, M. y Vogel, I. (1986), *Prosodic Phonology*, Kluwer Academic Press, Dordrecht, Holanda. (Citado en las páginas 112, 113, 115, 116, 118, 119 y 120.)

[128] Nibert, H. (2000), *Phonetic and Phonological Evidence for Intermediate Phrasing in Spanish Intonation*, Phd, University of Illinois at Urbana-Champaign, Urbana-Champaign, Estados Unidos. (Citado en la página 117.)

[129] Nooteboom, S. (1999), «The Prosody of Speech: Melody and Rhythm», in Hardcastle, W. J. y Laver, J., editores, «The Handbook of Phonetic Sciences», pp. 169–226, Wiley-Blackwell, Oxford, Inglaterra. (Citado en las páginas 100 y 104.)

[130] Norris, D. (1994), «Shortlist: A Connectionist Model of Continuous Speech Recognition», *Cognition*, vol. 52, pp. 189–234. (Citado en las páginas 24 y 26.)

[131] Nöth, E., Batliner, A., Kiessling, A. y Kompe, R. (2000), «VERBMOBIL: The Use of Prosody in the Linguistic Components of a Speech Understanding System», *IEEE Transactions on Speech and Audio Processing*, vol. 8 (5), pp. 519–532. (Citado en la página 50.)

[132] Oppenheim, A. V. y Schafer, R. W. (1968), «Homomorphic Analysis of Speech», *IEEE Transactions on Audio and Electroacoustics*, vol. 16 (2), pp. 221–226. (Citado en la página 30.)

[133] O'Rourke, E. (2005), *Intonation and Language Contact: A Case Study of Two Varieties of Peruvian Spanish*, Phd, University of Illinois at Urbana-Champaign, Urbana-Champaign, Estados Unidos. (Citado en la página 117.)

[134] Ortega-Llebaria, M. y Prieto, P. (2007), «Disentangling Stress from Accent in Spanish: Production Patterns of the Stress Contrast in De-Accented Syllables», in Prieto, P., Mascaró, J. y Solé, M.-J., editores, «Segmental and Prosodic Issues in Romance Phonology. Current Issues in Linguistic Theory», cap. 18, pp. 155–176, John Benjamins, Amsterdam, Holanda. (Citado en la página 117.)

[135] Pieraccini, R. (1986), «Lexical Stress and Speech Recognition», in «Proceedings of the 112th Meeting of the Acoustic Society of America», pp. 103–107, Anaheim, Estados Unidos. (Citado en la página 45.)

[136] Pike, K. L. (1945), *The intonation of American English*, Ann Arbor : University of Michigan Press, Michigan, Estados Unidos. (Citado en la página 108.)

[137] Polzin, T. y Waibel, A. (1998), «Pronunciation Variations in Emotional Speech», in «Proceedings of the European Speech Communication Association (ESCA) Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition», pp. 103–107, Kerkrade, Holanda. (Citado en la página 28.)

[138] Prieto, P. (2005), «Syntactic and Eurhythmic Constraints on Phrasing Decisions in Catalan», *Studia Lingüística*, vol. 59 (2-3), pp. 194–202. (Citado en la página 117.)

[139] Prieto, P. (2006), «Phonological Phrasing in Spanish», in Colina, S. y F., M.-G., editores, «Optimality-Theoretic Advances in Spanish Phonology», pp. 39–60, John Benjamins, Amsterdam, Holanda. (Citado en la página 119.)

[140] Prince, A. y Smolensky, P. (2004), *Optimality Theory Constraint Interaction in Generative Grammar*, Blackwell Publishing, Oxford, Inglaterra. (Citado en la página 113.)

[141] Quilis, A. (1993), *Tratado de Fonología y Fonética Españolas*, Editorial Gredos, Madrid, España. (Citado en las páginas 51, 99, 106, 108, 116, 117 y 173.)

[142] Rabiner, L. y Juang, B.-H. (1993), *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, Estados Unidos. (Citado en la página 61.)

[143] Ramus, F. y Mehler, J. (1999), «Language Identification with Suprasegmental Cues: A Study Based on Speech Resynthesis», *Journal of the Acoustical Society of America*, vol. 105 (1), pp. 512–521. (Citado en la página 48.)

[144] Ramus, F., Nespor, M. y Mehler, J. (1999), «Correlates of Linguistic Rhythm in the Speech Signal», *Cognition*, vol. 73 (3), pp. 265–292. (Citado en la página 109.)

[145] Rao, R. (2007), «On the Phonological Phrasing Patterns in the Spanish of Lima, Per£», *Southwest Journal of Linguistics*, vol. 26, pp. 81–111. (Citado en la página 119.)

[146] Ravinshakar, M. (1996), *Efficient Algorithms for Speech Recognition*, Phd, School of Computer Science, Computer Science Division, Carnegie Mellon University. (Citado en la página 176.)

[147] Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C. y Zavaliagkos, G. (1999), «Stochastic Pronunciation Modelling from Hand-Labelled Phonetic Corpora», *Speech Communication*, vol. 29 (2-4), pp. 209–224. (Citado en la página 65.)

[148] Rosenberg, A. (2009), *Automatic Detection and Classification of Prosodic Events*, Phd, Columbia University, Nueva York, Estados Unidos. (Citado en la página 151.)

[149] Sakurai, A., Hirose, K. y Minematsu, N. (2003), «Data-driven Generation of Fo Contours Using a Superpositional Model», *Speech Communication*, vol. 40 (4), pp. 535–549. (Citado en la página 186.)

[150] Scharenborg, O. (2007), «Reaching Over the Gap: A Review of Efforts to Link Human and Automatic Speech Recognition Research», *Speech Communication*, vol. 49 (5), pp. 336–347. (Citado en la página 3.)

[151] Schiel, F., Kipp, A. y Tillmann, H. (1998), «Statistical Modelling of Pronunciation: It's not the Model, it's the Data», in Strik, H., Kessens, J. y Wester, M., editores, «Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition», pp. 131–136, Rolduc, Kerkrade. (Citado en la página 65.)

[152] Selkirk, E. (1993), *Phonology and Syntax: The Relation Between Sound and Structure*, The MIT Press, Cambridge, Estados Unidos. (Citado en las páginas 113, 115 y 120.)

[153] Selkirk, E. O. (2000), «The Interaction of Constraints on Prosodic Phrasing», in Horne, M., editor, «Prosody: Theory and Experiment», pp. 231–262, Kluwer Academic Publishing, Dordrecht, Holanda. (Citado en la página 119.)

[154] Seppi, D., Falavigna, D., Stemmer, G. y Gretter, R. (2007), «Word Duration Modeling for Word Graph Rescoring in LVCSR», in «Proceedings of the INTERSPEECH», pp. 1805–1808, Antwerp, Bélgica. (Citado en la página 181.)

[155] Shattuck-Hufnagel, S. y Turk, A. E. (1996), «A Prosody Tutorial for Investigators of Auditory Sentence Processing», *Journal of Psycholinguistic Research*, vol. 25 (2), pp. 193–247. (Citado en la página 119.)

[156] Shriberg, E. y Stolcke, A. (2004), «Direct Modeling of Prosody: An Overview of Applications to Automatic Speech Processing», in «Proceedings of the International Conference on Speech Prosody», pp. 575–582, Nara, Japón. (Citado en la página 51.)

[157] Shriberg, E. y Stolcke, A. (2004), «Prosody Modeling for Automatic Speech Recognition and Understanding», in Johnson, M., Khudanpur, S., Ostendorf, M. y R., R., editores, «Mathematical Foundations of Speech and Language Processing. Volume 138 in IMA Volumes in Mathematics and its Applications», pp. 105–114, Springer-Verlag, Nueva York, Estados Unidos. (Citado en la página 51.)

[158] Shukla, M. (2006), *Prosodic Constraints on Statistical Strategies in Segmenting Fluent Speech*, Phd, Scuola Internazionale Superiore di Studi Anvanzati, Trieste, Italy. (Citado en las páginas 112 y 114.)

[159] Silbernagl, S. y Despopoulus, A. (2009), *Color Athlas of Physiology, 6º Ed.*, Thieme, Ludwigsburg, Alemania. (Citado en las páginas 12, 14, 18 y 22.)

[160] Sosa, J. (1999), *La Entonación del Español: su Estructura Fónica, Variabilidad y Dialectología*, Cátedra, Madrid, España. (Citado en las páginas 51 y 117.)

[161] Stolcke, A. (2002), «SRILM - An Extensible Language Modeling Toolkit», in «Proceedings of the International Conference on Spoken Language Processing (ICSLP)», pp. 901–904, Denver, Estados Unidos. (Citado en la página 192.)

[162] Stolcke, A., Shriberg, E., Hakkani-Tür, D. y Tür, G. (1999), «Modeling the Prosody of Hidden Events for Improved Word Recognition», in «Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech)», pp. 307–310, Budapest, Hungría. (Citado en la página 51.)

[163] Stolcke, A., Bates, R., Coccaro, N., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R. y Meteer, M. (2000), «Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech», *Computational Linguistics*, vol. 26 (3), pp. 339–371. (Citado en la página 140.)

[164] Strik, H. y Cucchiarini, C. (1999), «Modeling Pronunciation Variation for ASR: A Survey of the Literature», *Speech Communication*, vol. 29 (2-4), pp. 225–246. (Citado en las páginas 33 y 65.)

[165] 't Hart, J., René, C. y Antonie, C. (1990), *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody*, Cambridge University Press, Cambridge, Inglaterra. (Citado en las páginas 103 y 125.)

[166] Talkin, D. (1995), «A Robust Algorithm for Pitch Tracking (RAPT)», in Klejin, W. B. y Paliwal, K., editores, «Speech Coding and Synthesis», cap. 14, pp. 495–518, Elsevier, Amsterdam, Holanda. (Citado en la página 129.)

[167] Tatham, M. y Morton, K. (2005), *Developments in Speech Synthesis*, John Wiley and Sons, Chichester, Inglaterra. (Citado en la página 48.)

[168] Taylor, P. (2000), «Analysis and Synthesis of Intonation Using the Tilt Model», *Journal of the Acoustical Society of America*, vol. 107 (3), pp. 1697–1714. (Citado en la página 122.)

[169] Taylor, P. (2009), *Text-to-Speech Synthesis*, Cambridge University Press, Cambridge, Inglaterra. (Citado en la página 48.)

[170] Terhardt, E. (1974), «Pitch, Consonance, and Harmony», *Journal of the Acoustical Society of America*, vol. 55 (5), pp. 1061–1069. (Citado en la página 131.)

[171] Toledo, G. A. (1988), *El ritmo en el Español*, Gredos, Madrid, España. (Citado en la página 170.)

[172] Toledo, G. A. (2007), «Fraseo en Español Peninsular y Modelo Autosegmental y Métrico», *Estudios Filológicos*, vol. 42, pp. 227–243. (Citado en la página 117.)

[173] Toledo, G. A. (2008), «Fonología de la Frase Entonativa», *Estudios Filológicos*, vol. 43, pp. 207–222. (Citado en las páginas 117 y 119.)

[174] Torres, H. (2008), *Generación Automática de la Prosodia para un Sistema de Conversión de Texto a Habla*, Phd, Facultad de Ingeniería, Universidad de Buenos Aires, Argentina. (Citado en la página 183.)

[175] Torres, H. y Gurlekian, J. (2004), «Automatic Determination of Phrase Breaks for Argentine Spanish», in «Proceedings of the International Conference on Speech Prosody», pp. 553–556, Nara, Japón. (Citado en la página 51.)

[176] Torres, H. y Gurlekian, J. (2009), «Parameter Estimation and Prediction from Text for a Superpositional Intonation Model», in «Proceedings of the 20 Konferenz Elektronische Sprachsignalverarbeitung (ESSV)», pp. 238–247, TUDpress Verlag der Wissenschaften, Dresden, Alemania. (Citado en las páginas 184 y 186.)

[177] Truckenbrodt, H. (1999), «On the Relation between Syntactic Phrases and Phonological Phrases», *Linguistic Inquiry*, vol. 30, pp. 219–255. (Citado en la página 113.)

[178] Truckenbrodt, H. (2007), «The Syntax Phonology Interface», in de Lacy, P., editor, «The Cambridge Handbook of Phonology», cap. 18, pp. 435–456, Cambridge University Press, Cambridge, Estados Unidos. (Citado en las páginas 117 y 119.)

[179] Umeda, N. y Wedmore, T. (1994), «A Rythm Theory for Spontaneous Speech: the Role of vowel amplitude in the Rhythmic Structure», in «Proceedings of the International Conference on Spoken Language Processing (ICSLP)», pp. 1095–1098, Yokohama, Japón. (Citado en la página 109.)

[180] Univaso, P., Gurlekian, J. y Evin, D. (2009), «Reconocedor de Habla Continua Independiente del Contexto para el Español de Argentina», *Revista Clepsidra*, vol. 8, pp. 13–22. (Citado en la página 169.)

[181] Vaissière, J. (1988), «The Use of Prosodic Parameters in Automatic Speech Recognition», in Niemann, H., editor, «Recent Advances in Speech Understanding and Dialog Systems», pp. 71–99, Springer-Verlag, Berlín, Alemania. (Citado en la página 39.)

[182] van den Heuvel, H., van Kuijk, D. y Boves, L. (2003), «Modeling Lexical Stress In Continuous Speech Recognition For Dutch», *Speech Communication*, vol. 40 (3), pp. 335–350. (Citado en la página 170.)

[183] Varga, A. y Steeneken, H. (1993), «Assessment for Automatic Speech Recognition, NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems», *Speech Communication*, vol. 12 (3), pp. 247–251. (Citado en la página 191.)

[184] Vergyri, D., Stolcke, A., Gadde, V., Ferrer, L. y Shriberg, E. (2003), «Prosodic Knowledge Sources For Automatic Speech Recognition», in «Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)», pp. 208–211, Hong Kong, China. (Citado en la página 181.)

[185] Vicsi, K. y Szaszak, G. (2010), «Using Prosody to Improve Automatic Speech Recognition», *Speech Communication*, vol. 52 (5), pp. 413–426. (Citado en la página 181.)

[186] Vidal de Battini, B. (1964), *El Español de Argentina*, Consejo Nacional de Educación, Buenos Aires, Argentina. (Citado en la página 170.)

[187] Vigário, M. (2003), *The Prosodic Word in European Portuguese*, Mouton de Gruyter, Berlín, Alemania. (Citado en la página 113.)

[188] Waibel, A. y Lee, K. (1990), *Readings in Speech Recognition*, Morgan Kaufmann, San Mateo, Estados Unidos. (Citado en la página 34.)

[189] Wang, C. y Seneff, S. (2001), «Lexical Stress Modeling for Improved Speech Recognition of Spontaneous Telephone Speech in the Jupiter Domain», in «Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)», pp. 2761–2765, Aalborg, Dinamarca. (Citado en la página 169.)

[190] Watson, D. y Gibson, E. (2004), «The Relationship between Intonational Phrasing and Syntactic Structure in Language Production.», *Language and Cognitive Processes*, vol. 19 (6), pp. 713–755. (Citado en la página 120.)

[191] Weintraub, M., Taussig, K., Hunicke-Smith, K. y Snodgrass, A. (1996), «Effect of Speaking Style on LVCSR Performance», in «Proceedings of the International Conference on Spoken Language Processing (ICSLP)», pp. S16–S19, Filadelfia, Estados Unidos. (Citado en la página 28.)

[192] Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A. y Kollmeier, B. (2005), «Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines», in «Proceedings of the INTERSPEECH 2005», pp. 1273–1276, Lisboa, Portugal. (Citado en la página 3.)

[193] Wester, M. (2002), *Pronunciation Variation Modeling for Dutch Automatic Speech Recognition*, Phd, University of Nijmegen, Holanda. (Citado en la página 51.)

[194] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtech, V. y Wooland, P. (2006), *The HTK Book*, Cambridge University Press, Cambridge, Inglaterra. (Citado en las páginas 173 y 193.)

[195] Zhang, J. y Hirose, K. (2004), «Tone Nucleus Modeling for Chinese Lexical Tone Recognition», *Speech Communication*, vol. 42 (3-4), pp. 447–466. (Citado en la página 49.)

ÍNDICE ALFABÉTICO