Sistema de análisis prosódico y reconocimiento automático del habla

por

Albornoz, Enrique Marcelo

Director: Dr. Diego H. Milone

PROYECTO FINAL DE CARRERA

Ingeniería Informática

Grupo de investigación en señales e inteligencia computacional

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

Índice general

Prefacio **V**

1. Introducción **1**
1.1. Reconocimiento Automático del Habla 1
1.1.1. El Habla 2
1.1.2. Organización estructural del habla 6
1.1.3. El RAH mediante modelos ocultos de Markov 8
1.1.4. Aplicaciones y dificultades 10
1.2. Prosodia y RAH 11
1.3. Objetivos 12

2. RAH con información prosódica **15**
2.1. Clasificación prosódica basada en histogramas . . . 16
2.1.1. Segmentación 16
2.1.2. Extracción de rasgos prosódicos 17
2.1.3. Generación de histogramas 18
2.2. Incorporación de la prosodia al RAH 19
2.2.1. Redes de palabras 19
2.2.2. Obtención de la probabilidad de penalización prosódica 23
2.2.3. Modificación de la probabilidad en la red de palabras 24

3. Análisis y Diseño del Software **25**
3.1. Requerimientos 25
3.1.1. Equipo de trabajo 26
3.1.2. Hechos 26
3.1.3. Modelos 26
3.1.4. Clasificación de los requerimientos 33
3.1.5. Racionalización y Prioridades 36
3.1.6. Integración y validación 36
3.2. Diseño 37
3.2.1. Análisis y Diseño con el DC 37
3.2.2. Desarrollo de DC durante el análisis y diseño 38

3.2.3. Diagrama de Clases 38

4. Resultados y discusión **43**
4.1. Clasificación de estructuras prosódicas 44
4.2. Resultados de RAH 50

5. Conclusiones y trabajos futuros **57**

A. Orientación a Objetos **63**
A.1. Fundamentos del modelado OO 63
A.1.1. Objetos 64
A.1.2. Identidad 64
A.1.3. Estado 65
A.1.4. Comportamiento 65
A.1.5. Persistencia 65
A.1.6. Comunicación 66
A.1.7. Mensaje 66
A.1.8. Mensaje y Estímulo 66
A.1.9. Operaciones y Métodos 66

B. UML **67**
B.1. Modelos y Diagramas 67
B.2. Diagramas de UML 68
B.2.1. Diagramas de Casos de Uso 68
B.2.2. Escenarios 69
B.2.3. Diagramas de Interacción 70
B.2.4. Diagramas de Secuencia 70
B.2.5. Diagrama de Clases 71

Índice de figuras

1.1. Etapas básicas de un sistema de RAH 2
1.2. Espectro de la vocal /a/ con $F_0 \approx 250$ Hz. 4
1.3. Modelo compuesto. 9

2.1. Clases prosódicas para la palabra *dime* en el rasgo prosódico media de energía 20
2.2. Clases prosódicas para la palabra *longitud* en el rasgo prosódico mínimo de energía 20
2.3. Incorporación prosódica 21
2.4. Modelo de Lenguaje bigramática 22
2.5. Red de palabras de hipótesis 23

3.1. Caso de Uso general 27
3.2. Caso de Uso detallado 28
3.3. Diagrama de Secuencia 32
3.4. Diagrama de la clase Manejador_palabras 39
3.5. Diagrama de la clase Manejador_ToFy 39
3.6. Diagrama de la clase núcleo reconocedor 40
3.7. Diagrama de la clase Red_de_palabras 41
3.8. Diagrama de la clase M_Htk 41
3.9. Diagrama de clases del sistema 42

4.1. Clases prosódicas para la palabra *valenciana* en el rasgo prosódico media de F_0 45
4.2. Clases prosódicas para la palabra *desemboca* en el rasgo prosódico máximo de energía 46
4.3. Clases prosódicas para la palabra *metros* en el rasgo prosódico media de F_0 46
4.4. Clases prosódicas para la palabra *superior* en el rasgo prosódico media de energía 47
4.5. Clases prosódicas para la palabra *cúbicos* en el rasgo prosódico mínimo de energía 48
4.6. Clases prosódicas para la palabra *comunidad* en el rasgo prosódico mínimo de energía 48

4.7. Clases prosódicas para la palabra *valencia* en el rasgo prosódico media de F_0 49

B.1. Caso de Uso para manejo de un televisor 69
B.2. Ejemplo de un diagrama de secuencia 71
B.3. Diagrama de la clase Persona 72

Prefacio

Este proyecto pertenece al área de reconocimiento automático del habla. La aplicaciones de reconocimiento automático del habla presentan un problema multidisciplinar, relacionado con: procesamiento de señales, acústica, teoría de la comunicación y de la información, estadística, matemática, lingüística, fisiología, reconocimiento de formas e inteligencia artificial, etc. Las aplicaciones en las que el reconocimiento automático del habla tiene incumbencia son muchas, se podrían nombrar: ayuda a discapacitados, dictado automático, transcripción y traducción voz a voz, operaciones de máquinas a través de la voz, control manos-libres en aplicaciones industriales, educación, sustituto de contraseñas en el acceso a equipos informáticos y de PIN en acceso a cajeros automáticos, etc. En las últimas décadas se han realizado aportes muy importantes en muchos niveles de los sistemas de reconocimiento automático del habla, aunque la prosodia no está completamente integrada en estos sistemas. El proyecto consiste en el desarrollo de un sistema que permita el análisis de señales de voz con sus rasgos prosódicos y brinde la posibilidad de incorporar este análisis a un sistema de reconocimiento automático del habla.

Con el primer Capítulo se pretende dar una pequeña revisión de los conceptos más importantes relacionados con el reconocimiento automático del habla. También, se incluyen dos apéndices que dan una pequeña reseña de los temas tratados en el Capítulo 3, sobre el diseño del softwate.

En el Capítulo 2 se propone un método para caracterizar a las palabras según sus estructuras prosódicas y, aplicando modelos de lenguaje variantes en el tiempo, se utiliza esta información para desambiguar hipótesis en el proceso de reconocimiento mediante modelos ocultos de Markov.

Los sistemas de reconocimiento automático del habla con prosodia se encuentran solamente como resultado de investigaciones científicas, por esta razón, en el Capítulo 3 se analiza, diseña e implementa un sistema orientado a objetos donde se permite analizar las señales de voz y sus rasgos prosódicos

y permite incorporar el análisis prosódico a un sistema de reconocimiento automático del habla ya entrenado.

El Capítulo 4 se presenta dividido en dos partes principales, en una de ellas se ven resultados y discusiones acerca de la aplicación del método de clasificación por histogramas y en la segunda parte se analizan los resultados obtenidos de incorporar ésta información al sistema de reconocimiento automático del habla.

En el Capítulo 5 se exponen las conclusiones respecto de los aportes del método y de su implementación en un sistema de reconocimiento automático del habla, así como también los trabajos futuros en torno a estas ideas.

1

Introducción

Se ha avanzado mucho en el reconocimiento automático del habla (RAH) y se ha incorporado información importante a distintos niveles de análisis del habla. Sin embargo, los rasgos prosódicos no son generalmente ejes en estos análisis y su incorporación al RAH aún es incipiente. Las investigaciones actuales [1, 2, 3] tienen como fin encontrar, dentro de estas manifestaciones prosódicas físicas, la información necesaria para mejorar el rendimiento de los sistemas de RAH.

1.1. Reconocimiento Automático del Habla

El RAH es una disciplina que se encarga de la concepción y realización de sistemas automáticos que convierten las señales acústicas procedentes de un locutor humano en (secuencias de) categorías lingüísticas de un universo dado [4]. Se puede decir que un sistema de RAH está compuesto de tres etapas básicas que se pueden observar en la Figura 1.1.

Para esta ciencia el problema es multidisciplinar y está relacionado con: procesamiento de señales, acústica, teoría de la comunicación y de la información, estadística, matemática, lingüística, fisiología, informática (especialmente reconocimiento de formas e inteligencia artificial), etc.

Cabe destacar que el principal problema con el que se enfrenta es el modelado de la variabilidad temporal de la señal de voz. Para ésto, la técnica más utilizada y con mejores resultados en RAH, es la de modelos ocultos de Markov [5].

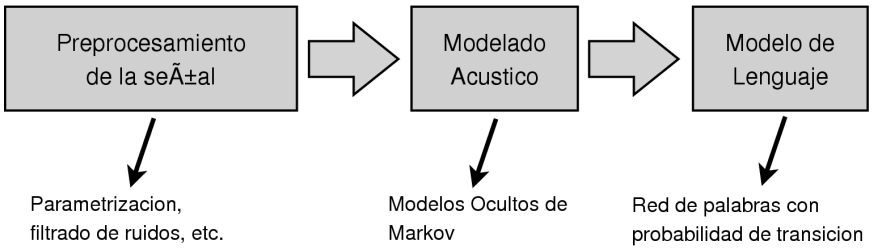


FIGURA 1.1: ETAPAS BÁSICAS DE UN SISTEMA DE RAH

1.1.1. EL HABLA

El ser humano logra el proceso de comunicación más avanzado entre los seres vivos por medio del habla, éste proceso es sencillo de explicar, aunque complejo de modelar computacionalmente. Debe haber un locutor y un oyente, el primero genera un mensaje y lo transmite por medio de señales acústicas (ondas sonoras) mientras que el otro las interpreta para entender el mensaje original.

ONDAS SONORAS

Las ondas sonoras se originan por el movimiento vibratorio de un cuerpo, en este caso las cuerdas vocales. La suma de varias ondas simples genera una onda compuesta y las armónicas de una onda de frecuencia f, son ondas con frecuencias múltiplos de f. El análisis de Fourier [6] trata de describir una onda compleja como suma de ondas simples, proporciona también el espectro de la onda compleja donde se muestra la amplitud de las ondas simples para cada armónico.

PRODUCCIÓN DE VOZ

La señal de voz es producida por el aparato fonador y se transmite mediante ondas de presión propagadas por el aire [7]. Este aparato consta de tres elementos principales:

1. **Un generador de energía: los pulmones.** Produce la diferencia de presión que crea el flujo de aire que activa la siguiente etapa.

2. **Un sistema vibrante: laringe y cuerdas vocales.** Cuando el flujo de aire pasa por la laringe hace vibrar las cuerdas vocales, creando una onda rica en armónicos que es modulada en la siguiente etapa. La frecuencia de vibración se denomina frecuencia fundamental (F_0), su sensación auditiva es el tono de la voz o entonación, la que varía con el hablante según sea mujer o varón, adulto o niño, etc.

3. **Una cavidad resonante: tracto vocal y cavidades.** La morfología del tracto vocal, la faringe, la boca y la nariz; es deformable por elementos articulatorios (lengua, labios, mandíbulas, velo del paladar), que determinan finalmente que frecuencias se atenúan y cuales se realzan.

Si las cuerdas vocales se encuentran relajadas y separadas el aire pasará por ellas sin provocar ningún sonido, pero si están tensas modularán el aire en pulsos (llamados pulsos glóticos) cuya frecuencia dependerá fundamentalmente de esta tensión y del tamaño del órgano. En el hombre, la frecuencia de vibración de las cuerdas vocales está entre 100 y 170 Hz, en las mujeres suele estar entre 180 y 280 Hz y en los niños puede superar los 300 Hz. Los valores de esta son los responsables de la F_0 producida al hablar [8].

De acuerdo con la zona en donde se genera el sonido se puede hacer una primera división de los sonidos de la voz: *Sonoros*, cuando en la generación intervienen las cuerdas vocales; y *Sordos* cuando el generador está en otra parte del tracto vocal (la nariz, la boca, etc.).

El sonido generado posee componentes frecuenciales que ocupan, en su conjunto, toda la banda del espectro sonoro del habla. Cuando este pasa por el tracto vocal recibe muchas modificaciones que dependen de la morfología del tracto. Si se grafica su espectro de frecuencias, se podrán ver algunos picos de resonancia y otros valles donde hubo predominantemente atenuaciones. Podemos imaginar al tracto vocal como un conjunto de resonadores que refuerzan o atenúan ciertas frecuencias según sea el sonido que se desea pronunciar. Si se excita al tracto vocal con los pulsos glóticos, las bandas de frecuencia que coincidan con la frecuencia de resonancia de alguno de sus resonadores no serán atenuadas. Como resultado, en la salida predominarán algunas ondas sinusoidales amortiguadas que se verán como picos en el espectro de frecuencias. Este es el concepto de formante, que puede definirse más precisamente como: *energía que se concentra en una banda de frecuencia por efecto de un resonador del tracto vocal.* Las formantes se notan con una F y un número que indica su orden de aparición desde las frecuencias más bajas. Casi siempre son distinguibles varias formantes en los sonidos vocálicos y ciertos sonidos consonánticos que conservan las formantes de su contexto vocálico.

En la Figura 1.2 se puede observar un análisis en frecuencia de la vocal /a/ con una aproximación del espectro en la que se pueden apreciar claramente las cuatro primeras formantes. La importancia de las formantes radica en que su posición identifica a los sonidos vocálicos; de hecho, si realizáramos un análisis para otra F_0, se podrían observar que las cuatro primeras formantes quedan prácticamente en el mismo lugar que estaban.

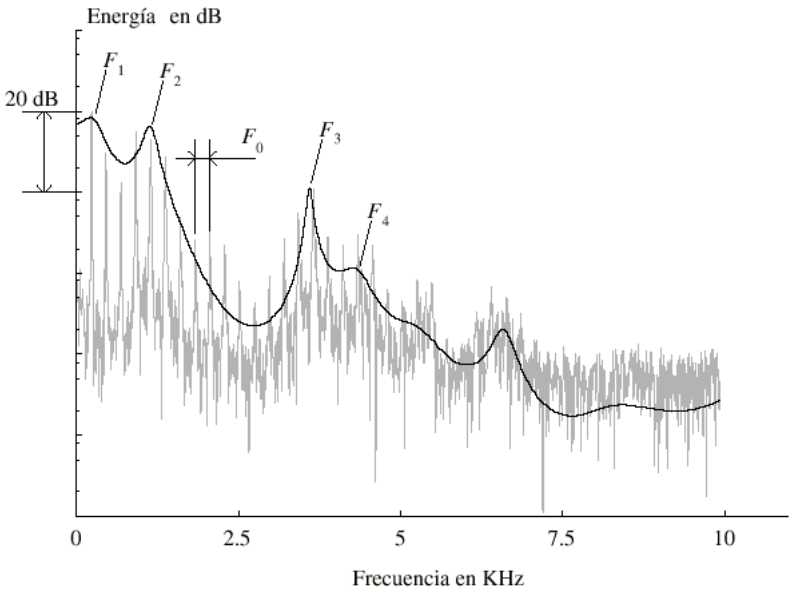


FIGURA 1.2: ESPECTRO DE LA VOCAL /A/ CON $F_0 \approx 250$ HZ. (ADAPTADA CON PERMISO DE [9])

ANÁLISIS DE LA SEÑAL DE VOZ

El tracto vocal presenta una variabilidad temporal y por ello la señal de voz es una señal no estacionaria. Debido a que la mayoría de los sistemas de análisis de voz y sistemas de RAH son implementados en computadoras que trabajan con datos digitales, se debe convertir la señal a través el proceso conocido como conversión analógico-digital de la señal [6].

Luego de digitalizarla y dado que no tendría sentido analizarla muestra a muestra y tampoco en períodos de varios segundos, se hace valer la hipótesis de estacionariedad por tramos de la señal en relación a la velocidad de variación de la morfología del tracto y ésta se la analiza en tramos de 10 a 30 milisegundos [10].

Se supone que la señal continua de voz, $v(t)$, es sometida a un muestreo uniforme con período T_v, entonces se representa la señal como $v(m)$ con $0 < m \leq N_v$. Aplicando la ventana de análisis obtenemos los tramos de voz:

$$v(t;n) = w(n;N_w)v(tN_d + n); \quad 0 < n \leq N_w \tag{1.1}$$

donde la señal $w(n;N_w)$ es una ventana de análisis definida para $0 < n \leq N_w$ y el paso del análisis está definido por $T_d = N_d T_v$.

Hay diversos tipos de ventanas, siendo la ventana cuadrada la más simple y la menos recomendable porque produce efectos indeseados. Para evitarlos se hace uso de ventanas de análisis que no son más que funciones con caracte-

rísticas especiales, que atenúan los efectos de rizado en el dominio frecuencial y minimizan las consecuencias que genera la aplicación de la ventana cuadrada.

Debido al costo computacional y a un compromiso subyacente entre la resolución frecuencial y la distorsión armónica se utiliza masivamente la ventana de Hamming. En (1.2) se expresa la definición de la ventana de Hamming, la más utilizada en este tipo de análisis. Una versión ampliada del uso de ventanas puede hallarse en [11].

$$w_H[n] = \frac{27}{50} - \frac{23}{50} cos(2\pi n/N) \tag{1.2}$$

En estos tramos de voz (1.1) ya se pueden distinguir los sonidos del habla, silencios y ruidos, por medio de herramientas de la física acústica.

Transformaciones de dominio

Las señales de voz no brindan información relevante en el dominio temporal, por ésto es necesario transformarlas (cambiar el punto de observación) y así poder extraer las características útiles al análisis de voz. Dentro de las transformaciones más utilizadas se encuentran los coeficientes cepstrales (CC) y los coeficientes ceptrales en escala de mel (CCEM).

La transformación de una señal en su cepstrum es una transformación homomórfica, y el concepto de cepstrum es una parte fundamental de la teoría de sistemas homomórficos para el procesamiento de señales que han sido convolucionadas [7, 12, 13]. Los CC están basados en la transformada discreta de fourier (TDF) y el cepstrum real de $v(m)$ es:

$$c(m) = \mathsf{TF}^{-1}\left\{\mathsf{log}|\mathsf{TF}\left\{v(m)\right\}|\right\} \tag{1.3}$$

Si se extiende la definición al análisis por tramos, se reemplaza la TF^{-1} por su definición, se puede observar que el argumento de la TF^{-1} es una secuencia real y par [7], por lo que la definición de cepstrum queda:

$$c(t;k) = \frac{1}{N_u} \sum_{z=1}^{N_u} \mathsf{log}|u(t;z)|\mathsf{cos}(2\pi/N_u)(z-1)(k-1) \tag{1.4}$$

Siendo $u(t;z) = TF(z)\left\{v(t;n)\right\}$

Cabe destacar que no se podrá recuperar la señal original ya que se ha descartado la información de la fase de la señal, por medio de la utilización del valor absoluto, y la transformación de $v(m)$ a $c(m)$ no es invertible.

Los CCEM se definieron con el fin de conservar las características de los CC integrando información relativa a la percepción humana. Un *mel* es una unidad de medida de la frecuencia fundamental percibida. Ésto no se corresponde linealmente con la frecuencia física del tono, debido a que éste no se percibe de manera lineal por el sistema auditivo humano. Mediante

investigaciones fisiológicas se llegó a plantear una ecuación que permite el mapeo entre las frecuencias en escala real (Hz) y en las frecuencias en escala perceptiva (mel)[7]:

$$F_{mel} = \frac{1000}{\log 2}\left[1 + \frac{F_{Hz}}{1000}\right] \tag{1.5}$$

En general, para el RAH sólo se utilizan los primeros 13 coeficientes de los CCEM, pues se descarta lo relativo al pulso glótico.

1.1.2. Organización estructural del habla

El habla puede organizarse según distintas estructuras jerárquicas de acuerdo con el aspecto que se considere como central. La lingüística provee de una jerarquía en base a la que se pueden desarrollar muchos otros estudios [14]. El objeto de estudio es principalmente la estructura del mensaje, despojándolo de los mecanismos que lo han generado. En este sentido, la fonética y la fonología estudian los sonidos elementales de una lengua tanto en lo que respecta a su acústica como a su función en el sistema de comunicación. Aquí no se considera el significado que transmiten estos sonidos y los símbolos asociados, y se analiza la prosodia a niveles de suprasegmentos y sílabas, aunque ésta abarque en su estudio niveles superiores en la estructura del habla. Las manifestaciones de los distintos niveles pueden ser unidades disímiles e independientes que ocurren sin modificar los rasgos característicos a cada nivel.

Fonemas, suprasegmentos y sílabas.

En la Tabla 1.1 se mencionan algunos de los niveles jerárquicos de la organización estructural del habla. Del análisis del proceso de generación y el resultado acústico se establecen modelos para los sonidos elementales del habla y se los denomina fonemas. Este es el nivel en el que pueden distinguirse las primeras unidades del habla.

Fonemas:	/e//s/ /u//n/ /b//a//R//k//o/
↓	
Acentuación:	/A/ /A/ /T/ /A/
↓	
Sílabas:	/es/ /un/ /bar/ /co/

Tabla 1.1: Algunos niveles de la organización jerárquica del habla.

En relación con los patrones de pronunciación, los modos articulatorios y los sonidos producidos, se hace una clasificación de los sonidos del habla en dos grandes grupos: los sonidos vocálicos o vocoides y los sonidos consonánticos o contoides. Las vocoides son las realizaciones acústicas de las vocales

y se definen como los sonidos que se producen sin modificar la morfología del tracto, por donde el aire circula desde los pulmones hasta el exterior. Las contoides son las realizaciones acústicas de las consonantes y corresponden a los sonidos producidos con algún estrechamiento u oclusión en la morfología del tracto vocal. De aquí se concluye que los sonidos vocálicos son producidos fundamentalmente por las cuerdas vocales (y por lo tanto son todos sonoros) y los consonánticos poseen más componentes generadas por turbulencias y oclusiones en el tracto vocal.

- **Los sonidos vocálicos**

  Las vocales en español son fácilmente identificables debido a sus formantes; pero tiene como desventaja que no aportan tanta información como la consonantes:

  - `e_ _e__o _i_ _o__o_a__e_ e_ _i_i_i_ _e e__e__e_`
  - `_l t_xt_ s_n c_ns_n_nt_s _s d_f_c_l d_ _nt_nd_r`

  Es apreciable también, que su duración es más extensa que la de otros sonidos.

  Las formantes $F1$, $F2$ y $F3$ son las más importantes para la caracterización de los sonidos vocálicos, incluso siendo posible realizar una buena clasificación con solamente las formantes $F1$ y $F2$. Las formantes superiores, con frecuencias generalmente mayores a los 3200 Hz, son bastante diferentes para distintos hablantes y caracterizan factores personales. En la Tabla 1.2 se exponen los valores habituales de las dos primeras formantes.

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

Tabla 1.2: Valores usuales de F_1 y F_2 para vocoides.

- **Los sonidos consonánticos**

  La variedad y las características que identifican a las contoides son mucho más amplias que en el caso de las vocoides. Una clasificación general es:

  - Oclusivas suaves: [b], [d], [g].
  - Oclusivas fuertes: [p], [t], [k].

- Nasales: [m], [n], [$\hat{n}$].
- Líquidas laterales: [l], [λ].
- Líquidas vibrantes: [r], [$\bar{r}$].
- Fricativas sordas: [f], [s], [∂], [x].
- Fricativas sonoras: [y], [β], [∂], [γ].
- Africadas o semioclusivas [$\hat{y}$], [$\hat{c}$].

Los suprasegmentos están relacionados con la expresión y representados principalmente por el acento, la cantidad y la entonación [14]. Estas estructuras poseen diversas manifestaciones físicas y sus correspondientes modelos y símbolos lingüísticos asociados. Las reglas que rigen su uso se agrupan bajo la denominación general de prosodia.

El suprasegmento es una estructura de duración mayor a la de fonemas y menor a la de morfemas o palabras, que es afectada por rasgos prosódicos comunes. En este rango de tiempo se encuentra la sílaba; que no es un suprasegmento, pero se le aproxima en su duración.

Una sílaba se constituye por un núcleo sonoro o vocálico y su contexto. El núcleo generalmente es el que posee la mayor apertura articulatoria y debe permitir la extensión de su duración. La división en sílabas del español está definida por un conjunto de reglas sencillas basadas en su representación ortográfica [15]. La acentuación refiere a una representación de los suprasegmentos, en la que las distintas sílabas, según sean acentuadas o no, se caracterizan como Tónicas (T) y Átonas (A) respectivamente.

Entonación.

El término entonación se utiliza, en un sentido amplio, para hacer referencia a un conjunto de fenómenos lingüísticos relacionados directamente con la F_0 de las emisiones de voz. La diversidad de niveles a los que se estudia la entonación incluye: F_0, tonema, grupo de entonación y curva melódica [16]. La F_0 se mide en cada tramo de análisis y constituye el nivel más elemental de estudio poseyendo la menor duración en el análisis. Además, es punto de partida del análisis en los niveles superiores [13].

1.1.3. El RAH mediante modelos ocultos de Markov

En esta aplicación el objetivo es comprender como se genera y entiende naturalmente el habla, aprender de este proceso y luego poder diseñar buenos sistemas de RAH. Hay buenas razones para suponer que el proceso del habla se puede modelar adecuadamente como un proceso estocástico:

- El mismo *sonido/fonema/palabra* suena diferente con cada pronunciación.

- Podemos suponer que, al hablar, se transita aleatoriamente entre diferentes configuraciones del tracto vocal y en cada configuración se emiten fonemas siguiendo alguna distribución de probabilidades.

Los modelos ocultos de Markov (MOM) son modelos estadísticos que proporcionan descripciones de secuencias de eventos. Para comprender la idea central, se ha de suponer que se tiene un modelo asociado a cada emisión oral y que éste logra una similar emisión. Entonces, se deberían considerar tantos modelos como distintas emisiones se tengan. Luego, al tomar una emisión de voz, se compara con las emitidas por los modelos con el objetivo de determinar cual es el que puede generarlo y así presentar el texto asociado a dicho modelo.

Para una aplicación real se debería tener una cantidad infinita de modelos, lo cual no es posible y además, puede que estos modelos no sean significativamente distintos entre sí. Utilizando esto último junto a la organización estructural del lenguaje, comentada en la sección anterior, se podría modelar componentes simples como fonemas, silabas, etc. y luego combinarlos para formar los componentes complejos que se requieren [10].

Modelo de lenguaje y modelo compuesto

A partir de aquí se dejan a un costado las características físicas de las señales, se dejan los fonemas para enfocar el estudio en los niveles de palabras y sus combinaciones para formar frases. Partiendo de la idea de los modelos de fonemas y enfocándolos en este nuevo nivel, se podría imaginar un modelo para las palabra. Estas estructuras son conocidas como *gramáticas*.

Sus secuencias de estados también puede verse como una cadena de Markov para la que se extienden los conceptos de los MOM [10]. Las descripciones fonéticas de cada palabra forman el *diccionario fonético*, y con este se pueden formar las palabras del modelo de lenguaje (ML) a partir de los modelos acústicos (MA) de los fonemas. Con todo esto se arma un modelo compuesto (MC) capaz de modelar cualquier frase. La Figura 1.3 se pueden ver los tres niveles: MA, diccionario fonético y ML.

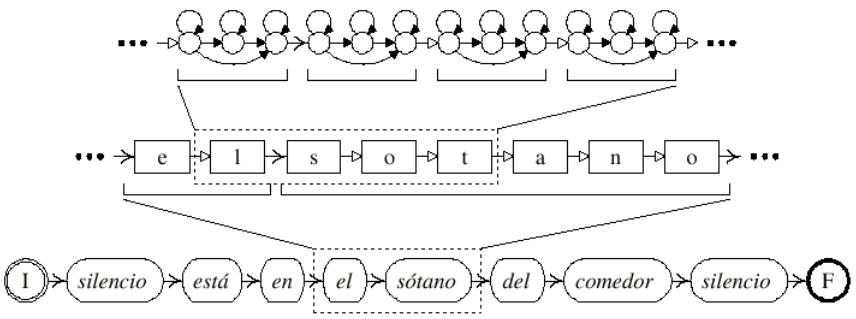


Figura 1.3: Modelo compuesto. (adaptada con permiso de [9])

El proceso de reconocimiento culmina eligiendo el modelo de la frase que mayor probabilidad posea[1], esto da como resultado el texto con que se formó la frase. Se darán más detalles de las técnicas utilizadas para el reconocimiento cuando en el Capítulo 2 se traten las redes de palabras.

Por último debe notarse que existen dos conjuntos de parámetros a estimar durante el entrenamiento: las probabilidades de transición y observación de los MA y las probabilidades de transición del ML. Estas estimaciones se realizan separadamente.

1.1.4. Aplicaciones y dificultades

Algunas de las aplicaciones más destacadas de los sistemas de RAH se citan a continuación:

Ayuda a personas con capacidades especiales. Aplicaciones diseñadas para ser utilizadas por personas que no pueden teclear o tienen dificultades para hacerlo, personas con dificultades de audición, personas temporalmente inmovilizadas, niños con trastornos del habla.

Dictado automático. En estas aplicaciones se sustituye el uso del teclado para la redacción de documentos, duplicando, triplicando y más las velocidades de escritura manual y mecanográfica.

Transcripción y traducción voz a voz. Este tipo de aplicaciones son utilizadas en transcripción automática de boletines de noticias, de discursos parlamentarios, de intervenciones en procesos judiciales, de conferencias, de clases, etc. Traductores universales de idiomas, acceso a boletines de noticias en lengua extranjera, a mensajes de megafonía en estaciones y aeropuertos, conferencias, etc.

Operación de máquinas a través de la voz. Podemos incluir aquí el control domótico, control por teléfono de contestadores automáticos, acceso a bases de datos, servicios interactivos, control de teléfonos móviles, ofimática, control de PDAs. El control "manos-libres" en aplicaciones industriales: trabajos para los que las manos del operario no estén disponibles o bien se presente algún peligro para éste, puede convenir usar la voz para controlar dispositivos.

Otras aplicaciones. En una categoría más general se podrían citar la identificación de idioma, la selección del reconocedor a utilizar, la selección del operador que debe atender una llamada en un servicio telefónico de emergencia, la identificación de locutor, la sustitución de contraseñas en el acceso a equipos informáticos o de PIN en acceso a cajeros automáticos, el control de juegos y juguetes, la realidad virtual, el control de dispositivos (DVD, video, etc).

[1] para este proceso se utiliza una extensión del algoritmo de Viterbi [10]

Se debe tener en cuenta que existen muchos factores que hacen que el RAH sea un problema complejo. A continuación se describen los más importantes:

Bidireccionalidad. Debido a que suele ser un proceso de diálogo el que se analiza, se presentan los problemas típicos de las transmisiones bidireccionales.

Incompletitud. Se intercambia más información de la transmitida. Esto se debe al mal uso del lenguaje, a la gesticulación, etc.

Continuidad. Cuando se analiza el habla continua, las marcas de separación de elementos (fonemas, sílabas, palabras, frases, etc.) que creemos percibir generalmente no existen.

Redundancia. Existe redundancia debido a que se transmiten unos 50 bits por segundo de información y la señal requiere alrededor de 100.000 bits por segundo.

Transitoriedad. Existe mucha información en zonas transitorias (consonantes, transiciones entre vocales, etc.).

Variabilidad. El habla es un fenómeno complejo y afectado por numerosas fuentes de variabilidad. Las unidades elementales (fonemas) son muy dependientes del contexto. El entorno, posición y características del micrófono también son fuente de variabilidad acústica. Diferentes locutores presentan diferencias fisiológicas y sociolingüísticas que hacen sus pronunciaciones muy diferentes. El estado físico y emocional del locutor afectan a su voz.

Ruido. Se denomina así a todas aquellas señales que no poseen información relacionada al fenómeno que estamos analizando. En la vida real siempre estamos en presencia de ruidos y estos provienen de diversas fuentes: otros hablantes, motores, vehículos, etc.

Ambientes del RAH. Las condiciones en las que se entrena un sistema RAH son, a menudo, muy disímiles a las de prueba o laboratorio. Mientras que las frases para entrenar un sistema RAH pueden estar leídas y de hecho bien pronunciadas para que el sistema saque provecho de éstas, cuando el sistema de RAH se emplea en un ambiente real, el discurso continuo que debe enfrentar es muy diferente y este es el mayor desafío.

1.2. Prosodia y RAH

Considerando a la señal acústica de la voz como el punto de partida, se podría pensar que en forma implícita todas las características del habla son

tenidas en cuenta. Sin embargo, las investigaciones [1] demuestran que la incorporación explícita de la información contenida en el habla a diferentes niveles de análisis favorece el rendimiento de todo el sistema de RAH. Es así como históricamente se han ido considerando progresivamente más y más características del habla. Los sistemas actuales de RAH incorporan muy diversos niveles de análisis del habla, desde el fonético hasta el gramatical. *Los rasgos prosódicos y la acentuación* se encuentran en uno de los niveles de análisis que aún no están completamente integrados al RAH.

Algunos investigadores han utilizado rasgos prosódicos en sistemas de traducción automática [17, 18], o para detectar eventos espúreos y fines de frases o palabras [19, 20, 21, 22]. En [23] se propuso un método para incorporar información adicional a un sistema de RAH mediante la penalización adaptativa del modelo de lenguaje. Luego, se utilizó esta técnica con éxito para incorporar información acentual en un sistema de RAH continua, pero se observó una débil asociación entre el acento prosódico y el acento ortográfico, lo que imponía una cota superior en las mejoras que podían realizarse [1]. En este trabajo se propone una nueva definición de la acentuación prosódica y su aplicación al RAH continua mediante el mismo método de penalización antes citado.

1.3. Objetivos

Se podría hablar de un objetivo general que comprende *el desarrollo de un sistema que permita el análisis de señales de voz con sus rasgos prosódicos y brinde la posibilidad de incorporar este análisis a un sistema de RAH.* Sin embargo, es menester detallar los objetivos específicos para establecer una línea de trabajo.

- Desarrollar rutinas de procesamiento de señales para el análisis de rasgos prosódicos.
- Evaluar las señales de voz con el fin de hallar rasgos prosódicos que sirvan para caracterizar a las palabras.
- Implementar de un sistema de RAH donde se aprecie la relevancia de la incorporación de rasgos prosódicos.
- Realizar el diseño de bibliotecas que permitan la aplicación de este método a un sistema de RAH ya entrenado.
- Realizar el análisis y el diseño del sistema utilizando los fundamentos de la orientación a objetos.
- Realizar las pruebas necesarias para verificar el funcionamiento del sistema.

- Analizar y exponer los resultados valorando las mejoras obtenidas.

2

Reconocedor automático del habla con información prosódica

En el español, la acentuación definida por las reglas ortográficas guarda una débil relación con las manifestaciones prosódicas del habla [1]. La idea principal de este trabajo es pasar a un segundo plano la información de la acentuación definida según las reglas ortográficas y hallar relaciones claras entre los rasgos prosódicos y las palabras que se pronunciaron, para poder definir una nueva forma de clasificar las prominencias acentuales del idioma. Una vez obtenida esta forma de clasificación, podremos incluirla en un sistema RAH para mejorar el reconocimiento con un método de penalización similar al propuesto en [23]. Para alcanzar estos objetivos se plantea una técnica de clasificación basada en histogramas. El método consiste en:

1. Identificar correctamente las posiciones de los fonemas en las frases: mediante un sistema de RAH previamente entrenado y, siendo conocidas las transcripciones, se realiza una alineación forzada con el algoritmo de Viterbi [24].

2. Extraer los rasgos prosódicos principales (F_0, energía, duración, etc) de cada señal: directamente aplicando los métodos clásicos de análisis por tramos de señales (ventanas con 10 ms de paso y 50 ms de ancho).

3. Seccionar las frases en sílabas y asociar los valores prosódicos correspondientes: en el punto 1 se obtuvo la segmentación en palabras y

fonemas [1]. Calcular para cada palabra los mínimos, medias y máximos prosódicos de cada sílaba.

4. Tomar todas las ocurrencias de la misma palabra y contabilizar las distintas estructuras que se presentan. A partir de estas, generar histogramas y clasificar las palabras según sus patrones prosódicos más característicos.

La incorporación de esta información al RAH se realiza por medio de las redes de palabras que se utilizan para el reconocimiento, más precisamente en el proceso de decodificación por el algoritmo de Viterbi. Las fases de esta etapa son:

1. Capturar la red de palabras de hipótesis del proceso de reconocimiento.
2. Identificar temporalmente cada hipótesis de palabra.
3. Identificar las posiciones de los fonemas en dichas palabras: mediante el sistema de RAH y, conocidas sus hipótesis de transcripciones, realizar la alineación forzada con el algoritmo de Viterbi.
4. Extraer los rasgos prosódicos principales de la frase que se está evaluando.
5. Seccionar las palabras en sílabas y asociar los valores prosódicos correspondientes.
6. Clasificar a las palabras según los patrones prosódicos presentados.
7. Comparar los patrones prosódicos de las hipótesis de palabras con las clasificaciones previas.
8. Penalizar la red de palabras y realizar el reconocimiento con esta red.

A continuación se detallan las características de todas las etapas antes listadas.

2.1. Clasificación prosódica basada en histogramas

2.1.1. Segmentación

Dentro de los muchos pasos que son más bien comunes a todos los desarrollos de sistemas RAH, a continuación se comentan los más destacados

[1] En el español la separación silábica se obtiene mediante la aplicación directa de las reglas ortográficas

y las modificaciones propias realizadas para la segmentación de la base de datos en este proyecto.

Las señales se analizaron con ventanas de 17.5 ms de ancho y con un paso de 10 ms. Los parámetros elegidos para ser extraídos fueron 12 coeficientes ceptrales en escala de mel y la energía. También se utilizaron coeficientes delta y aceleración. Las fórmulas de regresión para el cálculo de los coeficientes delta y aceleración se pueden revisar en [10] y como incluirlos a la parametrización se detalla en [25].

Para la creación y diseño del prototipo MOM se seleccionaron, como es habitual para fonemas, modelos de 5 estados. Las variables estadísticas que modelan la distribución de probabilidades de observación de estos estados con respecto a los coeficientes de las distintas frases de entrenamiento fueron *media* y *varianza*, modelando así con gaussianas esféricas en $\mathbb{R}^{39}$.

A partir de las transcripciones de las frases y las propias señales de voz se obtienen las segmentaciones en palabras y fonemas mediante alineación forzada por el algoritmo de Viterbi [26].

En un principio, las segmentaciones obtenidas no fueron las esperadas debido a que la rutina de entrenamiento no evalúa de manera precisa las pausas cortas (PC). Entonces se produjo una variación importante sobre el entrenamiento, realizándolo sin tener en cuenta la PC y se agregó el modelo de PC una vez entrenado completamente el MOM. El problema principal es definir la matriz de probabilidades de transición para este modelo, encontrar la más adecuada es el resultado de varias pruebas y un ajuste manual, debido a que la PC no se provee en las transcripciones de las frases de entrenamiento.

2.1.2. Extracción de rasgos prosódicos

Para el cálculo de la energía simplemente se realiza el producto interno de la señal consigo misma, lo que es igual al cuadrado de su norma-2 [11]:

$$E\{x(n)\} = \| x \|_2^2 = \sum_{i=1}^{N} |x_i|^2 \tag{2.1}$$

Para la extracción de la F_0 de la señal se utiliza un algoritmo basado en CC similar al de Noll [27], en éste se plantea como base el procesamiento homomórfico que se detalla a continuación.

Un sistema básico para generar sonidos de voz, consiste sólo en una señal fuente $s(t)$ pasando por el tracto vocal. El efecto del tracto está definido por su respuesta al impulso $h(t)$ y la salida $f(t)$ es igual a la convolución de $s(t)$ con $h(t)$ [27]:

$$f(t) = s(t) * h(t) \tag{2.2}$$

Ésto en el dominio frecuencial sería:

$$F(\omega) = S(\omega) \cdot H(\omega) \tag{2.3}$$

siendo $F(\omega) = \mathsf{TF}[f(t)]$, $H(\omega) = \mathsf{TF}[h(t)]$ y $S(\omega) = \mathsf{TF}[s(t)]$.
Aplicando valor absoluto y logaritmo se tiene:

$$C_y(\omega) = C_s(\omega) + C_h(\omega) \tag{2.4}$$

donde $C_y(\omega) = \mathsf{log}(|Y(\omega)|)$, $C_s(\omega) = \mathsf{log}(|S(\omega)|)$ y $C_h(\omega) = \mathsf{log}(|H(\omega)|)$.

Haciendo TF inversa

$$c_y(n) = c_s(n) + c_h(n) \tag{2.5}$$

Lo que se obtiene en esta última ecuación es el cepstrum de la secuencia de salida (señal de voz) que es igual a la suma del cepstrum de $s(t)$ y el cepstrum de $h(t)$. Por lo tanto, la convolución de dos secuencias en el dominio del tiempo se corresponde con la suma de las secuencias en el dominio del cepstrum [12]. La notable diferencia en la localización de las secuencias ceptrales ($c_s(n)$ y $c_h(n)$) permite que se puedan separar las dos componentes. Esto es debido a que, en el dominio de la frecuencia, la señal que corresponde al pulso glótico es la que se mueve más rápidamente y la correspondiente a la respuesta en frecuencia del tracto vocal es la que da la forma general del espectro (Figura 1.2). Es por esto que toda la información sobre el tracto vocal queda acumulada en los primeros coeficientes del cepstrum. Generalmente, mediante un *liftrado*[2] o simplemente tomando los primeros 30 coeficiente se puede obtener la información relativa al tracto vocal, debido a que sus componentes principales se encuentran en torno a valores pequeños de n. En RAH se suele descartar la información relativa a los pulsos.

En el cepstrum se observan picos localizados en el periodo fundamental de la señal y en múltiplos de éste que van decayendo en amplitud con n. Obteniendo del pico máximo dentro de los primeros 15 ms se podrá identificar el período fundamental y así determinar la F_0.

2.1.3. Generación de histogramas

Las palabras fueron preseleccionadas, eligiendo aquellas que tenían un número suficientemente alto de ocurrencias en la base de datos utilizada. Antes de continuar es menester explicar el proceso de clasificación de las palabras, introducido en la sección anterior. Una vez asociada la palabra a sus rasgos prosódicos, se calculan para éstos los valores máximos, mínimos y medios de cada sílaba para todos los sucesos de la palabra en la base de datos. Tomando los sucesos de cada palabra por separado y evaluando un rasgo particular se obtienen clases prosódicas codificadas en n dígitos, siendo n el número de sílabas de la palabra. Éste código indica en forma relativa la magnitud medida (por ejemplo: máximo de F_0) para cada sílaba. Paso siguiente, se contabilizan los sucesos de cada clase prosódica que poseen las palabras. A modo de ejemplo, la clase prosódica 321 (para palabras de tres

[2] denominación del filtrado en el dominio del cepstrum

sílabas) indica que la primer sílaba tiene valor máximo y la última valor mínimo; así la clase prosódica 213 indica que el valor máximo está en la última sílaba y mínimo en la segunda.

En primera instancia se puede ver que cada palabra, para sus distintos sucesos, y para cada rasgo prosódico, puede pertenecer a alguna de las $n!$ clases prosódicas que se forman de intercambiar las distribuciones de las cantidades en las n sílabas de la palabra. Afortunadamente para este método, casi todas las palabras pertenecen a unas pocas clases prosódicas y están caracterizadas en su mayoría por una única clase.

Para ejemplificar, se presentan algunas gráficas de valores relativos con resultados de la caracterización para distintos rasgos prosódicos. Cabe mencionar que, para simplificar las gráficas, se han eliminado las clases prosódicas para las que una palabra tiene cero sucesos.

En la Figura 2.1 se observan las clases prosódicas que caracterizan la palabra *dime* para la media de energía. Aquí se caracteriza completamente a la palabra con la clase prosódica *12* para 256 palabras computadas y este resultado se puede interpretar como: *la palabra dime se caracteriza por tener un valor mayor de energía media en la segunda sílaba.*

En la Figura 2.2 se ven las clases prosódicas que definen la palabra *longitud* para el rasgo prosódico mínimo de energía. Con 134 palabras computadas, se ve claramente que la clase prosódica *321* define completamente esta palabra. Este resultado se puede interpretar como: *la palabra longitud se caracteriza por poseer, para el rasgo mínimo de energía, el mayor valor en la primer sílaba y el menor valor en la tercer sílaba.*

2.2. Incorporación de la prosodia al RAH mediante redes de palabras

Ya se ha mencionado cuáles son las etapas necesarias para incorporar la información prosódica al RAH y que ésta se realiza por medio de las redes de palabras. En la Figura 2.3 se pueden apreciar gráficamente estas etapas y, a modo de ejemplo, como se realiza el proceso para una palabra *dime*.

A continuación se profundizan los conceptos y el método.

2.2.1. Redes de palabras

Una etapa muy importante para poder reconocer frases de palabras conectadas (discurso continuo) es la de definir el modelo de lenguaje que durante el proceso de reconocimiento da lugar a la red de palabras (RP). Existen varios tipos de ML, aquí se utilizó un modelo de bigramática [5] que puede verse en la Figura 2.4.

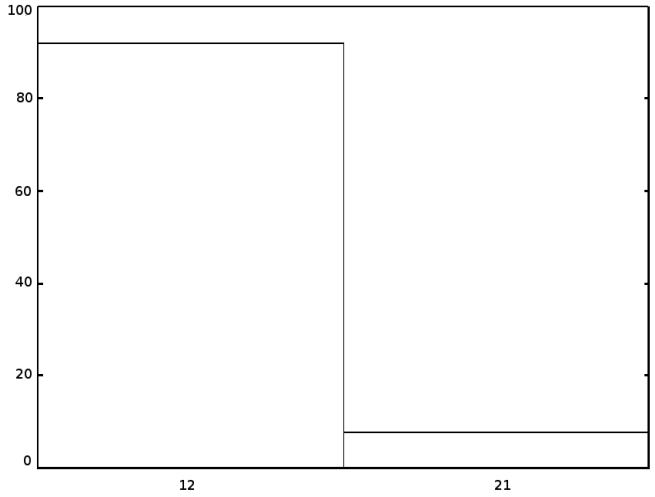


Figura 2.1: Clases prosódicas para la palabra *dime* en el rasgo prosódico media de energía

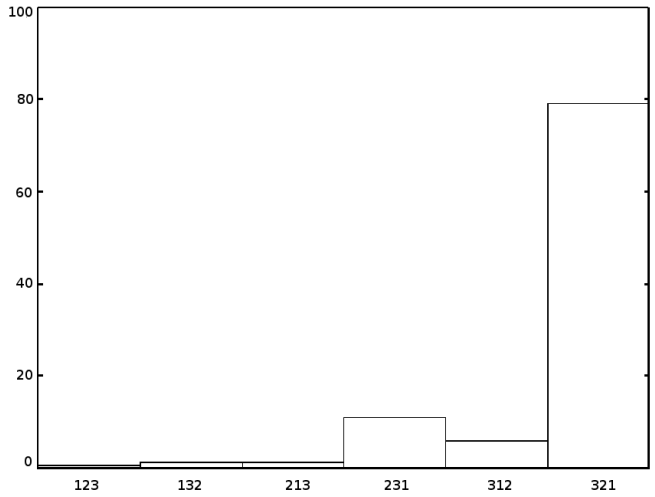


Figura 2.2: Clases prosódicas para la palabra *longitud* en el rasgo prosódico mínimo de energía

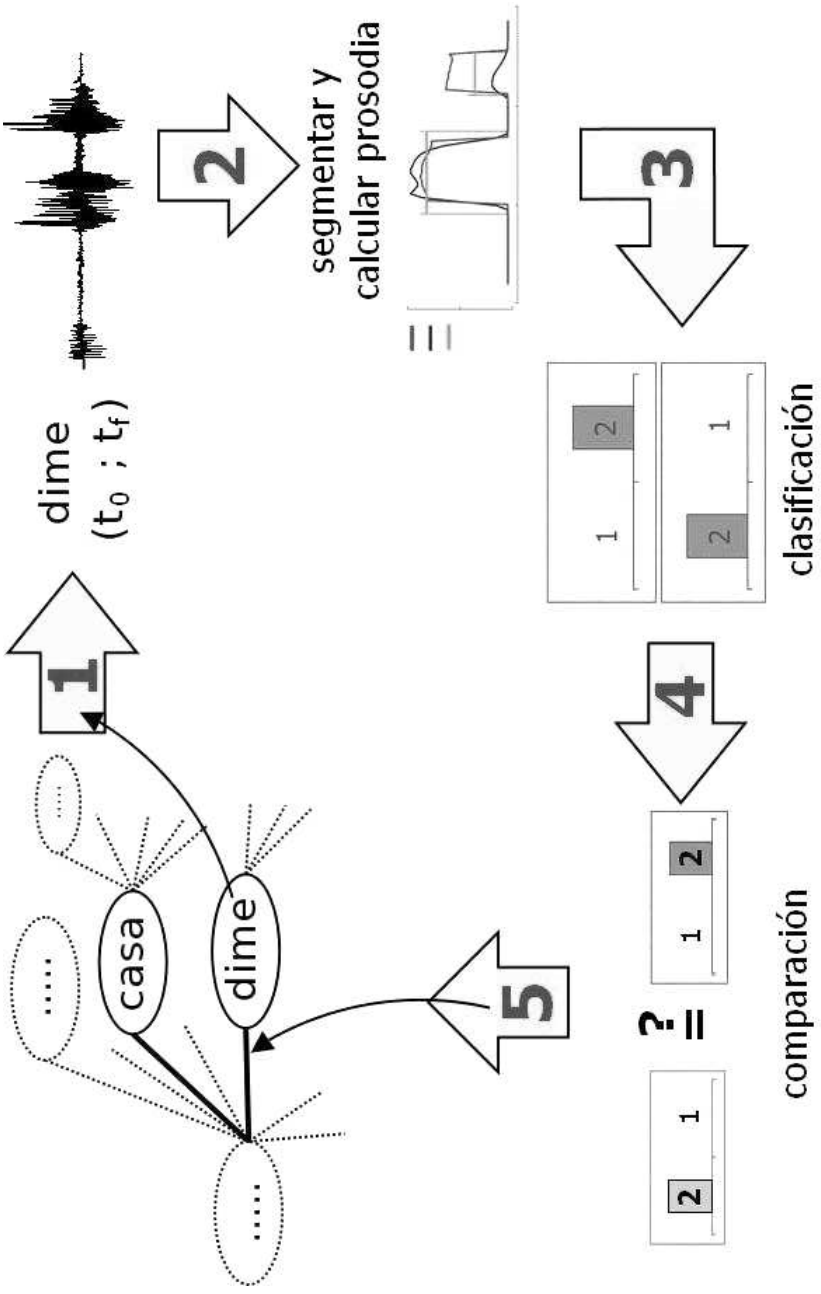


FIGURA 2.3: INCORPORACIÓN PROSÓDICA

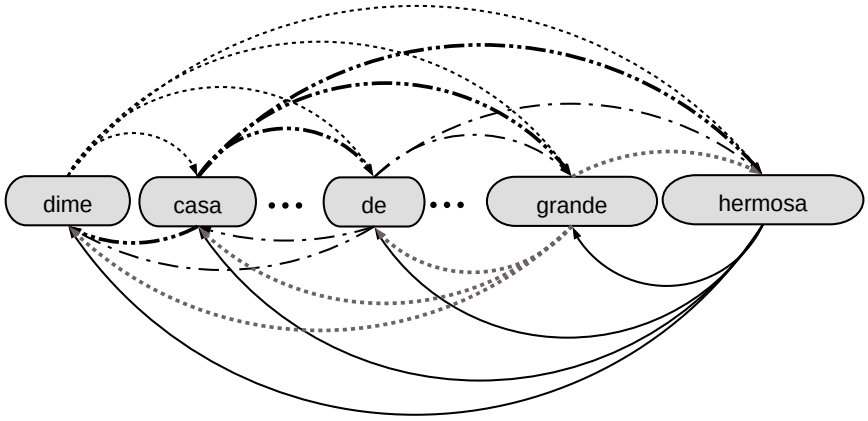


Figura 2.4: Modelo de Lenguaje bigramática

En principio se crea una RP, a partir del ML, que está compuesta por las M palabras distintas que conforman el corpus utilizado en el entrenamiento del sistema de RAH, y cada una de ellas conforma un *nodo* de la red. También son parte de la RP, para el ML bigramática, las M^2 conexiones posibles entre los nodos que se denominan *arcos*.

Una frase es entonces, desde este punto de vista, una secuencia "nodos" conectados por "arcos". Cada uno de los arcos mencionados tendrá asociado un valor que representa la probabilidad de transición entre las palabras que une (puede ser cero). Estos valores se calculan en la fase de entrenamiento a partir de las frases del corpus y una vez concluida esta fase se obtiene la RP indispensable para que el sistema de RAH reconozca habla continua.

Se comentó el proceso de generación de la RP, pero es menester comprender conceptualmente cómo se utiliza ésta en la etapa de reconocimiento para poder avanzar a la subsección 2.2.3. En la etapa de reconocimiento, el sistema de RAH asocia los modelos acústicos o MOM modelados en el entrenamiento, en una secuencia acústica representativa de la frase que se está reconociendo. Luego, utiliza los distintos modelos de fonemas para formar las palabras [3]. Superadas estas fases, el sistema de RAH genera una RP de *hipótesis* (RPH) ponderando la RP creada durante el entrenamiento y los resultados de las fases recién mencionadas. A esta RPH la podemos asociar con la gráfica de un árbol (Figura 2.5), con un inicio y un final, donde no necesariamente todas las palabras están conectadas entre si.

El último paso es utilizar esta RPH para obtener la secuencia de palabras con mayor probabilidad.

[3] se utiliza el diccionario fonético.

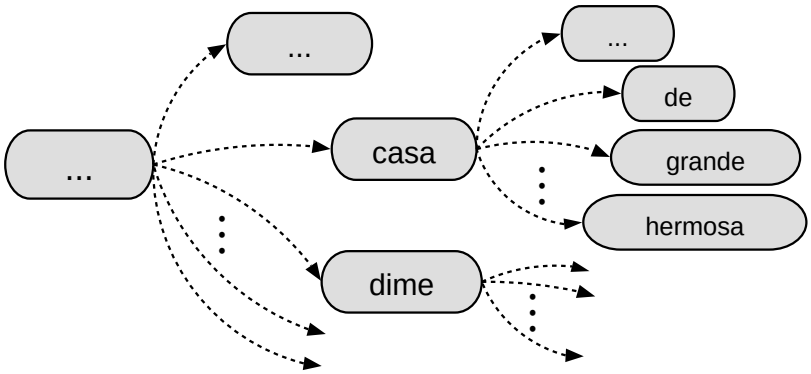


FIGURA 2.5: RED DE PALABRAS DE HIPÓTESIS

2.2.2. OBTENCIÓN DE LA PROBABILIDAD DE PENALIZACIÓN PROSÓDICA

Una vez en la etapa de reconocimiento, es preciso obtener la RPH que generó el sistema y de ésta se requiere obtener las palabras y sus localizaciones temporales. Con la lista de palabras de la RPH, sus segmentaciones y el archivo de audio se han de obtener los segmentos de audio parametrizados[4], para cada nodo de la RPH. Es preciso calcular los rasgos prosódicos del archivo de audio y luego asociar a cada nodo, según su ubicación temporal, el segmento de prosodia correspondiente. A partir de aquí se realiza la clasificación de los nodos de manera similar a la planteada en la subsección 2.1.3.

En la implementación, se ha incluido la información prosódica en los modelos de lenguaje para cada frase en particular. El método, a grandes rasgos, implica recorrer la RPH en búsqueda de palabras (nodos) que estén incluidas en la pre-clasificación realizada en una etapa anterior. En caso de hallar la palabra allí, se realiza el siguiente algoritmo que incluye pasos que fueron mencionados a grandes rasgos en el párrafo anterior:

1. Se extrae de la red de palabras la información de los tiempos en que está contenida la palabra. Se contemplan todas las hipótesis de palabras que da la red de palabras.

2. Se extrae la fracción de audio contenida en estos tiempos desde los archivos originales y para cada una se hace realiza el siguiente proceso:

   a) Se parametriza esta fracción de archivo con los parámetros que se usaron antes.

   b) Se utiliza el reconocedor, que ya estaba entrenado, para segmentar este tramo.

[4] se utilizan parámetros similares a los utilizados en la subsección 2.1.1

c) Se evalúan los rasgos prosódicos correspondientes a este lapso de tiempo. Y se asocian a las distintas sílabas.

d) Se clasifica esta palabra para cada rasgo prosódico, según la codificación de histogramas antes citada.

3. Se lleva a cabo la comparación correspondiente entre las clases prosódicas generales del modelo de histogramas prosódicos y las calculadas para esta palabra. En la implementación se adopta el criterio de indicar con un 0 que existe coincidencia y con un 1 el caso contrario.

Al finalizar la evaluación de toda la lista de p palabras se obtiene un vector, de tamaño p, de unos y ceros.

2.2.3. Modificación de la probabilidad en la red de palabras

A partir del vector de penalización esta etapa puede resumirse en lo siguiente:

> *En caso de no haber coincidencia (un 1 en el vector) se penaliza este suceso de palabra en la red de palabras y si existe coincidencia (un 0 en el vector) no se hacen modificaciones.*

Se debe tener en cuenta que el vector tendrá ceros también en aquellas posiciones que se correspondan con los nodos de la RPH que no pertenezcan al grupo de palabras pre-clasificadas.

Para realizar la penalización aún resta por definir la contante de ganancia que será multiplicada por el vector. Durante la implementación se han probado diversos valores llegando a la conclusión de que el valor óptimo es 2.

El sistema, que se plantea en el capítulo siguiente, implementa este método contemplando un diseño orientado a objetos.

3

Análisis y Diseño del Software

El objetivo de la ingeniería de requerimientos es brindar buenos requerimientos a las etapas posteriores de la ingeniería de software, de la cual forma parte. Es una disciplina que fuerza a considerar cuidadosamente a los requerimientos y a revisarlos dentro del contexto del problema. Asimismo, registra y refina los requerimientos y asegura la comunicación entre usuarios y analistas.

Esta etapa es necesaria e importante debido a que permite la detección y corrección de errores en forma temprana, y permite reducir la probabilidad de fallas.

> *El análisis de requerimientos es un proceso en que* ***lo que será hecho*** *se extrae y modela. Este proceso tiene que tratar con diferentes puntos de vista, y usa una combinación de métodos, herramientas, y actores.*
>
> [Leite]

3.1. Requerimientos

En la definición y análisis de los requerimientos se utilizó la metodología estándar en ingeniería del software [28]. Aunque en esta sección se hará especial énfasis en la fase de extracción de requerimientos y alguna otra, se han tenido en cuenta sobre éstas las consideraciones de las fases de análisis de

requerimientos, especificación de requerimientos y validación y certificación de requerimientos.

3.1.1. Equipo de trabajo

En la ingeniería de requerimientos (IR) se concibe como fase inicial la conformación de un grupo multidisciplinario bien estructurado para llevar a cabo las acciones de recolección de requerimientos. Para este proyecto, en relación a su extensión y a sus normas, el grupo comprende a una persona (el autor) que realizaría las tareas mencionadas y una persona que tomaría el rol de supervisor (director de PFC).

3.1.2. Hechos

El objetivo global es el desarrollo de un sistema que permita la incorporación de información prosódica en un RAH y brinde características didácticas, a los futuros usuarios, en el uso de la metodología planteada en la sección .Si bien las herramientas para manejar sistemas de RAH están disponibles, no se cuenta con herramientas que permitan la incorporación de la información prosódica al RAH. El desarrollo está orientado a la obtención de un software que permita la incorporación de esta información al RAH, o sea un sistema que permita el manejo de un reconocedor entrenado, de la información prosódica y de una interfaz de comunicación entre ambos.

3.1.3. Modelos

En esta sección se plantean modelos lógicos del sistema y las iteraciones del mismo. Éstos se utilizarán para identificar, analizar y validar los requerimientos. De la interacción entre los diagramas y la lista de requerimientos se obtiene una depuración de éstos últimos [29].

Los diagramas que se presentan a continuación son estándares y están definidos en el *lenguaje unificado de modelado* (UML[1]). La mayoría de los diagramas de UML y algunos símbolos complejos, son grafos que contienen formas conectadas por rutas. La información está sobre todo en la topología y no en el tamaño o la colocación de los símbolos. [2]

Casos de Uso

Los casos de uso (CU) se utilizan para obtener una perspectiva de los límites del sistema o subsistemas, sus actores (personas, dispositivos de hardware o sistemas de software con que se precisa interactuar), sus módulos y sus relaciones. En la Figura 3.1 se ve un CU que da idea de los actores

[1] del inglés Unified Modeling Language

[2] Para el lector no familiarizado con UML se sugiere consultar la reseña presentada en el Apéndice B.

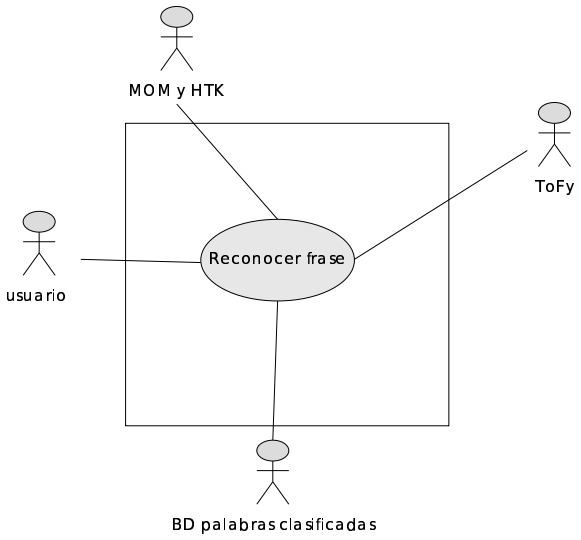


FIGURA 3.1: CASO DE USO GENERAL

externos y los límites del sistema. En la Figura 3.2, con un buen nivel de detalle, puede verse el contexto y también *qué debe hacer el sistema.*

En la Figura 3.2 pueden verse 4 actores externos:

- El usuario: es el que provee de la frase a reconocer al sistema
- MOM/HTK: refiere al reconocedor ya entrenado junto a la biblioteca del HTK
- ToFy: bibliotecas de rutinas para el manejo de archivos de audio y extracción de la prosodia
- BD palabras clasificadas: refiere a la base de datos (BD) de las palabras clasificadas por el método de histogramas[3]

También se ven los CU del sistema y como se relacionan. Las flechas con líneas de puntos indican inclusión. Los CU son:

- Reconocer frase: el usuario ingresa la frase a ser reconocida y espera el resultado del reconocimiento con y sin prosodia. Podría denominarse núcleo del sistema
- Generar RPH: refiere a la etapa de reconocimiento donde se genera una RP que son la hipótesis del actual proceso de reconocimiento
- Segmentar: refiere a la localización temporal de las palabras incluidas en la RP, a la identificación temporal de fonemas por el algoritmo de Viterbi y a la separación en sílabas de las palabras

[3]La BD fue obtenida en la investigación previa al desarrollo

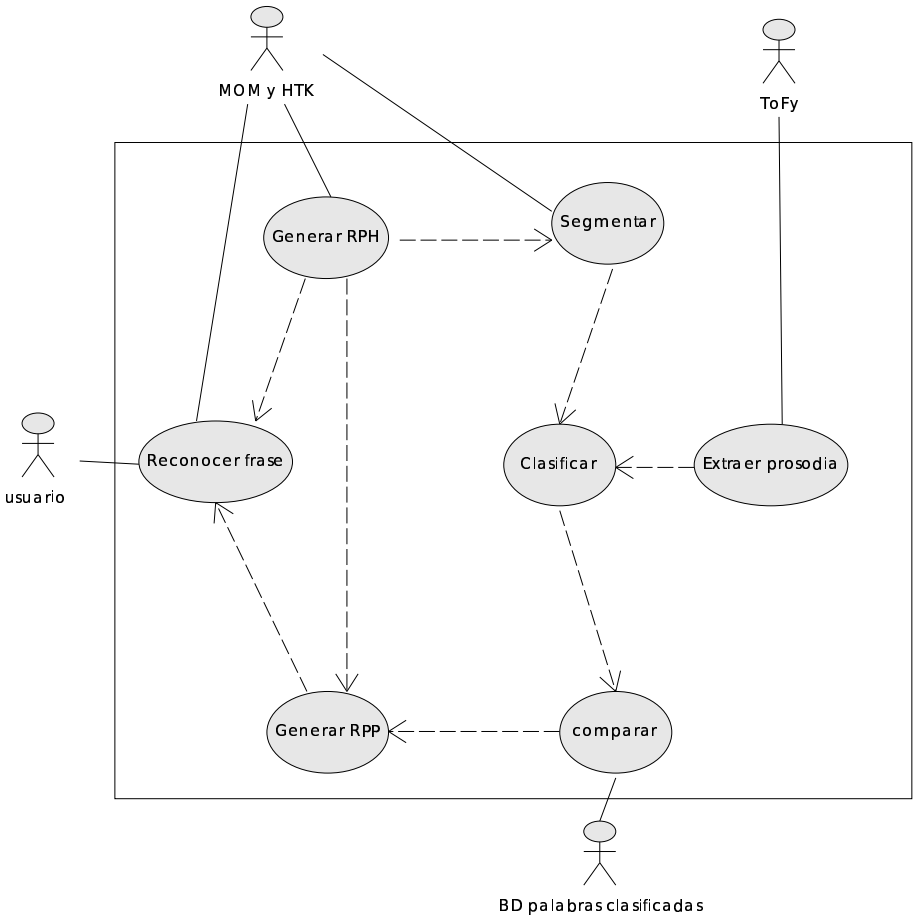


FIGURA 3.2: CASO DE USO DETALLADO

- Extraer prosodia: refiere a la utilización de ToFy para extraer la prosodia de la frase del usuario
- Clasificar: refiere al método de clasificación en base a histogramas, a partir de la estructura prosódica de las palabras
- Comparar: refiere al proceso de comparación que se realiza entre las palabras clasificadas, que se encuentran en la BD, y las que son hipótesis en el actual reconocimiento
- Generar red de palabras penalizada (RPP): refiere a la generación de una RP penalizada

Especificación de los CU:

Las especificaciones de los CU detallan las secuencias de actividades que se realizan normalmente y las alternativas según se requieran.[4]

[4]el CU *Reconocer frase* tiene dos especificaciones: reconocimiento con y sin prosodia

Nombre: **Generar RPH**	
Actor: **MOM y HTK**	
Caso Normal:	*Alternativas:*
1) Recibe el archivo de audio 2) Obtiene el MOM ya entrenado 3) Usa rutinas de HTK para obtener RPH 4) Devuelve la RPH	

Nombre: **Generar RPP**	
Actor:	
Caso Normal:	*Alternativas:*
1) Recibe la RPH 2) Recibe una lista de palabras a penalizar 3) Penaliza la RPH y obtiene la RPP 4) Devuelve la RPP	

Nombre: **Extraer Prosodia**	
Actor: **ToFy**	
Caso Normal:	*Alternativas:*
1) Se llama a ToFy para cálculo de prosodia 2) Se pasa al ToFy el archivo de audio 3) ToFy calcula prosodia 4) ToFy devuelve resultado	4.1) ToFy devuelve error

Nombre: **Segmentar**	
Actor: **MOM y HTK**	
Caso Normal:	*Alternativas:*
1) Recibe la RPH y el audio 2) Localiza cada palabra temporalmente 3) Usa algoritmo de Viterbi para identificar los fonemas de la palabra 4) Asocia fonemas para formar las sílabas y adjunta la extensión temporal de éstas 5) Devuelve las palabras separadas en sílabas y su localización temporal	3.1) El algoritmo devuelve error

Nombre: **Comparar**	
Actor: **BD palabras clasificadas** (BDpc)	
Caso Normal:	*Alternativas:*
1) BDpc levanta la BD 2) BDpc obtiene la palabra y su clasificación actual 3) BDpc busca la palabra en la BD 4) Encuentra la palabra 5) Compara las estructuras 6) Las estructuras son iguales: comunica que no hay error	 4.1) No encuentra la palabra 4.2) Termina la secuencia sin resultado 6.1) Las estructuras son distintas: comunica que debe penalizarse la palabra

Nombre: **Reconocer frase (1)**	
Actor: **Usuario - MOM y HTK**	
Caso Normal:	*Alternativas:*
1) Recibe el archivo de audio del usuario 2) Obtiene el MOM ya entrenado 3) Obtiene la RPH 4) Utiliza el MOM, la RPH y rutinas de HTK para reconocer 5) Devuelve frase reconocida **sin prosodia** al usuario	 1.1) Error de formato de audio

Nombre: **Reconocer frase (2)**	
Actor: **Usuario - MOM y HTK**	
Caso Normal:	*Alternativas:*
1) Recibe el archivo de audio del usuario 2) Obtiene el MOM ya entrenado 3) Obtiene la RPP 4) Utiliza el MOM, la RPP y rutinas de HTK para reconocer 5) Devuelve frase reconocida **con prosodia** al usuario	 1.1) Error de formato de audio

Nombre: **Clasificar**	
Actor:	
Caso Normal:	*Alternativas:*
1) Recibe las palabras separadas en sílabas y su localización temporal 2) Recibe las estructuras prosódicas 3) Asocia a las sílabas de las palabras, la estructura prosódica correspondiente 4) Calcula valores mínimos, medios y máximos en cada sílaba 5) Clasifica cada palabra por el método de histogramas 6) Devuelve las palabras clasificadas	

Escenarios

Los Escenarios describen la situación externa al sistema de software y sirven para ayudar a definir los requerimientos funcionales y no funcionales mediante su análisis y el de sus consecuencias.

Un escenario de la operación normal del sistema se presenta en la Tabla 3.1. En éste pueden apreciarse tanto, el contexto necesario para realizar el reconocimiento, la información esperada como objetivos de la ejecución, los recursos que tienen carácter de indispensables y la secuencia estricta de pasos, así como también la excepción que podría manifestarse.

Diagrama de secuencia

> *Un diagrama de interacción muestra una interacción, que consiste en un conjunto de objetos y sus relaciones, incluyendo los mensajes que se pueden enviar entre ellos* [29].

Existen dos tipos de diagramas de interacción: diagrama de secuencia (DS) y diagrama de colaboración. Aquí se utiliza el primero debido a que ambos son *isomorfos*[5] y a que el DS resulta más ilustrativo. En la Figura 3.3 se presenta el DS del sistema donde, pueden observarse los distintos actores y el flujo de información entre ellos. Los procedimientos y los mensajes tiene un ordenamiento temporal que permite comprender la dinámica del sistema.

[5]Se dice esto pues se puede convertir de uno a otro sin pérdida de información

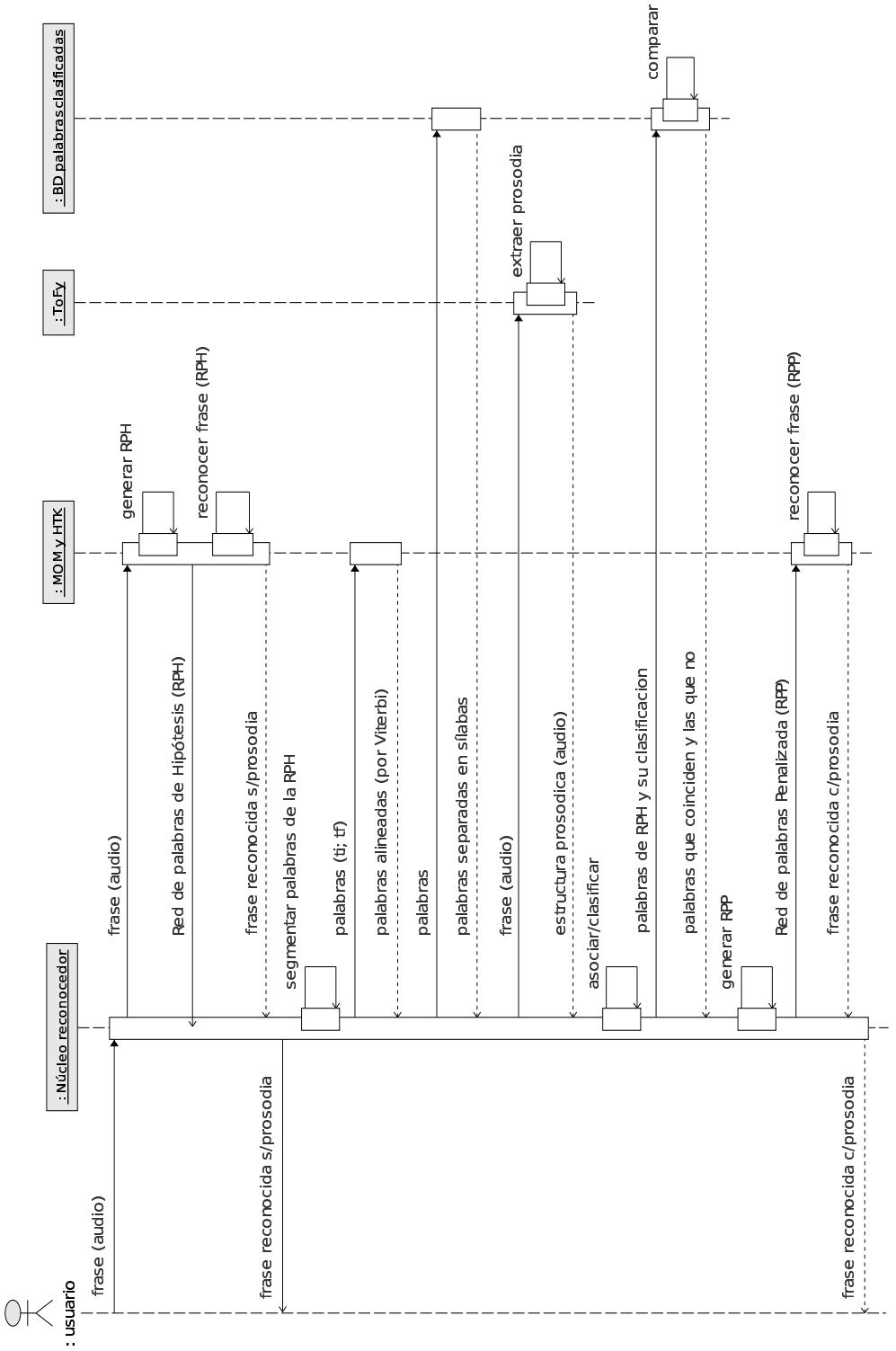


FIGURA 3.3: DIAGRAMA DE SECUENCIA

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Sistema de análisis prosódico y reconocimiento automático del habla. (Undergraduate project)"
Facultad de Ingeniería y Ciencias Hídricas - Universidad Nacional del Litoral, 2006.

	Reconocer una frase (archivo de audio)
Objetivo:	
	⋇ Reconocer la frase
	⋇ Reconocer la frase incorporando información prosódica
	⋇ Presentar resultados al usuario (frases, errores, etc)
Contexto:	
	→ El usuario posee la frase a reconocer
	→ Existe un reconocedor ya entrenado
	→ Existe una BD de palabras clasificadas prosódicamente
Actores:	
	★ Usuario
Recursos:	
	√ Frase (archivo de audio) del usuario
	√ Reconocedor entrenado
	√ Funciones del HTK
	√ Funciones del ToFy
	√ Funciones para clasificar y comparar palabras
Episodios:	
	1 - El usuario introduce en el sistema la frase
	2 - Se reconoce la frase con el reconocedor entrenado
	3 - Se incorpora al reconocedor la información prosódica
	4 - Se reconoce la frase con el reconocedor con prosodia
	5 - El usuario recibe los resultados del reconocimiento
Excepciones:	
	χ No se reconoce el formato del archivo de audio

Tabla 3.1: Escenario del sistema

3.1.4. Clasificación de los requerimientos

A partir de la interacción de los diagramas anteriores con los usuarios se definen los requerimientos o requisitos del sistema. Cabe destacar que los diagramas y las listas de requerimientos son de naturaleza dinámica, es decir, unos influyen sobre las actualizaciones y modificaciones de los otros. En este capítulo se presentan tanto los diagramas como los requerimientos finales ya validados y certificados.

Los requerimientos generalmente están formados por **requerimiento funcionales** y **requerimiento no funcionales**.

Existen otros requerimientos, llamados **de dominio**, que contemplan a los que derivan el dominio del sistema. Se pueden incluir aquí las tareas para llevar a cabo las distintas operaciones, las hipótesis y los parámetros que se utilizaron junto a las bibliotecas externas (HTK y ToFy).

Requerimientos Funcionales

Describen la funcionalidad o servicios que se espera que el sistema (o parte del sistema) provea. Dependen del tipo de software y del sistema que se desarrolle y de los posibles usuarios del software.

También declaran lo que el sistema no debe hacer.

Generales (punto de vista del usuario)

- Deberá cargar y utilizar un reconocedor entrenado (con HTK)
- Deberá cargar y utilizar una frase a elección del usuario
- Deberá ser capaz de reconocer la frase sin prosodia
- Deberá ser capaz de reconocer la frase con prosodia
- Proporcionará los resultados de reconocimiento con y sin prosodia (errores, aciertos, etc.)
- Proporcionará las frases como resultado de reconocimiento con y sin prosodia
- Deberá ser capaz de permitir el uso/expansión con otras BD
- No proveerá de herramientas para generar un reconocedor
- No entrenará un reconocedor
- No proveerá de herramientas para manipulación de archivos de audio y extracción de prosodia
- No proveerá de herramientas para generar la clasificación de palabras según sus histogramas prosódicos.

Específicos (punto de vista del sistema)

- Deberá permitir el cálculo de las variables prosódicas para la frase de usuario (archivo de audio)
- Deberá ser capaz de extraer la RPH del reconocedor
- Deberá ser capaz de seccionar cada hipótesis de palabra temporalmente
- Deberá ser capaz de identificar los fonemas temporalmente
- Deberá ser capaz de componer, temporalmente, palabras en sílabas de fonemas

- Deberá asignar los valores prosódicos a las sílabas de cada palabra según su localización temporal
- Deberá ser capaz de utilizar la técnica de histogramas para clasificar las palabras
- Debe permitir la manipulación de la BD de palabras clasificadas
- Deberá establecer un método para cotejar las palabras con la BD
- Deberá ser capaz de obtener una lista de nodos cuya estructura prosódica difiera a la establecida en la BD
- Deberá utilizar el método de penalización prosódica en la RPH utilizando la lista anteriormente obtenida y generar una RPP
- Deberá utilizar la RPP para el reconocimiento
- Deberá establecer un formato estándar para la BD de palabras clasificadas
- Deberá definir y agrupar a las funciones de manera simple y formar módulos/clases que faciliten la depuración, el mantenimiento y la extensión.
- Deberá definir interfaces simples entre módulos/clases
- Deberá ser capaz de definir los flujos de datos necesarios de manera simple y no redundantes. Éste debe considerarse como requisito básico en los dos items anteriores.

Requerimientos No Funcionales

No se refieren a las funciones especificas que brinda el sistema, sino a propiedades emergentes de éste y definen las restricciones de los servicios o funciones ofrecidos por el sistema.

- Portabilidad 1: el sistema utilizará la biblioteca HTK para el reconocimiento
- Portabilidad 2: el sistema utilizará un reconocedor ya entrenado con HTK
- Portabilidad 3: el sistema utilizará ToFy para manipulación de archivos de audio y extracción de prosodia
- Portabilidad 4: el sistema utilizará una BD de palabras ya clasificadas.
- Portabilidad 5 / usabilidad: el sistema proveerá soporte para ciertos formatos de archivos de audio. Éstos serán explicitados.

- Organizacional 1: en la implementación se utilizará el paradigma de orientación a objetos (OO).
- Organizacional 2: en la implementación se utilizará en un lenguaje de programación estándar.
- Organizacional 3: la definición de un estándar para la BD de palabras permitirá la manipulación (creación, extensión, modificación, etc.) y el uso para fines educativos esperados.
- Adaptabilidad/reusabilidad: para ésto se acompañará al paradigma de OO con documentación adecuada del código fuente.
- Éticos: la implementación del sistema y la documentación será GNU (*copyleft*). Se desarrolla con fines educativos y emplea bibliotecas con licencias que así lo permiten.

3.1.5. Racionalización y Prioridades

El costo principal para el sistema es la definición/desarrollo de módulos/clases que permitan la implementación de los requerimientos antes mencionados. También es menester la correcta definición de interfaces de comunicación entre los módulos. El riesgo principal recae en las definiciones/implementaciones de éstos, ya que las bibliotecas externas utilizadas han sido testeadas así como la BD de palabras, que es resultado de una investigación anterior [2].

En esta etapa, contando ya con requerimientos consistentes, se da un orden de prioridades, de manera tal que las necesidades de alta prioridad pueden ser encaradas primero, lo que permite definirlas y reexaminar los posibles cambios de los requerimientos, antes que los requerimientos de baja prioridad (que también pueden cambiar) sean implementados.

A partir de las definiciones del sistema, de sus límites y su tamaño se podría decir que, los requerimientos planteados poseen una prioridad similar para hacer posible el funcionamiento del sistema. Sin embargo, se podrían considerar algunos requerimientos cómo de baja prioridad y se incluyen otros que si bien no hacen al sistema, podrían considerarse como posibles extensiones (se presentan como módulos):
Ayuda del programa, Soporte para diversos formatos de audio, Módulo de interfaz para el manejo de la BD de palabras, Módulo de interfaz para el entrenamiento del RAH, Módulo de interfaz para la extracción de prosodia, Módulo para la generación de la BD de palabras.

3.1.6. Integración y validación

Esta tarea se lleva acabo de manera tal que sea posible obtener un conjunto de requerimientos, expresados en el lenguaje del usuario, de los cuales

se pueda validar la consistencia con respecto a las metas organizacionales obtenidas en la primera etapa.

- El sistema debe aceptar una frase del usuario y reconocerla utilizando un reconocedor ya entrenado
- El sistema debe utilizar una BD de palabras clasificadas (según método de histogramas) como condicionador en el reconocimiento de la frase del usuario y reconocerla utilizando un reconocedor con penalización prosódica
- El sistema proveerá ambos resultados al usuario
- Para el reconocimiento con prosodia se integrarán los requerimientos funcionales específicos.

3.2. Diseño

> *El diseño orientado a objetos es una estrategia de diseño en la cual los diseñadores del sistema piensan en términos de "cosas" en lugar de operaciones o funciones. El sistema se compone de objetos que interactúan entre ellos y que mantienen su propio estado local y suministran operaciones de esa información del estado* [28].

En esta sección se plantea el enfoque híbrido que se adoptó en el diseño de OO y se muestra el diagrama de clases (DC) del sistema. Para llegar a la definición final del DC se ha interactuado e iterado con los diferentes diagramas y requerimientos planteados en la sección anterior.

A continuación se plantean conceptos y definiciones propias de los DC y de su realización; luego se expone el DC de este sistema.

Se suponen entendidos conceptos generales relativos a diagramas UML [30] y Orientación a Objetos [29]. Para una revisión de éstos se sugiere ver Apéndices B y A.

3.2.1. Análisis y Diseño con el DC

El DC es el diagrama principal de diseño y análisis para un sistema. En él se especifica la estructura de clases[6] del sistema, con relaciones entre clases y estructuras de herencia [28, 29]. Durante el análisis del sistema, este diagrama se desarrolla buscando una solución ideal. Durante el diseño, se usa el mismo diagrama, y se modifica para satisfacer los detalles de las implementaciones.

[6] Una clase puede interpretarse como: "una plantilla para crear objetos"

3.2.2. Desarrollo de DC durante el análisis y diseño

A continuación se comentan aspectos tenidos en cuenta como parte de la metodología de OO que se utilizó durante el análisis y diseño del sistema.

- **Caso de Uso guiado**
  El diagrama de clases se desarrolla a través de información obtenida en los Casos de Uso, Diagramas de Secuencia y Diagramas de Colaboración. Los objetos encontrados durante el análisis son modelados en términos de la clase a la que instancian, y las interacciones entre objetos son referenciados a relaciones entre las clases instanciadas.

- **Diseño del sistema con Diagrama de Clases**
  El Diagrama de Clases se elabora para tener en cuenta los detalles concretos de la implementación del sistema. Se han considerado:

  - **Arquitecturas Multicapas.**
    La arquitectura del sistema incluye establecer si será un sistema simple diseñado para correr en una sola máquina, un sistema cliente y un servidor, o un sistema con módulos distribuidos.

    Este sistema no presenta mayores inconvenientes puesto que pertenece a la categoría de sistemas simples.

  - **Diseño de Componentes.**
    Un componente es un grupo de objetos o componentes más pequeños que interaccionan entre ellos y se combinan para dar un servicio que se especifica por la interfaz. El desarrollo es el proceso de ensamblar la combinación correcta de componentes en la configuración apropiada para llevar acabo la funcionalidad deseada.

  - **Análisis y diseño iterativo.**
    El DC se puede desarrollar en una forma iterativa, a través de un ciclo repetido de análisis, diseño e implementación, y después vuelta al análisis, para empezar el ciclo de nuevo. Este proceso se suele llamar 'round-trip engineering'.

3.2.3. Diagrama de Clases

Los DC son importantes no sólo para visualizar, especificar y documentar modelos, sino también para construir modelos ejecutables.

Si bien el sistema no es completamente orientado a objetos, pues se incluyen bibliotecas de funciones y estructuras de datos, cabe destacar que las clases planteadas son potencialmente reutilizables.

El DC general que se muestra en la Figura 3.9 da una visión global de las clases del sistema y como se relacionan. Sólo se ven las operaciones y atributos de las clases declarados como *públicos* y los nombres de las estructuras

de datos utilizadas. A continuación se mencionan las características más relevantes de cada clase y en las Figuras (3.4, 3.5, 3.6, 3.7 y 3.8) se pueden observar sus diagramas detallados.

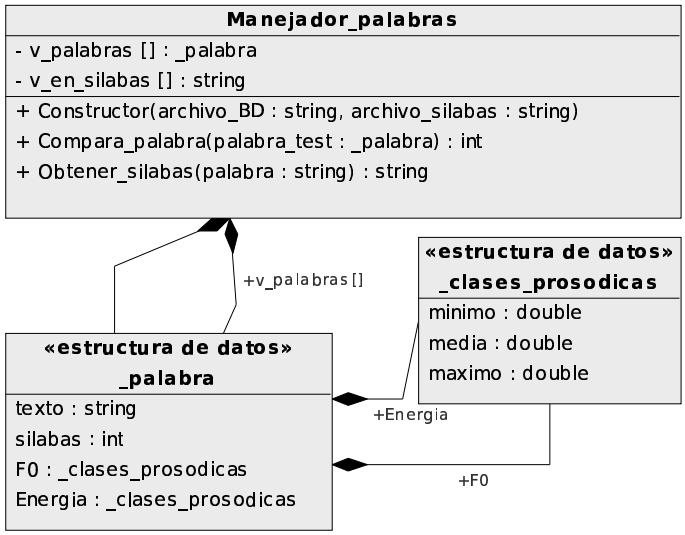


Figura 3.4: Diagrama de la clase Manejador_palabras

La clase *Manejador_ palabras* (Figura 3.4) tiene como principal responsabilidad el manejo de la BD de las palabras clasificadas. La operación **"Compara_palabra()"** se encarga de comparar una cierta palabra, con su clasificación prosódica asociada, con la misma palabra clasificada existente en la BD de palabras. Otra funcionalidad es la de aceptar una palabra y devolverla separada en sílabas en fonemas.

La clase *Manejador_ ToFy* permite la utilización de la biblioteca ToFY para el cálculo de los rasgos prosódicos y brinda una manera sencilla de obtener estos resultados (Figura 3.5).

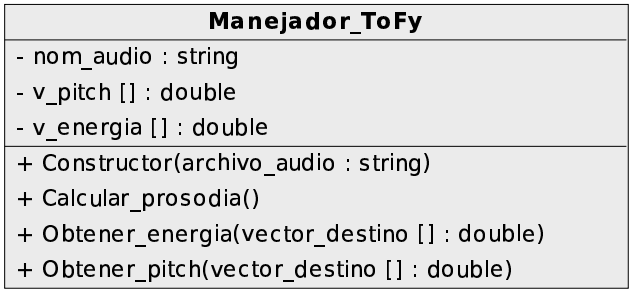


Figura 3.5: Diagrama de la clase Manejador_ToFy

En la Figura 3.6 se ve la clase *núcleo reconocedor* y las estructuras de datos asociadas. Esta clase instancia y utiliza los métodos de las demás clases.

Está compuesta de varios atributos (variables y constantes) que permiten la adecuada comunicación con las otras clases. Implementa como operación (**"Clasificador_histogramas()"**) el método de clasificación propuesto en [2].

La estructura de datos "**_palabra**" permite tener una palabra y su clasificación prosódica en una misma estructura. La estructura de datos "**_resultado**" permite manejar la información estadística del reconocimiento.

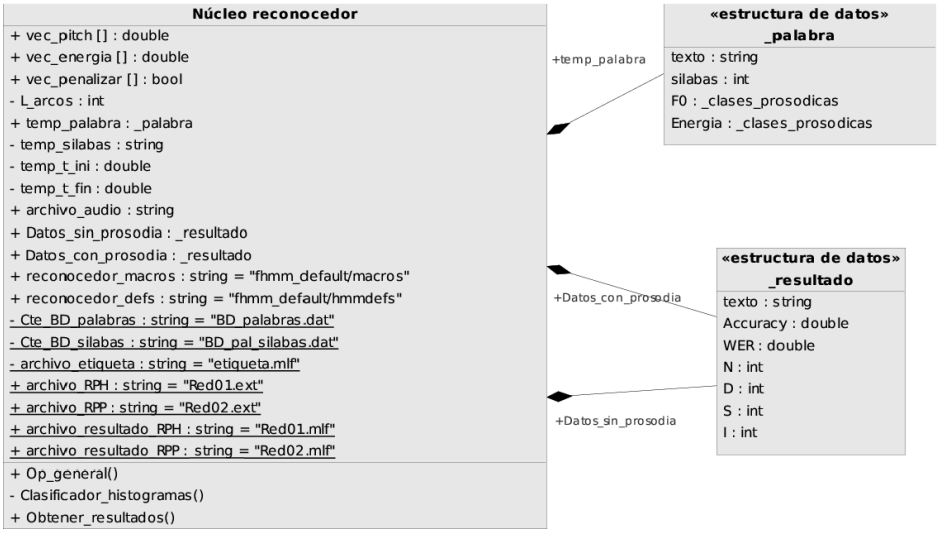


FIGURA 3.6: DIAGRAMA DE LA CLASE NÚCLEO RECONOCEDOR

La clase *Red_de_palabras* se puede ver en la Figura 3.7. La tarea principal de ésta es el manejo de la red de palabras, permite la penalización de los distintos arcos y tiene la capacidad de crear una red de palabras a partir de la estructura que posee en un determinado instante.

Las estructuras de datos "**_arco**" y "**_nodo**" permiten la manipulación de los parámetros de arcos y nodos respectivamente.

En la Figura 3.8 se ve la clase *M_Htk* que implementa operaciones de comunicación con la biblioteca del HTK. Permite utilizar el reconocedor para generar la RPH, reconocer una frase especificando la red de palabras, extraer las estadísticas de reconocimiento en una estructura de datos simple, utilizar el reconocedor para la alineación de fonemas, etc.

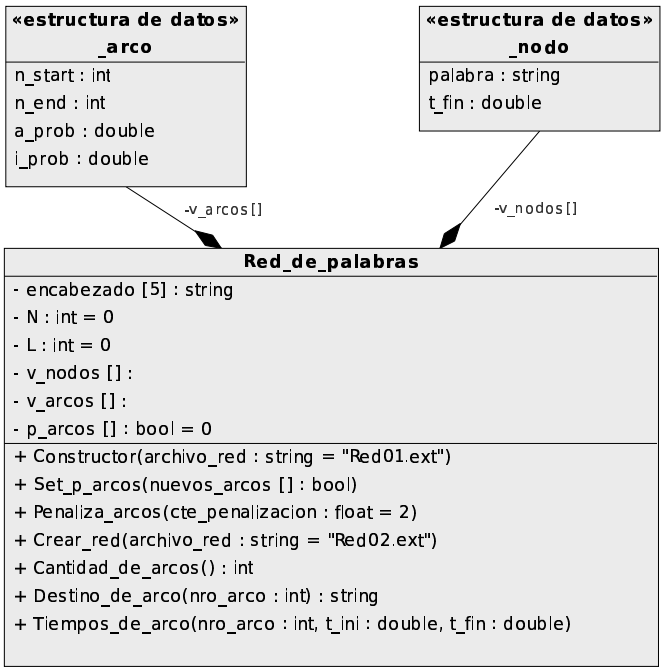


FIGURA 3.7: DIAGRAMA DE LA CLASE RED_DE_PALABRAS

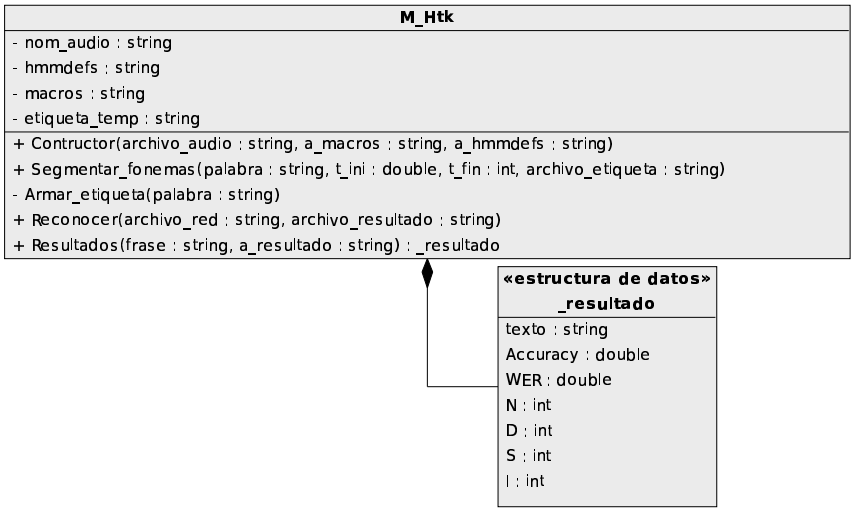


FIGURA 3.8: DIAGRAMA DE LA CLASE M_HTK

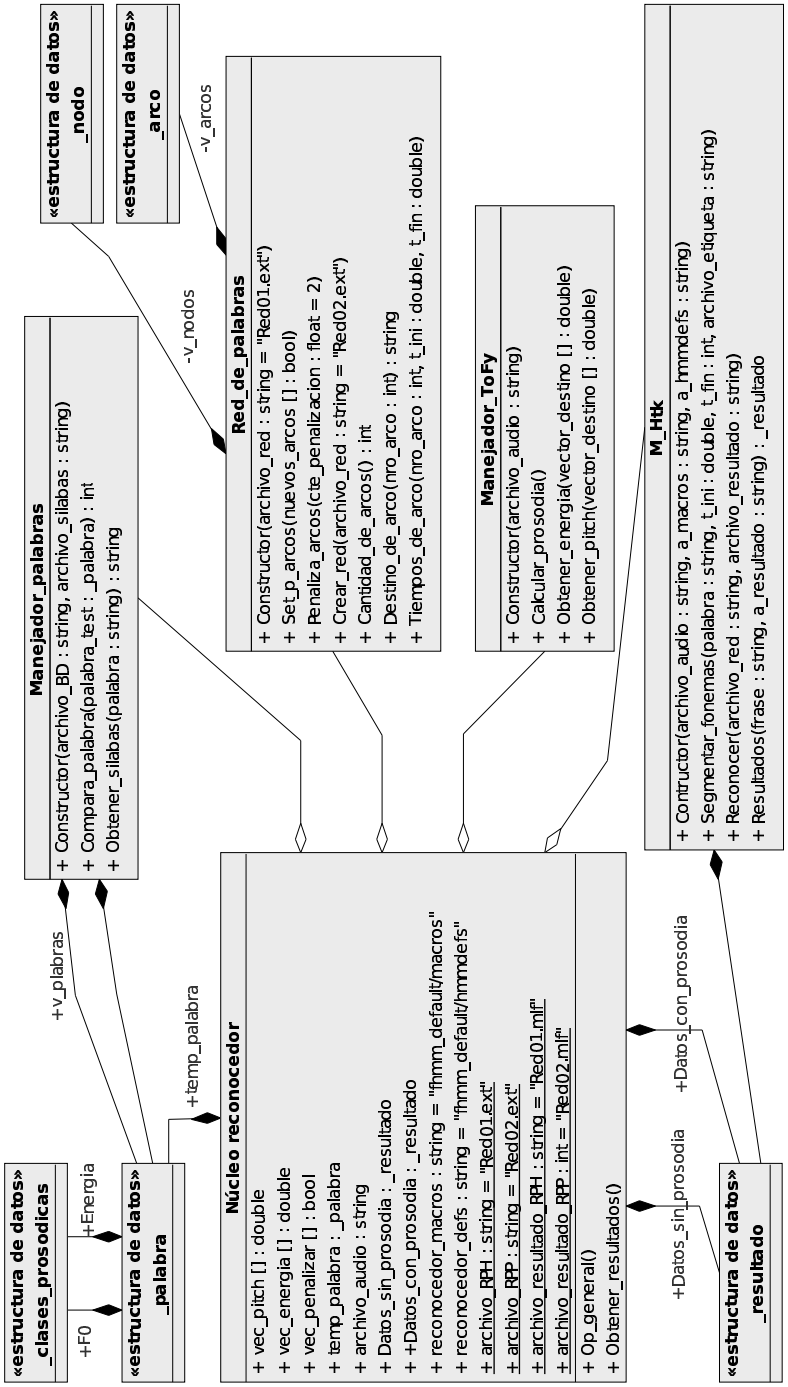


Figura 3.9: Diagrama de clases del sistema

4

Resultados y discusión

En este Capítulo se presentan los materiales utilizados en el desarrollo del proyecto y luego se analizan los resultados obtenidos en las secciones siguientes. La segregación de los resultados es menester debido a que se han obtenido resultados provenientes de la clasificación prosódica de las palabras (es decir de la aplicación del método descripto en la subsección 2.1.3) y resultados derivados del reconocimiento con penalización prosódica explicado en la Sección 2.2.

Para el desarrollo e implementación de los MOM se utilizó un conjunto de herramientas denominado *Hidden Markov Toolkit* (HTK)[1] [25]. Para la extracción de energía y F_0 de las señales se utilizaron rutinas del ToFy[2] y otras rutinas para el cálculo y las estadísticas fueron implementadas en *GNU C++* y *Free Pascal*.

Las frases utilizadas fueron extraídas de la base de datos Albayzin [31], creada por cinco Universidades españolas. Ésta se desarrolló con el objetivo de contribuir al desarrollo y la evaluación de sistemas de reconocimiento y procesamiento del habla. Los hablantes pertenecen a la variedad central del castellano, en su mayor parte de las comunidades de Castilla-La Mancha, Castilla-León, Cantabria y Madrid, con mujeres y varones de entre 18 y 55 años de edad. En la Tabla 4.1 se muestran los datos del subconjunto utilizado de esta base de datos.

[1] Desarrollado en el Speech and Vision Robotics Group en la Universidad de Cambridge, disponible en `http://htk.eng.cam.ac.uk`

[2] Desarrolladas en el Laboratorio de Cibernética de la Universidad Nacional de Entre Ríos (Argentina), disponible en `http://www.milone.tk`.

Total de elocuciones	1000
Total de frases con textos diferentes	500
Total de palabras	9448
Total de palabras diferentes	277
Hablantes femeninos	6
Hablantes masculinos	6

Tabla 4.1: Características el subconjunto de frases Minigeo 2.

4.1. Clasificación de estructuras prosódicas

Aquí se trataran los resultados de la caracterización en base a histogramas de las distintas palabras discriminadas por grupos según su cantidad de sílabas. En la Tabla 4.2 se puede observar, para el subconjunto de frases Minigeo 2, la cantidad de palabras que integran los grupos a los que se hace referencia. Se debe recordar que los monosílabos no son tenidos en cuenta por el método.

Descripción de grupo	Cantidad de palabras
Monosílabas	44
Bisílabas	115
Trisílabas	67
Quatrisílabas	44
Pentasílabas	6
Sextisílabas	1

Tabla 4.2: Grupos silábicos del subconjunto de frases Minigeo 2

Se puede observar un resumen de los resultados en la Tabla 4.3. La columna *Total de palabras evaluadas* da cuenta de la cantidad total de palabras de cada estructura silábica que posee la base de datos; la columna *Palabras diferentes* muestra cuántas palabras del total cumplieron los requisitos y se evaluaron; la columna *Porcentaje* muestra el porcentaje de las palabras que el método puede clasificar correctamente respecto de las que se evaluaron. La columna *Diferencia* indica cual es la diferencia relativa suficiente que ha de presentar la clase prosódica dominante sobre cualquier otra, para que se considere que ésta caracteriza a la palabra. Se ve que la capacidad de discriminación del método decrece marcadamente con la cantidad de sílabas, esto puede deberse a que no contamos con palabras de muchas sílabas que son representables por sus rasgos prosódicos o bien el método no es eficiente para esas cantidades de sílabas. Esto último podría comprobarse analizando una base de datos más extensa.

Cantidad de sílabas	Total de palabras evaluadas	Palabras diferentes	Porcentaje	Diferencia
2 sílabas:	1646	25	96.00 %	≥80 %
3 sílabas:	398	6	83.33 %	≥40 %
4 sílabas:	792	9	77.78 %	≥30 %
5 sílabas:	196	1	100.00 %	≥20 %

Tabla 4.3: Porcentajes de caracterización

Con la intención acercar al lector a la utilización y a la evaluación de los resultados del método, se presentan a continuación algunos ejemplos de histogramas.

En la Figura 4.1 se pueden observar las clases prosódicas para la media de F_0 de la palabra *valenciana*, que queda caracterizada por la clase prosódica *4321* para 34 palabras computadas. Se ve que más del 50 % de éstas pertenecen a la clase prosódica mencionada y ésta presenta una *diferencia*, respecto de las otras clases prosódicas, mayor al 30 %.

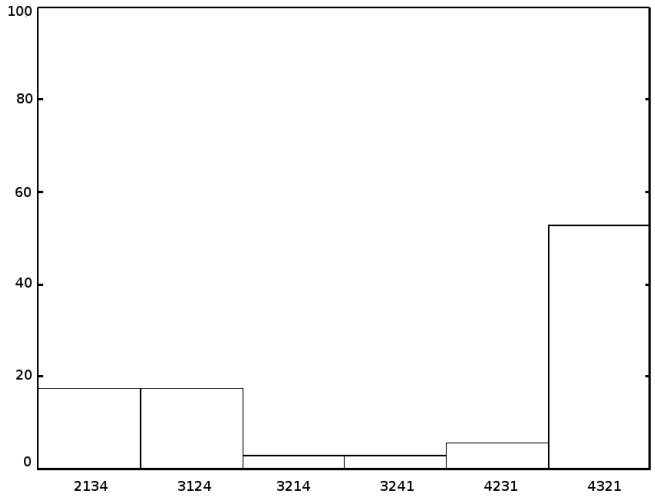


Figura 4.1: Clases prosódicas para la palabra *valenciana* en el rasgo prosódico media de F_0

En la Figura 4.2 puede verse a la clase prosódica *2134* caracterizando a la palabra *desemboca*, para el rasgo máximo de energía. De los 52 sucesos de la palabra, más del 60 % pertenecen a la clase prosódica *2134* y existe una diferencia mayor al 40 % sobre cualquiera de las otras 6 clases prosódicas.

Otro ejemplo puede apreciarse en la Figura 4.3 donde se observa que la clase prosódica *21* caracteriza a la palabra *metros*, para la media de F_0. De los 60 sucesos de la palabra, más del 90 % pertenecen a la clase prosódica *21*.

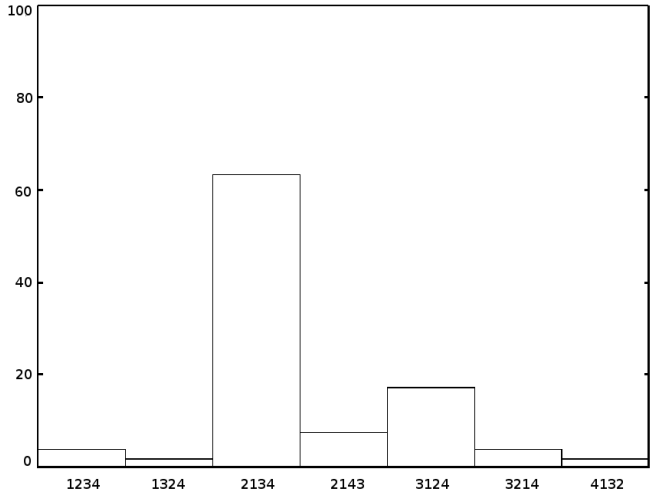


FIGURA 4.2: CLASES PROSÓDICAS PARA LA PALABRA *desemboca* EN EL RASGO PROSÓDICO MÁXIMO DE ENERGÍA

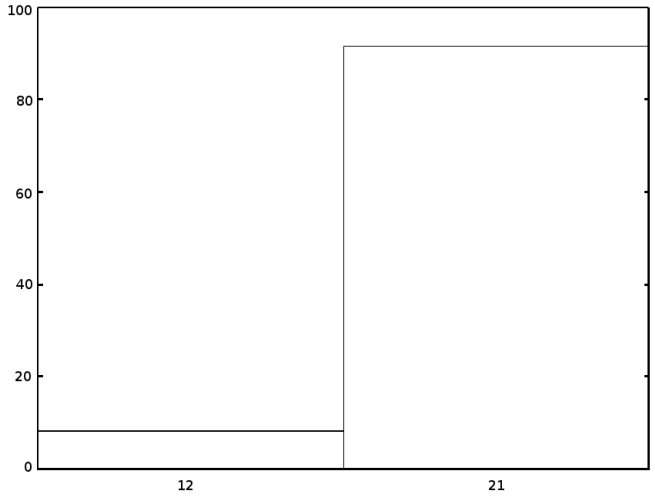


FIGURA 4.3: CLASES PROSÓDICAS PARA LA PALABRA *metros* EN EL RASGO PROSÓDICO MEDIA DE F_0

En la Figura 4.4 se ve que la clase prosódica *123* caracteriza a la palabra *superior*, para el rasgo media de energía. Cerca del 90 % de los 48 sucesos de la palabra pertenecen a la clase prosódica *123*.

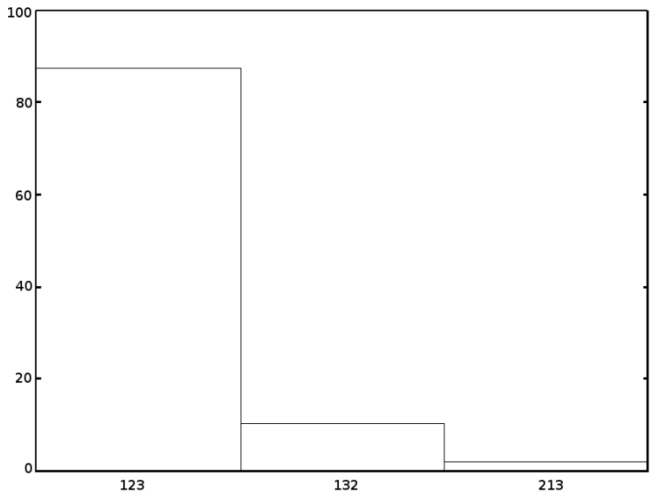


FIGURA 4.4: CLASES PROSÓDICAS PARA LA PALABRA *superior* EN EL RASGO PROSÓDICO MEDIA DE ENERGÍA

También para este método se ven palabras que no son clasificadas por ninguno de los rasgos prosódicos propuestos. Por ejemplo, para palabra *cúbicos* no se encontró una clase prosódica, de al menos un rasgo prosódico, que la caracterice. En la Figura 4.5 se observa el rasgo mínimo de energía para esta palabra, se ve que los 62 sucesos que presenta ésta se encuentran distribuidos entre dos clases prosódicas y que ninguna de éstas la caracteriza definitivamente, aunque estas dos clases juntas podrían caracterizar a la palabra considerando las otras 4 clases prosódicas para las que no hay ningún caso.

Un caso similar se presenta con la palabra *comunidad*, en la Figura 4.6 se ven las clases prosódicas para el rasgo mínimo de energía, con 256 sucesos computados de la palabra.

La palabra *valencia*, aunque está clasificada para la media y máximo de energía, en la Figura 4.7 se puede ver que las clases prosódicas para la media de F_0 no la clasifican. Se puede apreciar una distribución casi uniforme, de los 29 sucesos de la palabra, en el histograma.

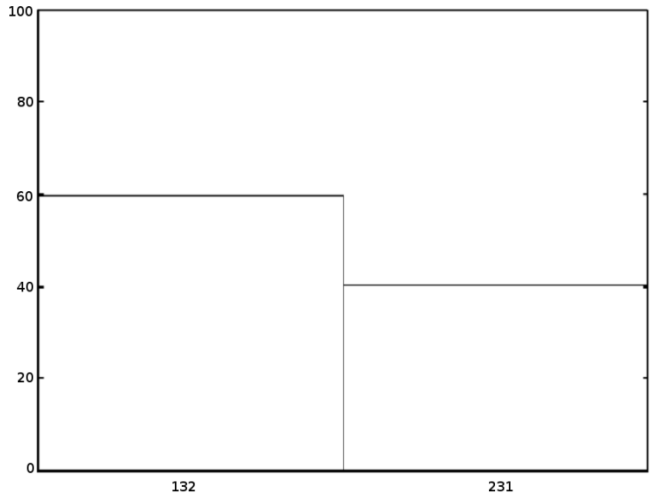


FIGURA 4.5: CLASES PROSÓDICAS PARA LA PALABRA *cúbicos* EN EL RASGO PROSÓDICO MÍNIMO DE ENERGÍA

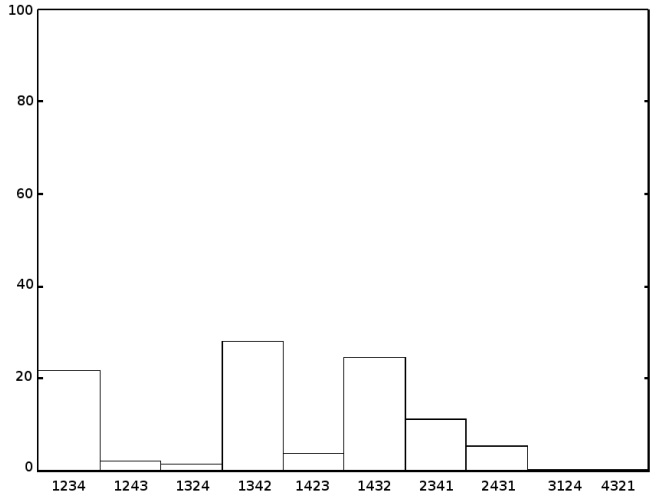


FIGURA 4.6: CLASES PROSÓDICAS PARA LA PALABRA *comunidad* EN EL RASGO PROSÓDICO MÍNIMO DE ENERGÍA

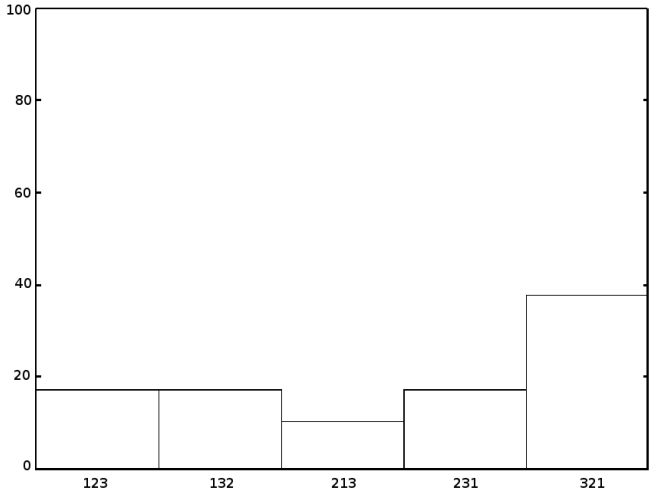


FIGURA 4.7: CLASES PROSÓDICAS PARA LA PALABRA *valencia* EN EL RASGO PROSÓDICO MEDIA DE F_0

Se vió que acorde al incremento del número de sílabas de la palabra, se incrementa la cantidad de clases prosódicas a la que puede pertenecer la palabra y entonces se debe reducir la tolerancia en las diferencias. Una característica interesante detectada en el análisis y que se da más frecuentemente en las palabras de tres o más sílabas, es que aparecen 2 o 3 clases prosódicas dominantes en las palabras para un parámetro prosódico determinado, esto es, de las $n!$ clases prosódicas que caracterizan a una palabra hay 2 o 3 que claramente se destacan de las otras. Esta última observación puede ser muy interesante desde el punto de vista que restringe también la caracterización de una palabra, si bien no a una, a varias clases prosódicas bien definidas.

Una extensión interesante de esta técnica es la combinación de los histogramas, esto es: se toman dos o más medidas por ejemplo: F_0 y energía, para las palabras de, por ejemplo, dos sílabas. Ahora bien, si se analiza por ejemplo la F_0 existen 2 secuencias posibles: que sea máxima la primera sílaba o que sea máxima la segunda sílaba. Si esto se combina con la energía, entonces se podría analizar para cada secuencia de F_0, dos secuencias de energía, o sea, un total de 4 combinaciones posibles que se traducen en cuatro posibilidades de clasificación para estas palabras de dos sílabas. Sin problemas se podría variar el número de sílabas y el número de variables prosódicas para dar $N = (s!)^r$ clases prosódicas en el histograma, siendo s el número de sílabas y r el número de de variables prosódicas.

4.2. Resultados de RAH

Para la prueba del método propuesto se ha utilizado como sistema de referencia un reconocedor del habla continua con idénticas características al descripto en la subsección 2.1.1.

Para medir la exactitud en el reconocimiento de palabras, se emplea un procedimiento basado en programación dinámica que alinea las etiquetas (palabras) para la comparación. Los resultados arrojados por la rutina del *HTK* son: el número de etiquetas correctas C, el número de etiquetas borradas B, el número de substituciones S, el número de inserciones I y el número total de etiquetas en el archivo de definición de las transcripciones N. El porcentaje de palabras reconocidas correctamente, *WER* (del inglés Word Error Rate), está dado por:

$$WER = \frac{N - B - S}{N} * 100\,\% \tag{4.1}$$

De las palabras expuestas en la Tabla 4.3 aquí sólo se contemplaron 35, debido a que la *duración del núcleo vocálico* que caracteriza a algunas de ellas no se ha tenido en cuenta y tampoco se consideran aquellas representadas por 2 o más clases prosódicas. En la Tabla 4.4 se exhiben las palabras utilizadas en la implementación y se muestra qué rasgo la clasifica. Ésta exhibe en su primer columna las palabras que actualmente integran la BD de palabras clasificadas, se presentan ordenadas por cantidad de sílabas y orden alfabético. Las demás columnas hacen referencia a los rasgos prosódicos utilizados en la etapa de clasificación. El número 1, en las columnas de Tabla, indica que existe una única clase prosódica (del rasgo que encabeza la columna) que caracteriza a la palabra.

Se han realizado pruebas preliminares sobre el subconjunto de frases Minigeo 2 y se obtuvieron buenos resultados. Con el reconocedor entrenado se re-evaluaron los datos. En el proceso normal de reconocimiento, el reconocedor arrojo un error de 5,99 % sobre el total de palabras correctas. Con la inclusión de la penalización prosódica se obtuvo un error de reconocimiento del 5,89 %. Estos valores se obtienen a partir de (4.1). Ahora se puede apreciar que la reducción relativa del error de reconocimiento es del 1,7 %.

Estas mejoras podrían parecer poco relevantes, pero no debemos dejar de considerar su contexto. Si bien no se realizó el análisis de significación estadística, se debe recordar que los experimentos se realizaron sobre 1000 elocuciones de 12 personas distintas (6 mujeres y 6 hombres), existen cerca de 9500 palabras de las cuales hay 277 palabras distintas y sólo 233 de éstas son capaces de ser contempladas por el método. De éstas últimas, 35 palabras tienen una clase prosódica que las define completamente. La posibilidad de incorporar más palabras pre-clasificadas, usando dos rasgos prosódicos o nuevos tipos de histogramas, redundará en beneficios directos en el sistema de RAH.

Debido al tipo de penalización que se ha planteado, el método claramente colabora con el sistema descartando las hipótesis de palabras que no coinciden con los rasgos prosódicos que presenta la frase analizada.

Para ilustrar las mejoras obtenidas mediante el método de penalización adaptativa del modelo de lenguaje, con los rasgos prosódicos propuestos en aquí, se presentan a modo de ejemplo algunas de las frases reconocidas. Se pueden identificar tres tipos de resultados: las mejoras totales, las mejoras parciales y las mejoras indirectas.

Resultados con mejoras totales

En este caso el método da una solución completa y permite un reconocimiento 100 % correcto de la frase.

- En la frase *bxge3113*, cuya transcripción correcta es:

  `Dime el nombre de los mares que bañan la Comunidad de Andalucía.`

  Mientras que el reconocedor sin información prosódica reconoce:

  `Dime el nombre de los mares que baña la Comunidad de Andalucía.`

  El reconocedor con prosodia reconoce:

  `Dime el nombre de los mares que bañan la Comunidad de Andalucía.`

  lo que es correcto, esto se debe a que la palabra `baña` es penalizada en la red y la hipótesis de `bañan` pasa a una tener mayor probabilidad.

- Una corrección similar se realiza en *euge0139*, cuya transcripción correcta es:

  `Dígame si hay algún río que pase por tres Comunidades Autónomas.`

  Mientras que el reconocedor sin prosodia reconoce:

  `Dígame segundo río que pase por tres Comunidades Autónomas.`

  El reconocedor con prosodia reconoce:

  `Dígame si hay algún río que pase por tres Comunidades Autónomas.`

  esto es correcto y se debe a que la palabra `segundo` es penalizada.

- En la frase *ruge0199*, cuya transcripción correcta es:

  `Dime los nombres de los ríos con más de cien kilómetros de longitud.`

  Mientras que el reconocedor sin prosodia reconoce:

  `Dime el nombre de los ríos con más de cien kilómetros longitud.`

  El reconocedor con prosodia reconoce correctamente:

  `Dime los nombres de los ríos con más de cien kilómetros de longitud.`

	Energía			F_0		
Palabras	Mínima	Media	Máxima	Mínima	Media	Máxima
baña	0	1	1	0	0	0
caudal	1	1	0	0	0	0
cuales	1	1	1	0	1	1
cuantos	1	1	1	0	0	0
dime	1	1	1	1	1	1
ebro	1	1	1	1	0	0
largo	1	1	1	0	0	0
leon	1	1	0	0	0	0
madrid	1	0	0	0	0	0
mares	1	1	1	0	0	0
mayor	1	0	0	0	0	0
menos	1	1	0	0	0	0
metros	1	1	0	1	1	0
nace	1	1	1	0	0	0
nacen	1	1	1	0	0	0
nombre	0	1	1	1	0	0
pasa	1	1	1	0	0	0
pasan	1	1	1	0	0	0
tajo	1	1	1	0	0	0
tiene	1	0	0	0	0	0
tienen	1	0	0	0	0	0
todos	1	0	0	0	0	0
una	0	1	1	0	0	0
longitud	1	1	0	0	0	0
segundo	1	0	0	0	0	0
valencia	0	1	1	0	0	0
superior	1	1	1	0	0	0
atlantico	1	0	0	0	0	0
autonoma	1	0	0	0	0	0
cantabrico	0	1	1	0	0	0
desemboca	0	1	1	0	0	0
desembocan	0	0	1	0	0	0
kilometros	1	0	0	0	0	0
valenciana	0	0	0	0	1	0
comunidades	1	0	0	0	0	0

Tabla 4.4: Tabla de clasificación prosódica

penaliza a `nombre` y soluciona la extensión de `kilómetros`, que con una buena segmentación permite incorporar a la palabra `de`.

- En la frase *vlge0251*, cuya transcripción correcta es:

  `Número de ríos con una longitud superior a los quinientos kilómetros.`

  Mientras que el reconocedor sin prosodia reconoce:

  `Nombre de ríos con una longitud superior a los quinientos kilómetros.`

  El reconocedor con prosodia reconoce correctamente:

  `Número de ríos con una longitud superior a los quinientos kilómetros.`

  Debido a que la palabra `nombre` está clasificada, se puede corregir, ya que no se corresponde la estructura hallada con su clasificación.

Resultados con mejoras parciales

En este caso el método da una solución en cuanto propone qué hipótesis de palabra no es correcta, aunque no corrige todos los errores de reconocimiento en la frase.

- Para *ilge0071*, cuya transcripción correcta es:

  `Hay algún río que nazca y desemboque en la misma Comunidad.`

  Mientras que el reconocedor sin prosodia reconoce:

  `Caudal de un río que nazca y desemboque en la misma Comunidad.`

  El reconocedor con prosodia reconoce:

  `Cada algún río que nazca y desemboque en la misma Comunidad.`

  Aquí se ve que el descartar la hipótesis de `caudal` ayuda a corregir otra palabra y un error por inserción.

- Para el archivo *ikge0064*, cuya transcripción correcta es:

  `En qué comunidad se halla el Duero?`

  Mientras que el reconocedor sin prosodia reconoce:

  `Que en que comunidad se halla el Ebro?`

  El reconocedor con prosodia reconoce:

  `Que en que comunidad se halla el Duero?`

  Aquí se ve que al penalizar la hipótesis de `Ebro` resulta más probable la palabra `Duero`.

- Para el archivo *xuge3040*, cuya transcripción correcta es:

  `En que comunidad desemboca el río ebro.`

  Mientras que el reconocedor sin prosodia reconoce:

  `Dime que comunidades. que desemboca el río ebro.`

  El reconocedor con prosodia reconoce:

  `En que comunidades. que desemboca el río ebro.`

  Debido a que la palabra `Dime` está clasificada, se puede reparar el error.

Resultados con mejoras indirectas

Entre estos casos, por ejemplo, se incluyen situaciones en que, si bien no se corrige una palabra prosódicamente incorrecta, se selecciona correctamente entre varias hipótesis de la misma palabra pero con distintos tiempos de inicio y fin, lo que indirectamente beneficia al resto de la red de palabras.

- Para la frase *nxge3161*, cuya transcripción correcta es:

  `Lugar donde desemboca el Jucar.`

  Mientras que el reconocedor sin prosodia reconoce:

  `Lugar donde desemboca Jucar`

  El reconocedor con prosodia reconoce:

  `Lugar donde desemboca el Jucar.`

  El reconocedor con prosodia reconoce correctamente, dado que la penalización de hipótesis mal ubicadas temporalmente de `desemboca` permite que se haga más probable la hipótesis que tiene a esta palabra con la ubicación correcta en la frase y así deja lugar para insertar la palabra `el`.

- Para el archivo *byge3122*, cuya transcripción correcta es:

  `Dime la comunidad en la que desemboca el río turia.`

  Mientras que el reconocedor sin prosodia reconoce:

  `Dime la comunidad de la que desemboca río turia.`

  El reconocedor con prosodia reconoce:

  `Dime la comunidad de la que desemboca el río turia.`

  Aquí no se corrige toda la frase, sin embargo se ve que al penalizar la hipótesis de `desemboca` mal ubicada en el tiempo puede corregir supresión de la palabra `el`. Es un caso similar al anterior.

- Para el archivo *nyge3082*, cuya transcripción correcta es:

  `Tienen la misma longitud y el mismo caudal el río guadiana y el río guadalquivir.`

  Mientras que el reconocedor sin prosodia reconoce:

  `Tiene la misma longitud tiene mas caudal del río guadiana y el río guadalquivir.`

  El reconocedor con prosodia reconoce:

  `Tiene la misma longitud cien mismo caudal del río guadiana y el río guadalquivir.`

  Si bien está lejos de corregir toda la frase, se ha corregido la palabra mismo. Esto se debe a que la palabra `tiene` está clasificada y la estructura prosódica que se encontró en ese lugar de la frase no coincide con ésta.

5

Conclusiones y trabajos futuros

Se ha presentado un método que analiza las señales de voz y produce una caracterización de las distintas palabras en base a histogramas, discriminadas por grupos según su separación silábica y luego clasificadas por su estructura prosódica. Este método se desarrolló de manera práctica y se validó experimentalmente. De los ejemplos presentados en las secciones 2.1.3 y 4.1, se hace evidente que esta caracterización distingue a las palabras y brinda esa información relevante buscada.

El Capítulo 3 de cuenta del análisis, diseño e implementación del sistema con orientación a objetos y del que se concluye que se ha construido una herramienta flexible que permite la aplicación del método propuesto.

La fase de implementación y prueba ha arrojado buenos resultados, se puede apreciar la relevancia de la incorporación del método al sistema de RAH en el Capítulo 4 donde los resultados han sido valorados y discutidos.

Como trabajos futuros se preve extender la técnica de histogramas a la combinación de éstos, tomando de a dos o más variable prosódicas y formando nuevos histogramas a partir de todas las combinaciones posibles de medidas de estas variables. Otra propuesta para clasificar mejor a las palabras es extraer más información prosódica de las palabras, como *cadencias, anticadencias y mesetas* de entonación [14]. También es necesario, para mejorar los efectos de la penalización en el reconocimiento, introducir las probabilidades relacionadas con la prosodia directamente en el algoritmo de Viterbi (y no como un pos-procesamiento basado en la red de palabras). Como tarea pendiente se planea extender los experimentos de reconocimiento a otros corpus

de habla, con diferentes acentos regionales, en condiciones de ruido y, más a largo plazo, realizar estos mismos estudios en otros idiomas.

Podría pensarse también en el desarrollo de una interfaz gráfica para el sistema y así simplificar su utilización a usuarios inexpertos.

Bibliografía

[1] D. H. Milone and A. J. Rubio, "Prosodic and accentual information for automatic speech recognition," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, pp. 321–333, Julio 2003.

[2] E. M. Albornoz and D. H. Milone, "Construcción de patrones prosódicos para el reconocimiento automático del habla," in *34ta Jornada argentina de informática e investigacion operativa*, (Rosario, Argentina), pp. 225–236, September 2005. Simposio ASAI.

[3] K. Chen, S. Borys, and M. Hasegawa-Johnson, "Prosody dependent speech recognition with explicit duration modelling at intonational phrase boundaries," in *Proceedings of the 8th European Conference on Speech Communication and Technology*, (Geneva), 2003.

[4] A. Marzal, "Reconocimiento automático del habla," tech. rep., España, Jun. 2002. Curso de RAH.

[5] F. Jelinek, *Statistical Methods for Speech Recognition.* Cambrige, Masachussets: MIT Press, 1999.

[6] A. V. Oppenheim and A. S. Wilsky, *Señales y Sistemas.* Prentice Hall, 1998.

[7] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals.* New York: Macmillan Publishing, 1993.

[8] A. M. B. Manrique, *Manual de Fonética Acústica.* Buenos Aires: Hachette, 1980.

[9] D. H. Milone, *Información Acentual para el Reconocimiento Automático del Habla.* PhD thesis, Universidad de Granada, Granada, España, Mar. 2003. Memoria de Tesis.

[10] D. H. Milone, "Fundamentos del reconocimiento automático del habla," tech. rep., UNL, Argentina, Feb. 2004.

[11] D. H. Milone and H. L. Rufiner, *Introducción a las señales y los sistemas discretos.* Argentina: UNER, 2004.

[12] J. G. Proakis and D. G. Manolakis, *Tratamiento digital de señales*. Madrid: Prentice Hall, 1998.

[13] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*. Prentice Hall, 1975.

[14] A. Quilis, *Tratado de Fonología y Fonética Españolas*. Biblioteca Románica Hispánica, Madrid: Editorial Gredos, 1993.

[15] E. A. Llorach, *Gramática de la Lengua Española*. Real Academia Española. Colección Nebrija y Bello, Madrid: Editorial Espasa Calpe, 1999.

[16] J. M. G. Almiñana, *Modelización de Patrones Melódicos del Español para la Síntesis y el Reconocimiento del Habla*. Barcelona: Servei de Publicacions de la Universitat Autònoma de Barcelona, Facultad de Filosofia i Lletres, Departament de Filologia Espanyola, 1991.

[17] J. Buckow, A. Batliner, R. Huber, E. Nöth, V. Warnke, and H. Niemann, "Dovetailing of acoustic and prosody in spontaneous speech recognition," in *Proceedings of 5th International Conference on Spoken Language Processing*, 1998. Prosody and Emotion 2.

[18] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "Verbmobil: The use of prosody in the linguistic components of a speech understanding system," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 5, pp. 519–532, 2000.

[19] S. Rajendran and B. Yegnanarayana, "Word boundary hypothesization for continuous speech in Hindi based on F0 patterns," *Speech Communication*, vol. 18, pp. 21–46, 1996.

[20] K. Hirose and K. Iwano, "Accent type recognition and syntactic boundary detection of japanese using statistical modeling of moraic transitions of fundamental frequency contours," in *Proceedings of the IEEE 23rd International Conference on Acoustics, Speech and Signal Processing*, vol. 1, (Seattle), pp. 25–28, 1998.

[21] S.-W. Lee and K. Hirose, "Dynamic beam-search strategy using prosodic-syntactic information," in *Workshop on Automatic Speech Recognition and Understanding*, pp. 189–192, 1999.

[22] A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," in *Proceedings of the 7th European Conference on Speech Communication and Technology*, vol. 1, pp. 311–314, 1999.

[23] D. H. Milone, A. J. Rubio, and R. López-Cózar, "Modelos de lenguaje variantes en el tiempo," in *Memorias del XXIV Congreso Nacional de Ingeniería Biomédica*, (Oaxtepec, México), SOMIB, Oct. 2001.

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

[25] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.1).* Cambridge University Engineering Department., Cambridge, Inglaterra, Dic. 2001.

[26] H. Ney and S. Ortmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 64–83, 1999.

[27] A. M. Noll, "Cepstrum pitch determination," *The Journal of the Acoustical Society of America*, vol. 41, pp. 179–195, Feb. 1967.

[28] I. Sommerville, *Ingeniería del software.* México: Addison-Wesley Iberoamericana, 6ta ed., 2002.

[29] J. Rumbaugh, I. Jacobson, and G. Booch, *El Lenguaje Unificado de Modelado.* España: Addison Wesley, 1999.

[30] J. Rumbaugh, I. Jacobson, and G. Booch, *Manual de Referencia: El Lenguaje Unificado de Modelado.* España: Addison Wesley, 2000. Rational Software Corporation.

[31] A. Moreno, D. Poch, A. Bonafonte, E.Lleida, J.Llisterri, J.B.Marino, and C. Nadeu, "Albayzin speech data base: design of the phonetic corpus," in *Proceedings of the 2th European Conference of Speech Communication and Technology*, (Berlin), pp. 175–178, September 1993.

[32] P. L. Torres, "Desarrollo de software orientado a objeto usando uml," tech. rep., Universidad Politécnica de Valencia, Departamento Sistemas Informáticos y Computación, 2003. www.dsic.upv.es/∼uml.

Apéndice A

Orientación a Objetos

La Orientación a Objetos (OO) presenta una proximidad de los conceptos de modelado respecto de las entidades del mundo real [32], esto

- Mejora la captura y validación de requisitos
- Acerca el espacio del problema y el espacio de la solución

El modelado integra las propiedades estáticas y dinámicas del ámbito del problema, entonces

- Facilita construcción, mantenimiento y reutilización

Los conceptos comunes de modelado durante el análisis, diseño e implementación

- Facilita la transición entre distintas fases
- Favorece el desarrollo iterativo del sistema
- Disipa la barrera entre el **qué** y el **cómo**

A.1. Fundamentos del modelado OO

La OO es un paradigma que depende de ciertos principios fundamentales. Para éstos se definen conceptos de los objetos y características que hacen a éstos. A continuación se explican algunos conceptos:

A.1.1. Objetos

Un **Objeto** es una entidad discreta con límites bien definidos y con identidad, es una unidad atómica que encapsula estado y comportamiento. La encapsulación en un objeto permite una alta **cohesión** y un bajo **acoplamiento**.

Objeto = Identidad + Estado + Comportamiento

- Un objeto puede caracterizar una entidad *física* con valor específico en determinado tiempo o *abstracta* que tiene distintos valores a lo largo del tiempo (ecuación matemática)
- El Modelado de Objetos permite representar el ciclo de vida de los objetos a través de sus interacciones
- El estado está representado por los valores de los atributos
- Un atributo toma un valor en un dominio concreto
- Algunos objetos están hechos de otros objetos, esto es, un objeto es un agregado de otros objetos.

Un objeto contiene datos y operaciones que operan sobre los datos. Un sistema construido con objetos que no cumplen estos requisitos, no es un sistema verdaderamente orientado a objetos. Éstos objetos se denominan **objetos degenerados** y se pueden distinguir dos tipos:

- Un objeto sin datos (que sería lo mismo que una biblioteca de funciones)
- Un objeto sin operaciones, sólo con operaciones del tipo crear, recuperar, actualizar y borrar (que se correspondería con las estructuras de datos tradicionales)

A.1.2. Identidad

Cada objeto posee un Object Identifier (oid). El oid establece la identidad del objeto y tiene las siguientes características:

- Constituye un identificador único y global para cada objeto dentro del sistema
- Es determinado en el momento de la creación del objeto
- Es independiente de la localización física del objeto, es decir, provee completa independencia de localización

- Es independiente de las propiedades del objeto, lo cual implica independencia de valor y de estructura
- No cambia durante toda la vida del objeto. Además, un oid no se reutiliza aunque el objeto deje de existir
- No se tiene ningún control sobre los oid y su manipulación resulta transparente

Sin embargo, es preciso contar con algún medio para hacer referencia a un objeto utilizando referencias del dominio (valores de atributos)

A.1.3. Estado

- El estado evoluciona con el tiempo
- Algunos atributos pueden ser constantes
- El comportamiento agrupa las competencias de un objeto y describe las acciones y reacciones de ese objeto
- Las operaciones de un objeto son consecuencia de un estímulo externo representado como mensaje enviado desde otro objeto

A.1.4. Comportamiento

- Los mensajes navegan por los enlaces, a priori en ambas direcciones
- Estado y comportamiento están relacionados
- Ejemplo: no es posible aterrizar un avión si no está volando. Está volando como consecuencia de haber despegado del suelo

A.1.5. Persistencia

- La persistencia de los objetos designa la capacidad de un objeto trascender en el espacio/tiempo
- Se pueden reconstruir, es decir, se toman de la memoria secundaria para utilizarlo en la ejecución (materialización del objeto)
- Los lenguajes OO no proponen soporte adecuado para la persistencia, la cual debería ser transparente, un objeto existe desde su creación hasta que se destruya

A.1.6. Comunicación

- El comportamiento global se basa en la comunicación entre los objetos que la componen
- La comunicación se lleva a cabo en base a las llamadas a métodos a través de las operaciones con sus parámetros específicos
- Un sistema informático puede verse como un conjunto de objetos autónomos y concurrentes que trabajan de manera coordinada en la consecución de un fin específico

A.1.7. Mensaje

- La unidad de comunicación entre objetos se llama mensaje
- El mensaje es el soporte de una comunicación que vincula dinámicamente los objetos que fueron separados previamente en el proceso de descomposición
- Adquiere toda su fuerza cuando se asocia al polimorfismo y al enlace dinámico

A.1.8. Mensaje y Estímulo

- Un estímulo causará la invocación de una operación, la creación o destrucción de un objeto o la aparición de una señal
- Un mensaje es la especificación de un estímulo

A.1.9. Operaciones y Métodos

- El término "operación" refiere a la especificación de una acción
- El término "método" se utiliza para referirse a la implementación de una operación.
- Una operación se puede especificar indicando su prototipo, la cual incluye el nombre, tipo y valores por defecto de todos los parámetros y (en el caso de funciones) un tipo de retorno.

Apéndice B

UML

El Lenguaje Unificado de Modelado es un lenguaje de modelado visual que se usa para especificar, visualizar, construir y documentar componentes de un sistema de software. Se usa para entender, diseñar, configurar, mantener y controlar la información sobre los sistemas a construir [30].

UML capta la información sobre la estructura estática y el comportamiento dinámico de un sistema. Un sistema se modela como una colección de objetos discretos que interactúan para realizar un trabajo que finalmente beneficia a un usuario externo. El lenguaje de modelado pretende unificar la experiencia pasada sobre técnicas de modelado e incorporar las mejores prácticas actuales en un acercamiento estándar. UML no es un lenguaje de programación. Es un lenguaje de propósito general para el modelado orientado a objetos.

B.1. Modelos y Diagramas

Un **modelo** captura una vista de un sistema del mundo real. Es una abstracción de dicho sistema, considerando un cierto propósito. Así, el modelo describe completamente aquellos aspectos del sistema que son relevantes al propósito del modelo, y a un apropiado nivel de detalle. Es siempre una abstracción a un cierto nivel, captura los aspectos esenciales de un sistema y omite algunos detalles. Las descripciones son su objetivo o significado.

Un **Diagrama** es una representación gráfica de una colección de elementos de modelado, a menudo dibujada como un grafo con vértices conectados por arcos.

- Un proceso de desarrollo de software debe ofrecer un conjunto de modelos que permitan expresar el producto desde cada una de las perspectivas de interés
- El código fuente es el modelo más detallado del sistema (y además es ejecutable). Sin embargo, se requieren otros modelos.
- Cada modelo es completo desde su punto de vista del sistema, sin embargo, existen relaciones de trazabilidad entre los diferentes modelos

B.2. Diagramas de UML

Los diagramas expresan gráficamente partes de un modelo. Los diagramas definidos en UML son:

- Diagrama de Casos de Uso
- Diagrama de Clases
- Diagrama de Objetos
- Diagramas de Comportamiento
  - Diagrama de Estados
  - Diagrama de Actividad
  - Diagramas de Interacción
    - Diagrama de Secuencia
    - Diagrama de Colaboración
- Diagramas de implementación
  - Diagrama de Componentes
  - Diagrama de Despliegue

A continuación se comentan algunos diagramas [30].

B.2.1. Diagramas de Casos de Uso

Dan una idea tanto al usuario (para entenderlo) como al analista/programador (para implementarlo) de como se comporta el sistema.

- Casos de Uso es una técnica para capturar información de cómo un sistema o negocio trabaja, o de cómo se desea que trabaje
- No pertenece estrictamente al enfoque orientado a objeto, es una técnica para captura de requisitos

En la Figura B.1 puede verse un ejemplo para manejar un televisor. El usuario interactúa con el televisor para encenderlo, apagarlo, cambiar el volumen y cambiar el canal. Se pueden apreciar los límites del sistema marcados por el recuadro exterior. Se ven dos relaciones internas: *«include»* (inclusión) y *«extend»*(extensión). La inclusión de un caso de uso también se conoce como *usar* un caso de uso, ésto nos permite ver en detalle que actividades se llevan a cabo en un caso de uso (siempre se realizan). La extensión permite especificar los casos de uso que se realizarán, siempre y cuando se cumplan dichas condiciones. En el ejemplo, *cambiar canal (secuencial)* debe verificar que el canal esté activo (no borrado) para poder avanzar secuencialmente, si está borrado saltará al próximo no borrado. Para el caso de uso *cambiar volumen* se extiende un caso de uso que sólo se llevará a cabo y comunicará al usuario que se ha llegado al mínimo o al máximo volumen posible.

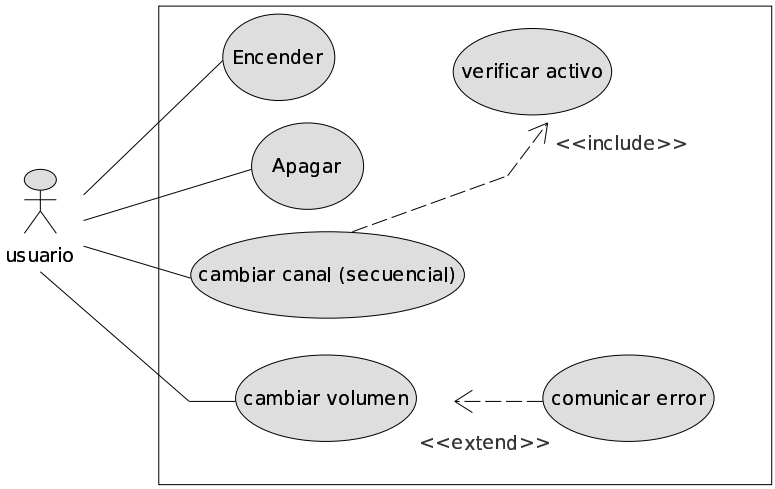


FIGURA B.1: CASO DE USO PARA MANEJO DE UN TELEVISOR

B.2.2. ESCENARIOS

Si bien los escenarios son herramientas descriptivas y no un tipo de diagramas UML, su uso junto a éstos es usual y muy relevante. La diferencia con CU es que los escenarios plantean el contexto global y no sólo la interfaz. De estos es posible derivar objetos y especificaciones orientadas a objetos. Al estar basados en interacciones reales o predichas, los usuarios los encuentran más simples que dar una definición de la funcionalidad y, por otro lado, al exponer un rango de posibles interacciones pueden llegar a revelar determinadas facilidades que se requieren del sistema. Además, los escenarios pueden ser pensados como historias que explican cómo el sistema es usado y son útiles para agregar detalles a un primer bosquejo de requerimientos. Son una clara herramienta descriptiva que permite, entre otras posibilidades:

- Una descripción del estado del sistema antes del escenario a describir.

- El flujo normal de eventos del escenario.
- Las excepciones al flujo normal.
- Información sobre otras actividades que pueden pasar en paralelo.
- Una descripción del sistema después de completar el escenario.

B.2.3. Diagramas de Interacción

- Describe secuencias de intercambios de mensajes entre los roles que implementan el comportamiento de un sistema
- Muestran cómo se comunican los objetos en una interacción

B.2.4. Diagramas de Secuencia

Un DS es un diagrama de interacción que destaca la ordenación temporal de los mensajes. Proporciona una idea de los aspectos dinámicos del sistema, del flujo de información y da una base para la construcción del sistema ejecutable.

El DS presenta un conjunto de objetos y los mensajes enviados y recibidos por ellos. Los objetos suelen ser instancias con nombres o anónimas de clases, pero también pueden representar instancias de otros elementos, tales como colaboraciones, componentes y nodos.

- Muestra la secuencia de mensajes entre objetos durante un escenario concreto
- Cada objeto viene dado por una barra vertical
- El tiempo transcurre de arriba hacia abajo
- Cuando existe demora entre el envío y la atención se puede indicar usando una línea oblicua

En la Figura B.2 se puede ver un ejemplo para este tipo de diagrama. Se destacan aquí los objetos y sus líneas de vida, la creación y destrucción de objetos, los mensajes y sus retornos, etc.

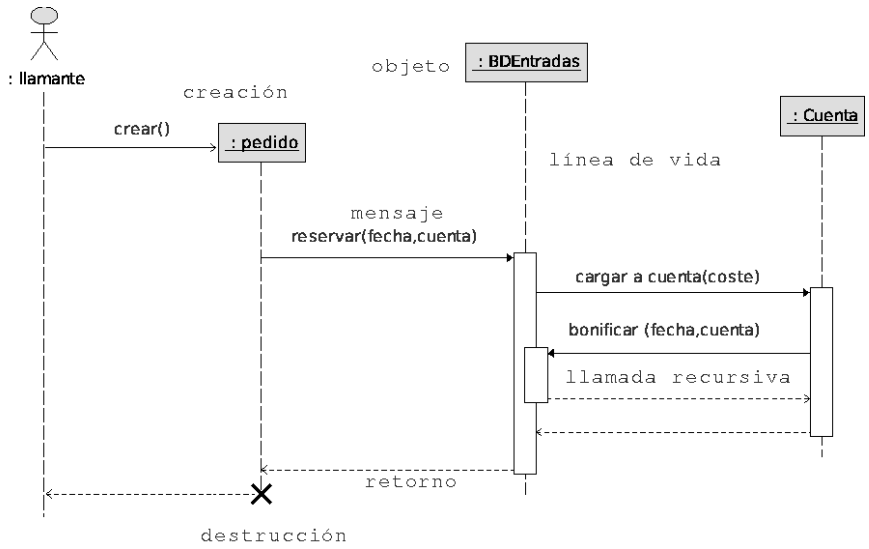


FIGURA B.2: EJEMPLO DE UN DIAGRAMA DE SECUENCIA

B.2.5. DIAGRAMA DE CLASES

Las clases son los bloques de construcción más importantes de cualquier sistema OO e implementan una o más interfaces. Se pueden utilizar para representar cosas que sean software, hardware o puramente conceptuales. Las clases son una abstracción de las cosas, a partir de estas se crean los objetos que son instancias de éstas.

- El Diagrama de Clases es el diagrama principal para el análisis y diseño
- Un diagrama de clases presenta las clases del sistema con sus relaciones estructurales y de herencia
- La definición de clase incluye definiciones para atributos y operaciones
- El modelo de casos de uso aporta información para establecer las clases, objetos, atributos y operaciones
- En UML, un objeto se representa por un rectángulo con un nombre subrayado

En la Figura B.3 puede observarse, a modo de ejemplo, el diagrama de la clase *Persona*, sus atributos y sus operaciones. Los símbolos a la izquierda de los atributos y de las operaciones están relacionado con los niveles de encapsulamiento: (−) *privado*, (+) *público* y (#) *protegido*. A la derecha de los nombres de los atributos se indican: dos puntos seguidos del tipo de dato del atributo y, opcionalmente, un signo igual (=) y luego el valor por defecto del atributo. Para las operaciones se indican entre paréntesis los parámetros

que recibe ésta, pueden ser de entrada, de salida o de entrada/salida. En caso de ser una función, la operación, tiene a la derecha dos puntos seguidos del tipo de dato del valor de retorno.

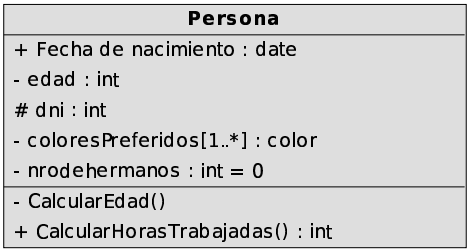
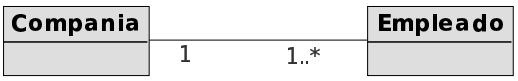


FIGURA B.3: DIAGRAMA DE LA CLASE PERSONA

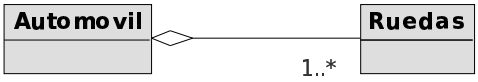
Las clases se pueden relacionar (ser asociadas con) con otras de diferentes maneras:

Generalización. Es el tipo de asociación que en OO se denomina *herencia*, una clase puede heredar los atributos y operaciones de otra. En UML está relación se representa como:

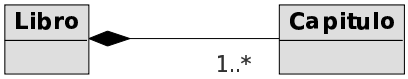
Asociación. Representa una relación entre clases, puede ser unidireccional o bidireccional (en relación al envío de mensajes) y los valores de multiplicidad (valor_mínimo[..valor_máximo]) expresan cuantos elementos de un lado pueden relacionarse con uno del otro lado de la asociación (el '*' como valor_máximo indica infinito). En UML ésto puede verse como:



Agregación. Es un tipo de asociación en la que los participantes no tienen igual status, describe como la clase que toma el papel de *el todo*, está compuesta de otras clases, las que son denominadas *partes*. La clase que actúa como *el todo* siempre tiene una multiplicidad uno. Un ejemplo en UML podría ser:



Composición. Son asociaciones que representan agregaciones **muy fuertes**. Las partes no pueden existir sin el todo y si éste es destruido, las partes se destruyen también. En UML se representa con un rombo sólido:



Este texto fue escrito con LaTeX, tipeado en "Kile". Las Figuras de los diagramas UML se realizaron con el modelador de UML "Umbrello" y las demás Figuras se generaron con los software gnuplot, Dia y Gimp.

Marzo de 2006