UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Modelado de estructuras prosódicas para el reconocimiento automático del habla

Enrique Marcelo Albornoz

Tesis remitida al Comité Académico del Doctorado
como parte de los requisitos para la obtención
del grado de
DOCTOR EN INGENIERIA
Mención Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2011

Comisión de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria,
Paraje "El Pozo", S3000, Santa Fe, Argentina.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz; "Modelado de estructuras prosódicas para el reconocimiento automático del habla"
Universidad Nacional del Litoral, 2011.

Doctorado en Ingeniería
Mención Inteligencia Computacional, Señales y Sistemas

Título de la obra:

Modelado de estructuras prosódicas para el reconocimiento automático del habla

Autor: Enrique Marcelo Albornoz

Director: Dr. Diego Humberto Milone

Codirector: Dr. Hugo Leonardo Rufiner

Lugar: Santa Fe, Argentina

Palabras Claves:

Modelado prosódico,
Reconocimiento automático del habla,
Modelo de lenguaje, redes de palabras,
Reconocimiento de emociones en el habla,
Clasificador jerárquico.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz; "Modelado de estructuras prosódicas para el reconocimiento automático del habla"
Universidad Nacional del Litoral, 2011.

Agradecimientos

Deseo expresar un muy profundo agradecimiento a mi director, Dr. Diego Milone, por acompañarme tan dedicadamente todos estos años y por ofrecerme sus conocimientos y experiencias para lograr esta Tesis. Por brindarme su confianza, sus consejos, su docencia constante y su amistad. De la misma forma quiero agradecer a mi codirector, Dr. Leonardo Rufiner, por contribuir con mi formación compartiendo sus experiencias, sus puntos de vista y su amistad.

Quisiera agradecer también al Dr. Ramón López-Cózar Delgado por su cordialidad, predisposición y sus valiosos puntos de vista durante mi estancia de investigación. Y a los integrantes del departamento de Lenguajes y Sistemas Informáticos de la Universidad de Granada (España) por su cordialidad y colaboración.

Al Dr. Leandro Di Persia, que junto al Dr. Diego Milone, fue promotor de mi iniciación en la investigación, siempre dispuesto a colaborar y a brindar su conocimiento. A mis amigos del **sinc**(i) Leandro y César por su compañía de tantos años,por brindarme siempre su ayuda e interesantes discusiones. También quisiera agradecer a todo el gran grupo humano del Centro **sinc**(i) por su constante apoyo, predisposición y amistad: Diego, Carlos, María Eugenia, Cecilia, Leonardo, Federico, Matías y Maximiliano.

Deseo agradecer a mi familia. En especial a mis padres Rubén y Silvia, por ser mis ejemplos de vida, por su constante cariño y apoyo a lo largo de toda mi vida. A mis hermanos por acompañarme siempre. A mis tíos, primos y a mi abuela Chicha por todo su amor incondicional. A mis amigos por estar siempre presentes y brindarme su afecto.

También expreso mi agradecimiento a:

- **sinc**(i): Centro de Investigación y Desarrollo en Señales, Sistemas e Inteligencia Computacional
- Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral
- Consejo Nacional de Investigaciones Científicas y Técnicas
- Departamento de Lenguajes y Sistemas Informáticos, E.T.S. de Ingenierías Informática y de Telecomunicación - Universidad de Granada (España)

Resumen

El término prosodia se utiliza para denominar a determinadas magnitudes físicas que pueden ser medidas en las señales de voz, como ser la energía, la frecuencia fundamental, la duración de sus unidades básicas, entre otras. Las características prosódicas de las señales de voz presentan información valiosa que, mediante métodos de reconocimiento de patrones, permiten la identificación y clasificación de diversos aspectos relativos a la producción de la voz. Algunas de las aplicaciones donde esta información tiene relevancia son el reconocimiento automático del habla (RAH), el reconocimiento de emociones y la identificación de hablante, etc.

El RAH es un área de estudio multidisciplinar que tiene como objetivo comprender el proceso de producción y de percepción del habla. El objetivo final es lograr que una máquina posea la capacidad de reconocer las palabras pronunciadas por un locutor e inclusive entender su significado, considerando cualquier hablante en cualquier entorno. Desde un punto de vista ingenieril, se podría decir que el RAH consiste en el diseño y construcción de un sistema que reconoce automáticamente las transcripciones asociadas a las elocuciones humanas. Los sistemas de RAH actuales utilizan modelos ocultos de Markov para realizar una caracterización fonética-acústica del habla. Estos modelos utilizan parametrizaciones de la señal que no contemplan la información prosódica de forma explícita. Para el Español no se ha adoptado un modelo estándar que de cuenta, en el discurso continuo, de la existencia de una relación clara y bien definida entre la acentuación establecida por las reglas ortográficas y las manifestaciones prosódicas del habla. En este trabajo se propone pasar a un segundo plano la información provista por la acentuación definida según las reglas ortográficas y hallar relaciones claras entre los rasgos prosódicos y las palabras que se pronuncian. De esta manera se busca definir una nueva forma de clasificar las prominencias acentuales del idioma. Para ésto, se definen modelos de palabras que las categorizan según la información prosódica que éstas presentan y se propone la incorporación de clasificadores prosódicos de hipótesis acústicas en la red de palabras generada por un sistema de RAH estándar. Además, y como resultado de los análisis del desempeño del método mencionado, se realiza un estudio profundo acerca de la localización acústica de las hipótesis de palabras (tanto

verdaderas como falsas) propuestas por un reconocedor con el fin de definir un nuevo corpus de datos. Este corpus contiene información precisa de las secuencias acústicas, luego asociadas a palabras, que presentan problemas al reconocedor. Esta información es usada para generar clasificadores prosódicos especializados para cada palabra.

Por otra parte, si se consideran las interacciones humanas de una forma general puede verse que existen muchas maneras en las que se intercambia información (voz, gestos corporales, gestos faciales, etc.). Un mensaje de voz, a través del que las personas se comunican, tiene una gran cantidad de información que se interpreta de forma natural e implícita. Esta información puede ser expresada o percibida en la entonación, el volumen y la velocidad de la voz, entre otros rasgos prosódicos. El estado emocional del hablante está estrechamente relacionado con esta información, debido a que puede presentarse implícitamente en estas variables y ésto motiva su estudio. En los últimos años, el reconocimiento de emociones se ha convertido en un espacio de investigación multidisciplinar que ha recibido gran interés. El reconocimiento automático del estado emocional tiene como objetivo lograr una interacción más natural entre seres humanos y máquinas. En esta Tesis también se aborda este tema. Inicialmente se realiza una exploración de clasificadores basados en mezclas de Gaussianas y modelos ocultos de Markov, con coeficientes ceptrales en escala de mel en la etapa de extracción de características. Se realizan análisis sobre las características prosódico-acústicas de las emociones que permiten agruparlas de forma no supervisada. De esta manera, se generan modelos de clasificación jerárquicos que contemplan los agrupamientos de emociones encontrados y permiten mejorar el rendimiento en relación a clasificadores estándar.

Prefacio

El trabajo de investigación que se presenta en esta Tesis se enmarca en el área de reconocimiento automático del habla. Las diferentes aplicaciones en este campo 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, inteligencia artificial, etc. Las aplicaciones en las que el reconocimiento automático del habla tiene incumbencia son muchas, entre las que 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, el acceso a cajeros automáticos, etc. En los últimos años, muchos investigadores han dedicado su esfuerzo a mejorar los sistemas de reconocimiento automático del habla. Cada nivel de análisis del habla se ha convertido en un campo de investigación, y en cada uno se ha provisto información importante al reconocedor. Las características prosódicas que se presentan en las señales de voz han ganado terreno en los análisis del habla y encontrar la mejor forma de incorporar esta información al reconocimiento automático del habla es el problema que se plantea en los análisis actuales.

En las comunicaciones humanas existen diversas formas de transmitir información. Si se considera la contribución de cada parte en una comunicación para los receptores del mensaje, se puede mencionar que las palabras sólo representan un 7 %, la información paralingüística un 38 % (prosodia, calidad de la voz, emociones, etc.) y la comunicación no verbal un 55 % [1]. Esta clasificación considera a la expresividad emocional como parte de la información paralingüística de la comunicación y en este trabajo sólo se considerará como tal. Sin embargo, existen numerosos estudios que demuestran su relación con las palabras y con la comunicación no verbal. Existe mucho interés en investigar y desarrollar aplicaciones relacionadas con el reconocimiento de emociones, mientras que el estudio de su vínculo con la prosodia se ha hecho muy atractivo para los investigadores.

En este trabajo se pretende obtener un método que permita el análisis de señales de voz con sus rasgos prosódicos y brinde la posibilidad de incorporar este tipo de

análisis en sistemas de reconocimiento automático del habla y de reconocimiento de emociones. En ambos casos, se pretende desarrollar un sistema que mejore el rendimiento en relación a los reconocedores actualmente conocidos, basado en análisis de las características prosódicas.

OBJETIVOS

El objetivo general de esta Tesis es hallar información prosódica relevante en las señales de voz, modelarla y/o utilizarla de forma explícita, con el fin de mejorar el desempeño de los sistemas de reconocimiento basados en la voz.

Las aplicaciones elegidas para explorar la utilidad de la información prosódica son el reconocimiento automático del habla propiamente dicho y el reconocimiento de emociones. En este marco se proponen diversas técnicas de análisis de señales de voz y de reconocimiento de patrones, así como la forma más efectiva de emplearlas para la clasificación.

OBJETIVOS ESPECÍFICOS PARA RAH

- Hallar relaciones claras entre las manifestaciones físicas más importantes de la prosodia (frecuencia fundamental (F_0), energía, extensión del nucleo vocálico) y las palabras que se pronuncian.
- Definir formas nuevas de clasificar las prominencias acentuales del idioma.
- Analizar las mejores formas de utilizar esta información en el modelado estadístico del lenguaje para mejorar el rendimiento de un sistema de reconocimiento automático del habla en discurso continuo.
- Realizar las adaptaciones necesarias para utilizar esta información cuando se trabaja en condiciones de ruido.

OBJETIVOS ESPECÍFICOS PARA RECONOCIMIENTO DE EMOCIONES

- Evaluar técnicas conocidas de reconocimiento de patrones para clasificar emociones.
- Análizar las características espectrales y prosódicas de las señales para caracterizar las emociones.
- Hallar similaridades acústico-prosódicas en los diferentes tipos de emociones que permitan definir nuevos grupos de emociones.

- Utilizar los grupos de emociones para definir nuevos modelos de clasificación y comparar su desempeño respecto de clasificadores tradicionales.

En el primer capítulo se pretende proporcionar una pequeña revisión de los conceptos más importantes relacionados al procesamiento de señales y al reconocimiento de patrones que se utilizan posteriormente. También se introducen los aspectos que serán más relevantes para la investigación en el reconocimiento automático del habla y se presenta una introducción al reconocimiento automático de emociones.

En el Capítulo 2 se propone un método para caracterizar a las palabras según sus estructuras prosódicas y, aplicando modelos de lenguaje, se utiliza esta información para desambiguar hipótesis en el proceso de reconocimiento mediante modelos ocultos de Markov. Se analiza la utilización de este método de clasificación por histogramas y se discuten los resultados obtenidos al incorporar esta información en el sistema de reconocimiento automático del habla. Luego, se propone un nuevo enfoque que permite clasificar las hipótesis obtenidas por un reconocedor estándar durante el proceso de reconocimiento. Éste permite distinguir las hipótesis verdaderas de las falsas para un segmento acústico dado, utilizando información prosódica. El objetivo final es corregir de forma más precisa los errores cometidos por el modelo actual.

El Capítulo 3 se presenta dividido en tres partes principales. En primer lugar, se desarrollan y exponen resultados y discusiones acerca de la utilización de métodos estándar para la clasificación de emociones. En la segunda parte se presentan diferentes análisis prosódico-acústicos de la información presente en las señales emocionales con el fin de identificarlas y agruparlas. Luego, en la última parte, se propone un clasificador jerárquico basado en esta información de grupos que logra un mejor desempeño que los clasificadores tradicionales.

En el Capítulo 4 se exponen las conclusiones respecto de los aportes realizados en las dos líneas que se siguen en la Tesis, así como también los trabajos futuros en torno a estas ideas.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Nociones preliminares y marco conceptual

1.1. Reconocimiento de patrones

El reconocimiento de patrones es un área que explora las diferentes técnicas aplicadas al análisis y clasificación de datos. Esta disciplina se interesa en las técnicas automáticas para descubrir las uniformidades y singularidades en los datos, así como también en los métodos computacionales que permitan clasificarlas [2]. La mayoría de los casos prácticos incluyen 3 etapas bien definidas:

1. el *preprocesamiento* o *etapa de extracción de características* es donde usualmente se analizan y procesan los datos, generalmente de naturaleza continua, para generar vectores discretos de dimensión n con las características más relevantes ($\boldsymbol{x}_i = [x_1, x_2, \ldots x_n]$),

2. la *fase de entrenamiento* o *aprendizaje* es donde se ajustan la estructura y los parámetros del modelo, tal que para una entrada dada $\boldsymbol{x}_i$ se obtenga la salida deseada d_i. Ésto puede plantearse simplificadamente como $d_i = f(\boldsymbol{x}_i)$ donde se deben encontrar la estructura y los parámetros que rigen el comportamiento del sistema f.

3. en la *fase de prueba* se presentan, al modelo entrenado, datos que nunca ha observado para comprobar si éste tiene la capacidad de identificarlos. En general, sólo se posee un pequeño conjunto de datos representativo de las entradas para entrenar el modelo, por lo que es menester que el sistema tenga esta capacidad de *generalización*.

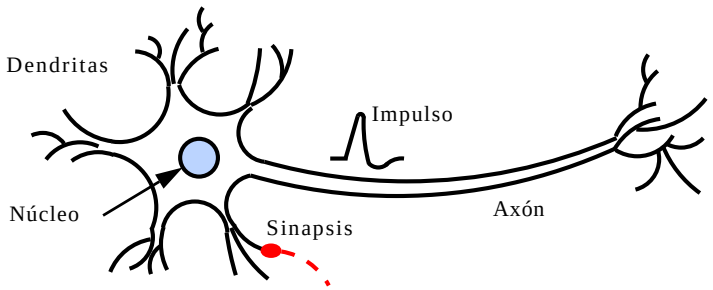

FIGURA 1.1: Esquema simplificado de una neurona.

En este trabajo se utilizan técnicas estándar para el preprocesamiento de señales de voz y son descriptas en la sección 1.2.3. Por otra parte, para la clasificación y agrupamiento de patrones se emplean métodos tomados de la inteligencia computacional y de la estadística.

El concepto de red neuronal artificial está inspirado, como tantos otros, en conceptos biológicos, en este caso las redes neuronales biológicas. Desde hace muchas décadas los investigadores han estudiado la estructura y el comportamiento del cerebro. Si bien aún no se tienen detalles del funcionamiento de las funciones mentales complejas, se ha avanzado mucho en este campo [3]. Se ha determinado que el sistema nervioso del cerebro está formado por alrededor de 10^{11} neuronas interconectadas. Cabe destacar que se han considerado diferentes enfoques para estudiar este sistema biológico, a partir de los cuales se propuso una amplia variedad de redes neuronales. En este capítulo se introducirán aquellas utilizadas en el presente trabajo.

Por otra parte, se encuentran los métodos estadísticos con los que se pueden modelar las características de las señales. La clase de modelo a emplear es dependiente del tipo de señal analizada. Mientras que una señal estacionaria es aquella que mantiene sus características estadísticas en el tiempo, las no-estacionarias presentan variaciones temporales en éstas. Los modelos pueden ser clasificados por su capacidad de capturar o no las dinámicas temporales en las distribuciones de los datos. En este capítulo se presentan dos métodos estadísticos, uno estático y otro dinámico.

1.1.1. PERCEPTRON MULTI-CAPA

Las neuronas son células y pueden comprenderse, desde una perspectiva computacional, como pequeñas unidades de procesamiento. Aunque cada una de éstas funciona mucho más lentamente que una pequeña unidad de silicio, el enorme número de neuronas y la gran cantidad de conexiones logran que el conjunto tenga una gran capacidad de cómputo. Entre los elementos básicos de una neurona podemos destacar un *núcleo*, las *dendritas*, las *sinapsis* y el *axón* (Figura 1.1). Las dentritas son las encarga-

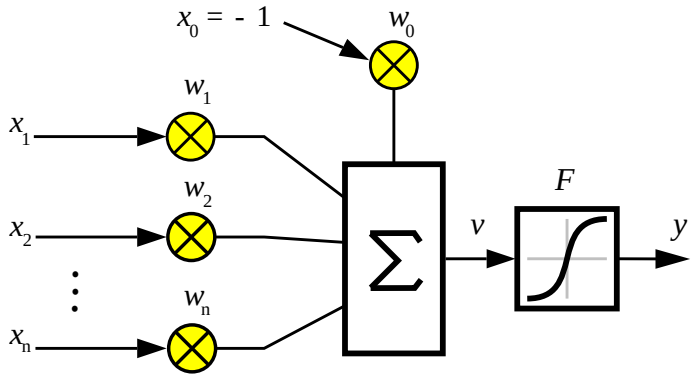

Figura 1.2: Esquema del modelo matemático de una neurona.

das de llevar la información al núcleo de la célula y por el axón la neurona se comunica mediante impulsos. Una neurona recibe información de miles de otras neuronas a partir de la que genera una respuesta. Los enlaces entre neuronas se denominan sinapsis y éstos tienen una importancia proporcional a la actividad que han tenido las células.

A partir de las interpretaciones del funcionamiento individual de las neuronas, se ha desarrollado un modelo matemático que trata de formalizar el comportamiento de estas células y se denomina *perceptron simple*. Si se consideran n entradas a la neurona, se tiene $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ y un vector de pesos asociado $\boldsymbol{w} = [w_1, w_2, \ldots, w_n]$ que representan los pesos sinápticos. En el núcleo se produce una sumatoria de las entradas pesadas que se puede expresar como $\sum_{i=1}^{n} x_i w_i$ y se realiza una diferencia con un cierto umbral θ. De esta forma la salida lineal de la neurona se puede expresar como $v = \sum_{i=1}^{n} x_i w_i - \theta$. Luego, v es procesada de alguna forma para obtener el valor final de salida (activación) de la neurona y puede expresarse como $y = F(v)$. El término θ es considerado un grado de libertad más, término de tendencia, sesgo o *bias*. Usualmente se incorpora a los vectores $\boldsymbol{x}$ y $\boldsymbol{w}$ como $x_0 = -1$ y la sinapsis $w_0 = \theta$ [3]. Estas analogías pueden verse comparando las Figuras 1.1 y 1.2. La expresión final del perceptron simple está dada por

$$y = F\left(\sum_{i=1}^{n} x_i w_i - \theta\right) = F\left(\sum_{i=0}^{n} x_i w_i\right) = F\left(v\right) \tag{1.1}$$

Una vez definido el perceptron, éste es entrenado ajustando sus pesos mediante algoritmos de entrenamiento *supervisado* o *activo* [3]. Básicamente, consiste en presentar los vectores de entrada al perceptron y obtener la salida, que se compara con la salida deseada (d) y se ajustan los pesos en base al error detectado utilizando algún método de gradiente descendiente [3]. La medida del error se define como el error cuadrático

medio

$$E = \frac{1}{2}(d - y)^2, \tag{1.2}$$

y la actualización de los pesos está dada por

$$w_i(t+1) = w_i(t) - \eta \frac{\partial E}{\partial w_i}, \tag{1.3}$$

donde

$$\frac{\partial E}{\partial w_i} = \frac{\partial E}{\partial v} \cdot \frac{\partial v}{\partial w_i} \tag{1.4}$$

y como v es lineal respecto a los pesos, se tiene que $\partial v/\partial w_i = x_i$. Por otra parte, si se considera que en este caso simplemente usamos $F(v) = v$ se tiene que $\partial E/\partial v = -(d-v)$ y la fórmula de actualización de los pesos queda definida por

$$w_i(t+1) = w_i(t) + \eta(d - v)x_i \tag{1.5}$$

donde d es la salida deseada, t el número de iteración y η un factor de ganancia o velocidad de aprendizaje (entre $(0, 1)$).

En algunos casos se requiere que las salidas y sólo tomen dos estados: [0 o 1] o bien [−1 o 1]. Así, la $F(v)$ se define como una función signo y en (1.5) se modifica sutilmente el factor η durante el aprendizaje. Por otra parte, es deseable que la función de activación sea continua y derivable. Generalmente, estas funciones son definidas como sigmoideas porque su derivada es simple [4] (necesaria para $\partial E/\partial v$). Las fórmulas generales de una sigmoidea unipolar y de su derivada respectivamente son

$$f(x) = \frac{1}{1 + e^{-\beta x}}, \tag{1.6}$$

$$f'(x) = \beta f(x)(1 - f(x)), \tag{1.7}$$

donde el parámetro β determina la pendiente.

Para este método está demostrado que, si las clases son linealmente separables, se puede alcanzar $y = d$ [4]. Sin embargo, este modelo neuronal sólo permite dividir el espacio de las entradas de forma lineal en dos partes, por lo que resulta de utilidad muy limitada en aplicaciones prácticas.

Usualmente las aplicaciones requieren que el espacio sea dividido en más de dos partes y de forma no-lineal. Para ésto es posible definir redes de perceptrones o bien *perceptrones multi-capa*. Estas redes están conformadas por conjuntos de perceptrones dispuestos en capas. Los nodos están completamente conectados entre capas (con sus respectivos pesos asociados) y sin conexiones en la misma capa (ver Figura 1.3). El vector de entrada o vector de características alimenta cada uno de los perceptrones de

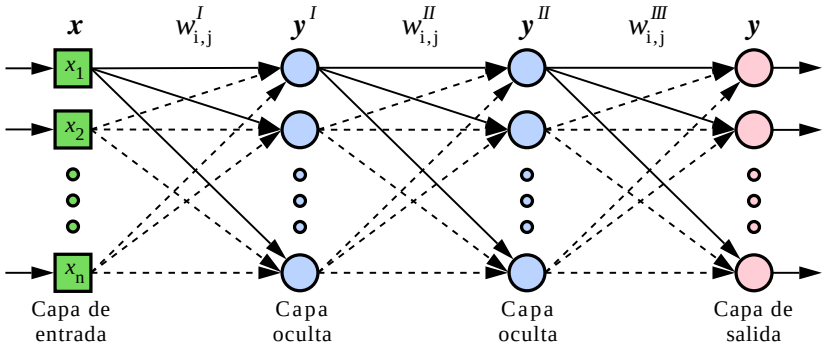

Figura 1.3: Esquema genérico de un perceptron multi-capa.

la primer capa, la salida de esta capa alimenta la segunda capa, y así sucesivamente. Por este motivo los perceptrones multi-capa son denominados redes de alimentación hacia adelante, más conocidas por su nombre en inglés *feed-forward networks* [3]. La ecuación que define las salidas de cada neurona se puede definir como

$$y_j^k = F\left(\sum_{i=0}^{m^{k-1}} w_{ij}^k y_i^{k-1}\right) \tag{1.8}$$

donde y_j^k es la salida de la neurona j de la capa k, m^{k-1} es la cantidad de neuronas de la capa $k-1$ y w_{ij}^k son los pesos asociados a las conexiones entre las salidas i de la capa $k-1$ y la neurona j de la capa k.

En esta arquitectura también su utiliza el paradigma de aprendizaje supervisado en base a la corrección del error entre la salida actual y la salida deseada. Si bien se conoce la salida esperada, para el caso de una red con múltiples capas no se conoce cuáles son las salidas deseadas para los nodos internos. Por este motivo el error debe ser propagado por la red hacia atrás desde la capa de salida y se hace por medio del algoritmo conocido, por su nombre en inglés, como *Back-propagation* [4]. Éste utiliza un método de gradiente descendiente para minimizar el error cuadrático medio de la función de error respecto de todos los pesos de la red. Los pesos de la red se actualizan en la dirección del gradiente negativo del error según

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \nabla E\left(\boldsymbol{w}^{(t)}\right) \tag{1.9}$$

donde $\boldsymbol{w}^{(t)}$ son los pesos en el tiempo t y η es la tasa de aprendizaje.

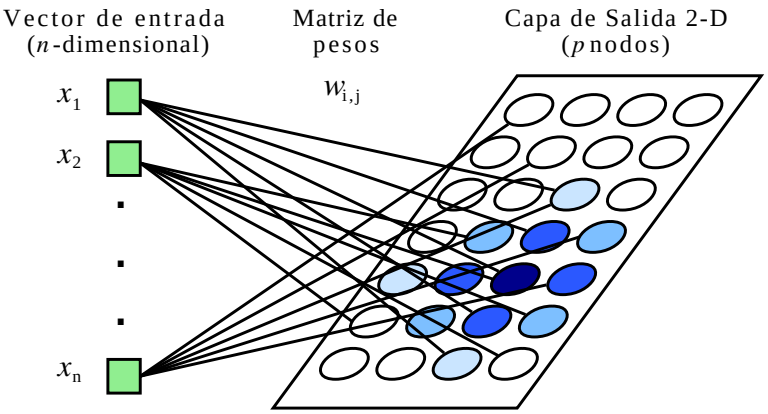

FIGURA 1.4: Esquema de un mapa auto-organizativo bidimensional. Cada patrón mapeado en el plano 2-D activa la neurona ganadora y las de su entorno.

1.1.2. MAPAS AUTO-ORGANIZATIVOS

Un mapa auto-organizativo (SOM, del inglés, *Self-Organizing Map*) es un tipo de red neuronal artificial cuyo entrenamiento se realiza mediante *aprendizaje no supervisado* y con reglas de *aprendizaje competitivo* [3]. La característica principal de un SOM es que puede preservar las propiedades topológicas del espacio de entrada, por lo tanto, los patrones cercanos en el espacio de entrada son mapeados preservando sus relaciones de vecindad en un espacio (de reducidas dimensiones) de salida. Esta red suele utilizarse para descubrir similaridades entre los patrones de entrada a partir de las estructuras subyacentes que no pueden apreciarse en la dimensión de entrada. Los resultados suelen ser muy útiles como punto de partida para la definición de agrupamientos o clases de los vectores de entrada, por ésto suele considerarse un *método de agrupamiento*.

El SOM consiste en una capa de entrada y una capa de salida, con conexiones direccionales de la entrada a la salida y conexiones entre las neuronas de la capa de salida. La red mapea los patrones de entrada ($\boldsymbol{x}$) n-dimensionales a un mapa discreto q-dimensional [5]. En la Figura 1.4 puede verse un esquema de un SOM bidimensional.

Si el mapa tiene p nodos o neuronas, entonces la matriz de pesos w_{ji} es de tamaño $p \times n$. Para cada vector $\boldsymbol{x}$ se mide la *distancia* a cada vector de pesos $\boldsymbol{w}_j$ (asociado a la neurona j). En general se utiliza la distancia Euclidea definida como [4]

$$d(\boldsymbol{x}, \boldsymbol{w}_j) = \| \boldsymbol{x} - \boldsymbol{w}_j \| = \sqrt{\sum_{i=1}^{n} (x_i - w_{ji})^2}. \tag{1.10}$$

Una vez calculadas todas las distancias, se selecciona la neurona ganadora como

aquella que tiene la menor distancia al vector $\boldsymbol{x}$. El paso siguiente es actualizar los pesos en base a la ganadora y esto puede realizarse por medio de dos técnicas: *el ganador toma todo* o *el ganador toma la mayor parte* [4]. En la primera sólo se actualizarán los pesos de la neurona ganadora, mientras que en la segunda se considera la vecindad de la ganadora, es decir que se considera que los nodos que están "muy" cerca interactuan de forma diferente que aquellos que están "más" lejos. Por lo tanto, la definición general para la actualización de los pesos de la neurona $\boldsymbol{w}_j$ (en la iteración $k+1$) es

$$\boldsymbol{w}_j^{(k+1)} = \boldsymbol{w}_j^{(k)} + \eta^{(k)} \Lambda_G^{(k)} \left[\boldsymbol{x}^{(k)} - \boldsymbol{w}_j^{(k)} \right] \tag{1.11}$$

donde $\eta^{(k)}$ es un parámetro de aprendizaje, $\Lambda_G^{(k)}$ es una función de vecindad de la ganadora y $\boldsymbol{x}^{(k)}$ es el patrón de entrada elegido en la iteración k. El parámetro $\eta^{(k)}$ decrece con el tiempo porque en las primeras iteraciones se necesitan correcciones más importantes. La función $\Lambda_G^{(k)}$ se encarga de otorgar mayor importancia a aquellas neuronas que se encuentren más próximas a la ganadora y generalmente es útil que el tamaño de la vecindad considerada sea también decrementado en el tiempo.

1.1.3. Modelos de mezcla de Gaussianas

Una de las distribuciones de probabilidad más conocida y ampliamente utilizada para modelar distribuciones de variables continuas es la denominada *distribución normal* o *distribución Gaussiana*. La morfología de esta densidad es la de una campana, simétrica respecto del valor medio de la distribución. Puede definirse en base a dos parámetros como

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \sigma), \tag{1.12}$$

donde μ es el valor medio y σ es la desviación estándar de los datos.

Aunque las propiedades analíticas de las distribuciones Gaussianas son importantes, éstas encuentran sus limitaciones cuando intentan modelar distribuciones de datos reales [2]. Si se considera un conjunto de datos reales que están concentrados en dos grupos bien separados, se puede notar que una distribución Gaussiana simple no puede capturar su estructura de forma apropiada. Sin embargo, la superposición de dos distribuciones podría ajustarse mucho mejor a la disposición real de los datos. Dichas superposiciones, formadas como una combinación lineal finita de distribuciones simples son usualmente denominadas modelo de mezcla de distribuciones o simplemente modelo de mezcla, y son ampliamente utilizadas en el reconocimiento estadístico de patrones. Si, como se menciona en el ejemplo, estas distribuciones son densidades Gaussianas simples, el modelo es llamado modelo de mezcla de Gaussianas (GMM, del inglés, *Gaussian Mixture Model*) [2] y es definido como

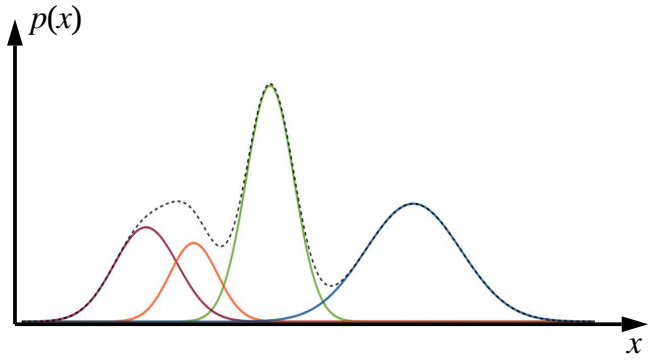

FIGURA 1.5: Gráfica de una mezcla de 4 gaussianas con pesos iguales. Con línea de puntos está graficada la mezcla de las 4 Gaussianas.

$$p(\mathbf{x}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1.13}$$

donde los coeficientes de la mezcla verifican $\sum_k \omega_k = 1$ y $0 \leq \omega_k \leq 1$ para todo k. En la Figura 1.5 se muestra un GMM unidimensional compuesto por 4 Gaussianas simples.

Si se define el modelo de una forma paramétrica como $\lambda = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \omega_k\}$ con $k = 1, 2, \ldots, K$, éste puede ser determinado por los vectores de medias $\boldsymbol{\mu}_k$, las matrices de covarianza $\boldsymbol{\Sigma}_k$ y el vector de coeficientes de la mezcla $\boldsymbol{\omega}$. Los parámetros pueden ser estimados utilizando el método de maximización de la esperanza (EM, del inglés, *Expectation-Maximization*) [2]. Este método comienza con una estimación inicial de los parámetros $\lambda(0)$ a partir de los cuales se estiman los nuevos parámetros del modelo $\lambda(1)$, y ésto se repite iterativamente hasta alcanzar algún criterio de convergencia.

Dada una observación $\mathbf{x}_n$, la ecuación anterior puede expresarse como

$$p(\mathbf{x}_n) = \sum_{k=1}^{K} p(k) p(\mathbf{x}_n|k), \tag{1.14}$$

donde $p(k) = \omega_k$ y $p(\mathbf{x}_n|k)$ es la k-ésima distribución normal. Entonces, utilizando el teorema de Bayes, la probabilidad a posteriori puede escribirse como

$$\gamma_{nk} \equiv p(k|\mathbf{x}_n) = \frac{p(k)p(\mathbf{x}_n|k)}{\sum_\ell p(\ell)p(\mathbf{x}_n|\ell)} = \frac{\omega_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_\ell \omega_\ell \mathcal{N}(\mathbf{x}_n|\mu_\ell, \Sigma_\ell)}. \tag{1.15}$$

Para modelar la distribución de un conjunto de observaciones $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ utilizando un GMM, se debe maximizar la función de costo $-\log p(\mathbf{X}|k)$. Las derivadas

respecto de μ_k, Σ_k y ω_k son igualadas a cero para poder obtener las fórmulas de reestimación

$$\tilde{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n, \tag{1.16}$$

$$\tilde{\boldsymbol{\Sigma}}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T, \tag{1.17}$$

$$\tilde{\omega}_k = \frac{N_k}{N} \tag{1.18}$$

con $N_k = \sum_{n=1}^{N} \gamma_{nk}$.

Si se utiliza un número suficiente de Gaussianas, y se ajustan sus medias, covarianzas y los coeficientes de la combinación lineal, es posible aproximar casi cualquier densidad continua con una precisión arbitraria [2].

1.1.4. Modelos ocultos de Markov

En la sección anterior se presentó un modelo estadístico interesante y valioso en el área de reconocimiento de patrones. Con los GMM es posible modelar las características *estáticas* de los datos, sin embargo, en algunas ocasiones es necesario capturar información acerca de la variabilidad temporal de los datos. Para poder modelar estos cambios es necesario un modelo más complejo que pueda tener en cuenta las características *dinámicas* de los datos.

Los modelos ocultos de Markov (HMM, del inglés, *Hidden Markov Models*) son básicamente modelos estadísticos que describen secuencias de eventos. Cuando son utilizados para clasificar se debe estimar un modelo para cada clase o tipo de señal. Durante la clasificación de una señal se calcula cual es la probabilidad que tiene cada modelo de haber generado ésta. El clasificador informa, como salida, cual es el modelo que obtiene la máxima probabilidad. Los HMM tienen dos elementos básicos: un proceso de Markov y un conjunto de distribuciones de probabilidad de salida [6]. Este modelo es considerado una máquina de estados finita, porque tiene un conjunto de estados que se conectan entre ellos por medio de arcos de transición que tienen probabilidades asociadas. El sistema puede estar, en algún tiempo y con alguna probabilidad, en cualquiera de los estados disponibles y a intervalos regulares de tiempo se produce una transición entre estados. Estas transiciones están condicionadas por las probabilidades de los arcos de transición.

En la Figura 1.6 se puede ver un HMM con una estructura conocida como *de izquierda-derecha*, dado que las transiciones sólo pueden darse en ese sentido. Se puede

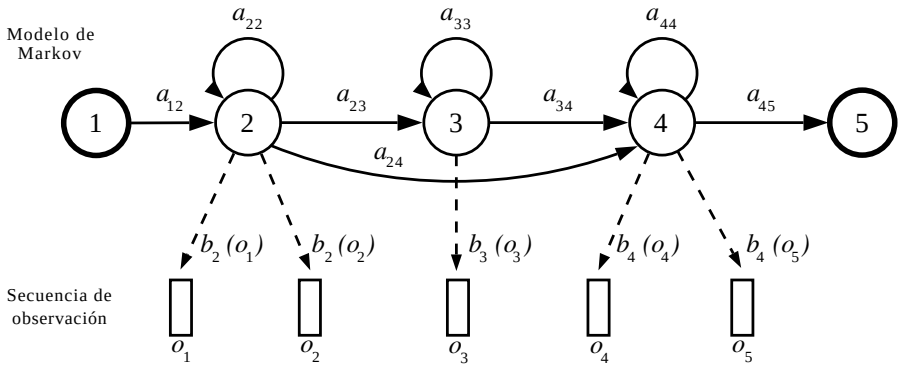

FIGURA 1.6: Esquema de un HMM de 5 estados, donde sólo 3 pueden observar.

ver que los estados pueden emitir algunos símbolos determinados. Los coeficientes a_{ji} son las probabilidades asociadas a las transiciones desde el estado j al estado i, mientras que los parámetros $b_j(o_\ell)$ representan la probabilidad de que el estado j observe el símbolo o_ℓ, del conjunto de símbolos observables. Además, en el modelo de la Figura 1.6 algunos estados no tienen posibilidad de observar algún símbolo (probabilidad cero) y para éstos no se graficaron arcos de probabilidad. Debido a que cada estado puede emitir cualquier símbolo, si se mira solamente la salida del modelo es imposible determinar cuál es el estado actual en el que se encuentra el modelo. Por esta razón, el comportamiento del modelo se encuentra "oculto" y ésto motiva su nombre. En este marco queda claro que no es posible determinar directamente cómo el modelo ha transitado por los diferentes estados para generar una salida particular y con qué probabilidad se genera ésta. Además como todos los estados emiten cualquier símbolo, varias podrían ser las secuencias de estados que la generen. Como puede apreciarse no es un problema trivial, sin embargo, generalmente es suficiente obtener la secuencia que genera la salida con mayor probabilidad asociada y existen muchos algoritmos optimizados para ésto. Entre los más utilizados está el algoritmo de *Viterbi*. Éste recorre todas las transiciones de estados en el tiempo y, para cada estado, sólo almacena la máxima probabilidad acumulada y el estado anterior con el que se logra ésta. El resultado es que muchas secuencias no se computan ya que sólo se continuan los caminos de máxima probabilidad y de esta forma no se pierde generalidad y se realizan menos cálculos [7].

Un HMM está definido por una estructura algebraica $\Theta = \langle \mathcal{Q}, \mathbb{O}, \mathbf{A}, \mathcal{B} \rangle$, donde $\mathcal{Q}$ es el conjunto de estados posibles, $\mathbb{O}$ es el espacio observable, $\mathbf{A}$ es la matriz de probabilidades de transición y $\mathcal{B}$ es el conjunto de distribuciones de probabilidad de observaciones (o emisiones) [6].

En los HMM las distribuciones de probabilidad $b_j(o_\ell)$ para cada símbolo o_ℓ son

discretas, mientras que para los HMM continuos (CHMM) las distribuciones de probabilidad generalmente son expresadas como una mezcla

$$b_j(\mathbf{x}) = \sum_{k=1}^{K} c_{jk} b_{jk}(\mathbf{x}) \tag{1.19}$$

donde K es el número de componentes de la mezcla y b_{jk} es la densidad de probabilidad dada por el componente k de la mezcla (generalmente es una distribución Gaussiana).

Dada una secuencia de evidencias acústicas $\mathbf{X}^T$, el entrenamiento se puede resumir en maximizar la función de costo

$$\mathcal{O}(\Theta, \tilde{\Theta}) \triangleq \frac{1}{p(\mathbf{X}^T|\Theta)} \sum_{\forall \mathbf{q}^T} p\left(\mathbf{X}^T, \mathbf{q}^T|\Theta\right) \log p\left(\mathbf{X}^T, \mathbf{q}^T|\Theta\right) \tag{1.20}$$

El objetivo es determinar los parámetros desconocidos a partir de los datos observables. En un CHMM, los parámetros pueden ser estimados de forma eficiente utilizando el algoritmo de avance-retroceso (del inglés, *forward-backward*) [8]. En los modelos donde todas las distribuciones son Gaussianas, y definiendo

$$\gamma_t(i,j) \triangleq \Pr\left(q_{t-1} = i, q_t = j \left| \mathbf{X}^T, \Theta\right.\right), \tag{1.21}$$

y

$$\psi_t(j,k) \triangleq \Pr\left(q_t = j, k_t = k \left| \mathbf{X}^T, \Theta\right.\right) \tag{1.22}$$

las fórmulas de reestimación de las transiciones de estados $\tilde{a}_{ij}$, los pesos de la distribución $\tilde{c}_{jk}$, los vectores de medias $\tilde{\boldsymbol{\mu}}_{jk}$ y las matrices de covarianza $\tilde{\boldsymbol{\Sigma}}_{jk}$ quedan definidas como [6]

$$\tilde{a}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{1.23}$$

$$\tilde{c}_{jk} = \frac{\sum_{t=1}^{T} \psi_t(j,k)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{1.24}$$

$$\tilde{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T} \psi_t(j,k)\mathbf{x}_t}{\sum_{t=1}^{T} \psi_t(j,k)} \tag{1.25}$$

$$\tilde{\boldsymbol{\Sigma}}_{jk}^{-1} = \frac{\sum_{t=1}^{T} \psi_t(j,k)(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_{jk})^{\mathrm{T}}}{\sum_{t=1}^{T} \psi_t(j,k)} \tag{1.26}$$

1.2. RECONOCIMIENTO AUTOMÁTICO DEL HABLA

En la sección anterior se presentó una introducción al reconocimiento de patrones y a las diversas técnicas que se utilizaron durante el desarrollo de esta Tesis. A continuación se explican los marcos conceptuales de las áreas de desarrollo que se abordaron: el reconocimiento automático del habla y el reconocimiento de emociones en la voz.

Durante su evolución, el ser humano ha desarrollado el proceso de comunicación más avanzado entre los seres vivos: el lenguaje. Ésta una de las características que lo distinguen como el ser más inteligente y el habla es una de las formas de llevar adelante esa comunicación. Las señales de voz son secuencias de sonidos y la forma en que se secuencian éstos está regida por las reglas del idioma. La lingüística es la ciencia que estudia cómo el ser humano utiliza estas reglas, y por otra parte, es la fonética la ciencia que estudia cómo son producidos los distintos sonidos.

El proceso de comunicación básicamente comienza con la idea del mensaje a comunicar que luego, por medio de acciones neurológicas y musculares, es transmitida por medio de ondas sonoras al receptor. Este proceso contempla las estructuras lingüísticas y las reglas gramaticales del idioma a la hora de formar sílabas, palabras, frases, etc. Finalmente, se enfatiza la información importante del mensaje incorporando características prosódicas en las distintas estructuras del habla [8].

El reconocimiento automático del habla (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. Por otra parte, el procesamiento digital de señales ha avanzado en gran medida, gracias a las investigaciones realizadas en el campo del procesamiento de voz [9]. Un sistema de reconocimiento automático del habla puede definirse con un esquema básico de 3 bloques como el que se ve en la Figura 1.7. Una señal de voz es la entrada

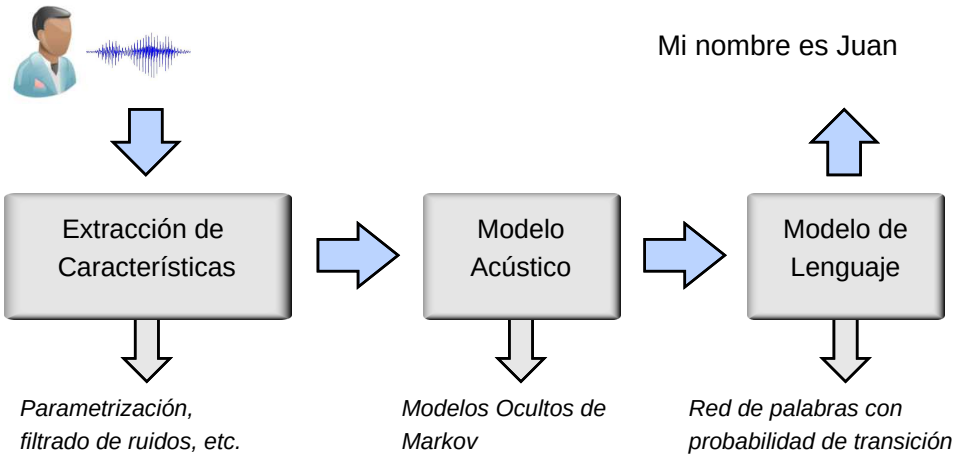

Figura 1.7: Esquema básico de un sistema de reconocimiento automático del habla.

al sistema y la salida es una transcripción de dicha señal. En un primer bloque la señal de voz es adquirida y sometida a diversos preprocesamientos como ser el filtrado de ruidos, la aplicación de técnicas de realce para determinada información de la señal y la extracción de parámetros, entre otras. En un segundo bloque se generan los modelos acústicos de las unidades elementales del habla, utilizando los parámetros ya extraídos. En esta etapa los sistemas estándar modelan fonemas utilizando los modelos ocultos de Markov, pues con éstos se obtienen los mejores resultados en RAH. Su ventaja principal es la capacidad de modelar bien la variabilidad temporal de la señal de voz [10]. Finalmente estos modelos se asocian para formar palabras. En un tercer bloque se incorpora información propia del lenguaje, de cómo y con qué probabilidades se pueden transitar diferentes secuencias de palabras.

1.2.1. Producción de la voz

La señal de voz es producida por el aparato fonador y se transmite mediante ondas de presión propagadas por el aire [8]. Los pulmones impulsan aire hacia afuera a través de la traquea, pasando por la glotis. En el camino está la laringe, que es una de las partes más importantes del tracto vocal, pues allí se encuentran las cuerdas vocales. Cuando las cuerdas vocales están tensas, éstas vibran y el aire es modulado generando pulsos periódicos. Por otra parte, cuando las cuerdas están relajadas el flujo de aire no sufre modificaciones en la glotis. Luego, el flujo de aire continúa su recorrido hacia el exterior pasando por la faringe y la lengua. Finalmente, y dependiendo de la posición del velo del paladar, el aire sale por la boca y/o la nariz y con la modificación recibida en este trayecto se percibe como habla.

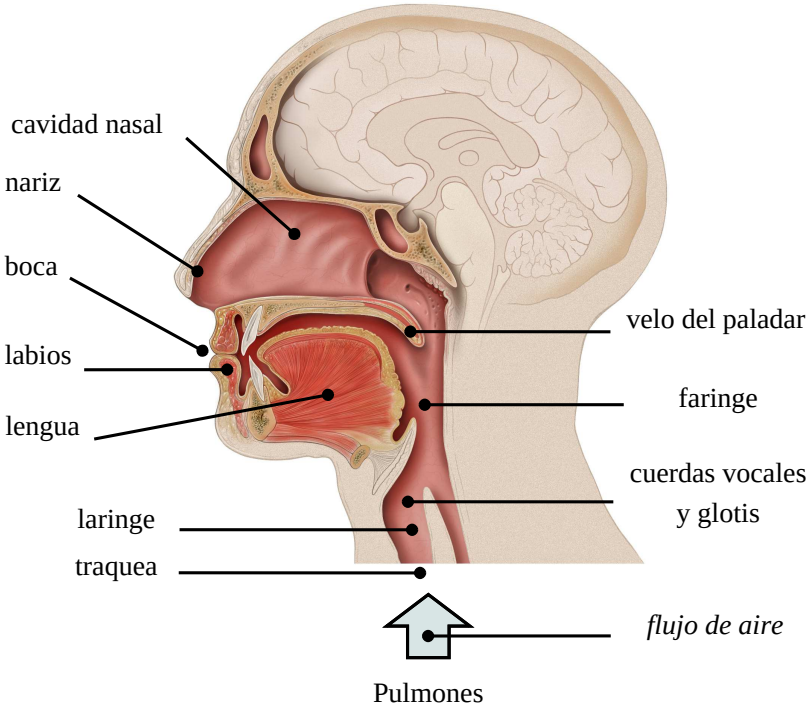

FIGURA 1.8: Sistema de producción de la voz. Imagen de Patrick J. Lynch and C. Carl Jaffe. Adaptada con licencia Creative Commons.

El tracto vocal puede verse como un tubo acústico no-uniforme que va desde la glotis hasta los labios [9]. En la Figura 1.8 pueden verse las partes involucradas en la producción de la voz en un corte lateral de la cabeza y cuello. Como el cambio de la estructura del tracto no es instantáneo, existen pequeños intervalos temporales donde la morfología y por lo tanto las características de la señal emitida se mantienen constantes. En otros periodos se producen ajustes del tracto para generar el siguiente sonido en la elocución (coarticulación). En una señal de voz los sonidos se encuentran en forma secuencial y continua, por lo que sus características intrínsecas son alteradas en sus márgenes y en relación a los sonidos del contexto.

Los sonidos pueden ser categorizados según la exitación presente en su generación como: sonoros o sordos. Como se mencionó previamente, si las cuerdas vocales se encuentran tensionadas se genera un tren de impulsos periódicos y ésto da lugar sonidos sonoros. En el caso de los fonemas sonoros (o tonales), la frecuencia con que vibran las cuerdas vocales se denomina frecuencia fundamental (en inglés: *pitch*, o simplemente F_0) y su sensación auditiva es el tono de la voz o entonación, la que varía con el hablante y el tipo de elocución [11]. Por otro lado, los sonidos sordos son generados mediante

Tabla 1.1: Valores usuales de F_1 y F_2 para vocales en español.

Vocal	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

oclusiones parciales o totales de ciertos sectores del tracto vocal. Sin embargo, existen otros sonidos que son combinación de éstos y algunos que son combinación de éstos y periodos de silencio.

Otra de la formas más generales de caracterizar a los sonidos es la que los divide en vocales y consonantes. En la producción de los primeros el flujo de aire no tiene obstrucciones, mientras que para los otros las características son adquiridas mediante variaciones y obstrucciones en las distintas partes del tracto. Las diferentes configuraciones que puede adoptar el tracto (incluidas las cuerdas vocales) y su propia morfología, dependiente del sexo, la edad, etc. del hablante, permiten que haya regiones donde pueden resonar ciertas frecuencias y otras donde las componentes frecuenciales son atenuadas [8]. En los espectros de la vocales puede notarse claramente la exitación periódica y la estructura armónica presente. Cada vocal tiene un conjunto de frecuencias resonantes que la caracteriza, y son conocidas como *formantes*. En las vocales es apreciable también que su duración es claramente más extensa que la de otros sonidos [12]. Como puede intuirse, y es más evidente en el idioma español, las vocales son fácilmente identificables debido a que son pocas y sus formantes están bien definidas. Como contrapartida, se podría remarcar que no aportan tanta información como la consonantes. Por ejemplo, considerese la siguiente frase:

- _a_ _o__o_a__e_ __i__a_ _a_ i__o__a_io_ _e_ _e__o
- l_s c_ns_n_nt_s br_nd_n m_s _nf_rm_c__n d_l t_xt_

Las 3 primeras formantes (F_1, F_2 y F_3) son las más importantes para la caracterización de los sonidos vocálicos, incluso sólo con las dos primeras se puede lograr una buena clasificación. Las formantes superiores, con frecuencias generalmente mayores a los 3200 Hz, son bastante diferentes para distintos hablantes y caracterizan factores personales. En la Tabla 1.1 se exponen los valores habituales de las dos primeras formantes [12].

Por otra parte, la variedad y las características que identifican a las consonantes son mucho más amplias que en el caso de las vocales. Entre las diversas clasificaciones

que existen, una de las más utilizadas es la que utiliza el *modo de articulación* para distinguirlas y su clasificación general es la siguiente [12]:

- Oclusivas suaves: [b], [d], [g]. Los organos articulatorios se cierran completamente y luego se produce una apertura instantánea. En las mismas condiciones son menos enérgicas que las *Oclusivas fuertes* ya que parte de la energía se usa para hacer vibrar las cuerdas vocales. Ejemplos: *bar, dedo.*
- Oclusivas fuertes: [p], [t], [k]. Los organos articulatorios se cierran completamente y luego se produce una apertura instantánea, son más enérgicas que las anteriores por ser consonantes sordas. Ejemplos: *cata, papá.*
- Nasales: [m], [n], [$\hat{n}$]. Se producen con la cavidad bucal cerrada y el pasaje nasal abierto. Ejemplos: *niño.*
- Líquidas laterales: [l], [λ]. La salida de aire se produce por uno o ambos lados de la cavidad bucal. Ejemplos: *lila.*
- Líquidas vibrantes: [r], [$\bar{r}$]. Existen vibraciones del extremo de la lengua contra el inicio del paladar. Ejemplos: *raro.*
- Fricativas sordas: [f], [s], [∂], [x]. Se produce un estrechamiento en los organos articulatorios y no involucran la vibración de las cuerdas vocales. Ejemplos: *feo, sosa.*
- Fricativas sonoras: [y], [β], [∂], [γ]. Se produce un estrechamiento en los organos articulatorios e involucran la vibración de las cuerdas vocales. Ejemplos: *yuyo.*
- Africadas o semioclusivas [$\hat{y}$], [$\hat{c}$]. Se produce un cierre total de los organos articulatorios seguido de una pequeña apertura de uno de estos donde se produce una fricación. Ejemplo: *chucho.*

1.2.2. 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 una jerarquía en base a la que se pueden desarrollar muchos otros estudios [12]. En este caso el objeto de estudio es principalmente la estructura del mensaje, despojándolo de los mecanismos que lo han generado. Según esta estructura 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. Las manifestaciones de los distintos niveles pueden ser unidades disímiles e independientes que ocurren sin modificar los rasgos característicos

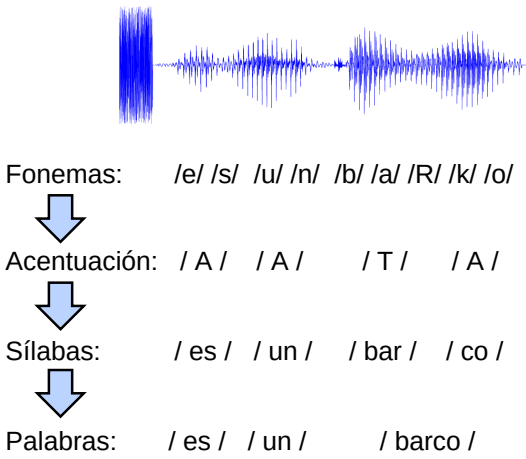

Figura 1.9: Estructuración básica del habla desde una perspectiva lingüística.

a cada nivel. Para este trabajo 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 en la mayoría de los estudios ésta abarque niveles superiores en la estructura del habla.

Fonemas, suprasegmentos y sílabas.

En la Figura 1.9 se pueden ver algunos de los niveles jerárquicos de la organización estructural del habla. Allí puede distinguirse en un primer nivel una señal de voz, silencio y su ruido. Luego, en un segundo nivel se identifica el habla y se determinan las unidades más elementales: los fonemas. Éstos son los modelos definidos para estos sonidos elementales que surgen del análisis del proceso de generación del habla y de la observación de su resultado acústico.

Los suprasegmentos están relacionados con la expresión y representados principalmente por el acento, la cantidad y la entonación [12]. Estas estructuras poseen diversas manifestaciones físicas y sus correspondientes modelos y símbolos lingüísticos asociados. 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 si bien no es un suprasegmento, se le aproxima en su duración. Aquí se define la acentuación como una representación de los suprasegmentos en la que las distintas sílabas, según sean acentuadas o no, se caracterizan como Tónicas (T) y Átonas (A).

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 [13].

PROSODIA

El término *prosodia* refiere a las características físicas que pueden medirse del lenguaje hablado, y entre las más conocidas se encuentran el tono (F_0), la energía y la duración del núcleo vocálico. El rango de variación de la F_0 usualmente está entre 50-250 Hz para los hombres y entre 120-500 Hz para las mujeres [8]. El nivel del tono puede desplazarse, hacia arriba o hacia abajo, según si está relacionado con la acentuación, la entonación o las emociones expresadas [8]. La *entonación* se define asociada al contorno del tono en el tiempo y es un factor importante, aunque no el único, que da forma a estructuras gramaticales como preguntas, afirmaciones, etc. Sin embargo, este término también es utilizado en los diferentes niveles lingüísticos de análisis asociados a la F_0. Entre los diferentes niveles en los que se estudia la entonación están: F_0, tonema, grupo de entonación y curva melódica [14]. La F_0 es calculada sobre los tramos de mínima duración definidos para el análisis de la voz y así constituye el nivel más elemental de estudio. De esta forma se convierte en el punto de partida del análisis en los niveles superiores [9]. Por ejemplo, para el nivel de sílabas o suprasegmentos, pueden definirse *cadencias de entonación*, *anticadencias de entonación* y *mesetas de entonación*. Los métodos utilizados para obtener esta información se detallan en la sección 2.2.1.

1.2.3. ANÁLISIS DE SEÑALES DE VOZ

Entre las características más importantes de las señales de voz están su naturaleza continua y su no-estacionariedad. Por lo tanto, sería incorrecto analizar esta señal en períodos de varios segundos y por otro lado carece de sentido hacer un análisis muestra a muestra. Si analizamos la forma en que se genera una señal de voz, es posible advertir que el tracto vocal tiene una velocidad máxima con la que puede alterar su morfología. Ésto permite plantear la hipótesis, ampliamente aceptada, de que la señal de voz permanece estacionaria durante ciertos intervalos de tiempo en relación a la velocidad de variación de la morfología del tracto. Más pecisamente se considera que la señal de voz es estacionaria en periodos de tiempo que son de aproximadamente 20 ms [8].

Para realizar el análisis de la señal de voz bajo esta hipótesis se recurre al *análisis por tramos*, para el cual es necesario extraer los segmentos de la señal mediante técnicas de ventaneo de la señal [7]. Existen muchos tipos de ventanas que pueden utilizarse en el

análisis por tramos, entre ellas podemos mencionar: la ventana rectangular, la ventana de Hamming y la ventana de Blackman [15]. La elección del tipo de ventana depende directamente de la aplicación y está basada en un compromiso que se asume entre la reducción del fenómeno de Gibbs y la resolución frecuencial requerida. La ventana de Hamming es la más popular en las aplicaciones de voz [8].

Por otra parte, cabe mencionar que sólo una parte de la información presente en las señales de voz es visible en el dominio temporal. Aquí se pueden extraer algunas características de la señal como ser la energía, los segmentos de silencio y los cruces por cero, entre otros. Para extraer características más relevantes para el análisis del habla es necesario cambiar el punto de observación de estas señales, es decir transformarlas. Considerando que el análisis se realiza a las señales ventaneadas, segmento a segmento, se presenta a continuación una pequeña introducción a las transformaciones usadas aquí, en su forma más general y para señales discretas.

Coeficientes cepstrales

El análisis cepstral es un caso especial de los métodos de procesamiento homomórfico [6]. Éste es aplicado en el análisis de señales de voz pues permite obtener de forma separada la información relativa a la señal de excitación y la información del filtro (tracto vocal). Las características obtenidas con esta transformación permiten modelar el tracto vocal que genero la señal. El cepstrum está basado en la Transformada de Fourier (FT, del inglés *Fourier Transform*) [16] y se define como

$$cc(n) = \mathrm{FT}^{-1}\{\log|\mathrm{FT}\,\{x(n)\}\,|\}. \tag{1.27}$$

Coeficientes cepstrales en escala de mel

Un *mel* es una unidad de medida de la frecuencia percibida de un tono. La representación de los Coeficientes cepstrales en escala de mel (MFCC, del inglés *Mel Frequency Cepstral Coefficients*) resulta de un análisis que combina las propiedades del cepstrum y resultados subjetivos de la percepción humana de tonos puros. La escala de mel fue determinada relacionando la escala frecuencial real (en Hz) y la escala de frecuencias percibidas (mel):

$$F_{mel} = 1000 \log_2 \left[1 + \frac{F_{Hz}}{1000}\right] \tag{1.28}$$

Para obtener los coeficientes MFCC se calcula la FT de la señal y luego se aplica al espectro un banco de filtros configurados según la escala de mel [8]. Luego, se toma el logaritmo de la potencia de cada una de las bandas de mel. Finalmente, se realiza la transformación inversa (FT^{-1}). Considerando que el argumento de la FT^{-1} es una

secuencia real y par, puede simplificarse su cómputo utilizando una transformación coseno.

1.2.4. El RAH mediante modelos ocultos de Markov

Para poder diseñar buenos sistemas de RAH es necesario entender como se genera y entiende naturalmente el habla. Por ésto, el RAH es un problema multidisciplinar, 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. 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.

La teoría de fonéticas acústicas postula la existencia de una cantidad finita de unidades fonéticas distintivas del lenguaje hablado. Éstas pueden ser representadas en base a sus propiedades manifestadas en la señal de voz [6]. Aunque estas propiedades son muy variables dependiendo del hablante y de los sonidos del contexto, se considera que hay reglas que rigen esta variabilidad. Utilizando ésto se puede conceptualizar el primer paso del proceso de reconocimiento del habla. En este paso se segmenta la señal en tramos, que mantienen ciertas propiedades acústicas, y se los asocian con partes de las transcripciones fonéticas de las señales. En la siguiente sección se comentá como se utiliza esta información para analizar el habla.

Modelos acústicos

Con los modelos acústicos (MA) se busca definir patrones que permitan analizar el habla. La idea central es suponer que se tiene un modelo asociado a cada emisión oral y que con éste se puede lograr una emisión similar. Siguiendo este razonamiento, se deberían considerar tantos modelos como emisiones distintas se tengan. Luego, al analizar una emisión de voz por segmentos, éstos son comparados con las emisiones de los modelos con el objetivo de determinar cual es el que puede generar esa porción acústica y así presentar el texto asociado a dicho modelo.

Para el caso de una aplicación real se debería contar con una cantidad infinita de modelos, lo cual no es posible y además, puede que estos modelos no sean significativamente distintos entre sí. Por lo tanto, y utilizando la organización estructural del habla

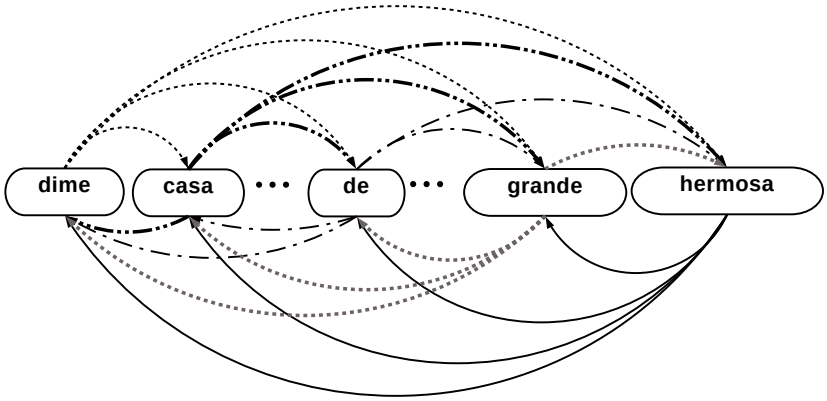

Figura 1.10: Esquema de un modelo de lenguaje. Con cada tipo de línea se representan todas las transiciones posibles para una palabra.

comentada anteriormente, se propone modelar las estructuras simples como fonemas, sílabas o palabras y luego combinarlos para formar los componentes complejos que se requieren [7].

Los modelos ocultos de Markov, descriptos en la sección 1.1.4, son modelos estadísticos que proporcionan descripciones de secuencias de eventos. Pueden entrenarse con muchas pronunciaciones y, al decodificar, el costo computacional depende básicamente del número de modelos y no del número de pronunciaciones con que fueron entrenados [6]. De esta forma se convierten en una alternativa adecuada para modelar los fonemas y actualmente son el estado del arte.

1.2.5. Modelos de lenguaje y modelo compuesto

Para definir el modelo de lenguaje (ML) se dejan de lado las características físicas de las señales, no se consideran los fonemas y se enfoca el estudio en las palabras y en como se combinan éstas para formar frases. El ML representa la gramática del lenguaje y en éste se definen todas las transiciones permitidas entre las palabras existentes en el ámbito de aplicación (almacenadas en el diccionario del reconocedor). En la Figura 1.10 se puede ver un esquema de esta red de gramática que está caracterizada por una lista de nodos (que representan a las palabras) y una lista de arcos que los une entre sí con una cierta probabilidad [17]. Estas probabilidades de transición entre palabras se calculan desde la estructura misma del lenguage, a partir de la definición de una estructura gramática particular. Entonces, una secuencia de n palabras $(p_1, \ldots, p_n)$ tendrá asociada una probabilidad de suceder $P(p_1, \ldots, p_n)$. Usualmente los ML utilizan secuencias de 2 o 3 palabras para calcular esta probabilidad y se denominan bi-gramática o tri-gramática respectivamente. Por ejemplo, para un ML bi-gramático

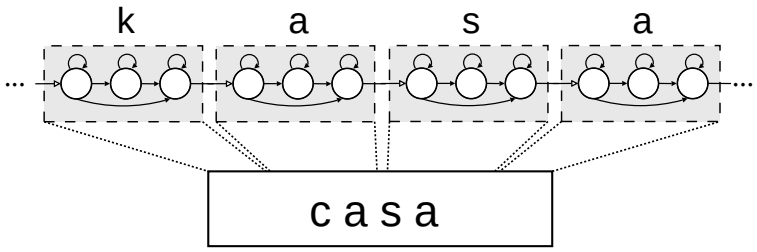

FIGURA 1.11: Diccionario fonético: ejemplo para la plabra *casa*.

la probabilidad de la frase "*Es un barco rojo*" puede aproximarse por

$$P(Es, un, barco, rojo) \approx P(Es| < c >)P(un|Es)P(barco|un)P(rojo|barco) \quad (1.29)$$

donde $< c >$ es simplemente un marcador de comienzo de frase.

Todas las palabras que integran el vocabulario junto a sus descripciones fonéticas conforman el *diccionario fonético*, y con este se pueden formar las palabras a partir de los modelos acústicos de los fonemas (Figura 1.11). Combinando el modelado acústico de fonemas y el modelo de lenguaje se logra un modelo compuesto (MC) capaz de modelar cualquier frase. En la Figura 1.12 se puede ver un MC formado por tres niveles: MA, diccionario fonético y ML.

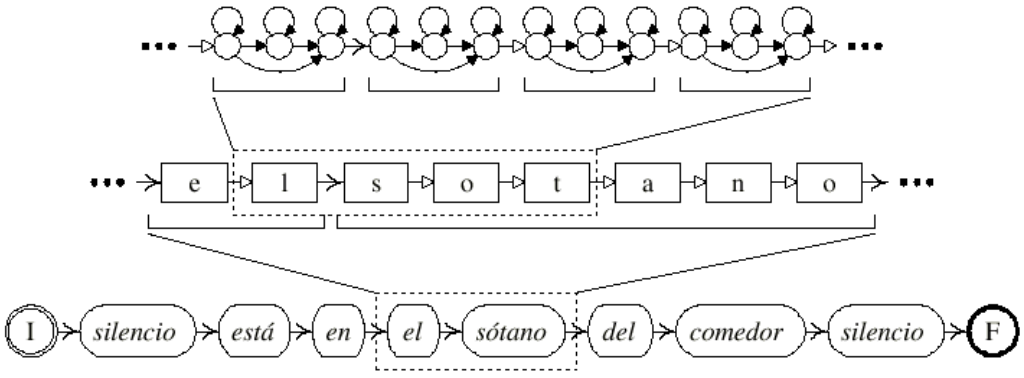

FIGURA 1.12: Modelo compuesto de 3 niveles.

Durante el entrenamiento de los modelos, existen dos conjuntos de parámetros a estimar: las probabilidades de transición y observación de los MA y las probabilidades de transición del ML. Estas estimaciones se pueden realizar en paralelo pues se pueden considerar independientes.

Durante el proceso de reconocimiento se utiliza el MC para eligir el modelo de la frase que posee mayor probabilidad[1], y se obtiene como resultado el texto con que se formó la frase.

1.3. Reconocimiento de emociones

Por medio de un mensaje de voz las personas pueden expresar ideas y comunicarse. Estos mensajes contienen información valiosa, más allá de las palabras, que puede expresarse en forma de prosodia, de estado emocional, etc. Mientras que las personas tienen la capacidad natural de reconocer e interpretar muy fácilmente la información proveniente de los estados emocionales, está vigente el desafio de lograr un modelo general que logre esto en la iteración humano-computadora. Es así que el reconocimiento de emociones ha despertado un gran interés en investigadores provenientes de múltiples disciplinas. Éste cumple un papel importante en el desarrollo del paradigma de interacción hombre-máquina, más aún, es un componente valioso dentro del nuevo paradigma de inteligencia ambiental [18].

1.3.1. Definición de emociones

Existen dos ideas, usualmente propuestas como antagónicas, acerca del origen de las emociones. Una de ellas explica las emociones desde el punto de vista de la psicología evolutiva y la otra las explica como construcciones sociales [19]. La teoría evolutiva fundamenta a favor de que las emociones son adaptaciones naturales y representan una especie de respuesta psicológica típica. En esta línea está ampliamente aceptado el término de emociones "básicas" para definir algunas emociones consideradas universales. El conjunto más popular de emociones básicas es el llamado "grandes seis" e incluye: alegría, enojo, miedo, sorpresa, tristeza y repulsión. Se adiciona a este grupo el estado neutral ya que se considera un estado de referencia que varía dependiendo de una gran variedad de factores. Ekman et al. [20] investigó ésto para argumentar a favor de la universalidad y el carácter innato de las emociones. Ellos encontraron que personas de sociedades diferentes (Nueva Guinea, Borneo, Estados Unidos, Brasil y Japón) reconocian la expresión de las mismas emociones básicas en las mismas fotografías. La idea acerca del origen "evolutivo" de las emociones no fue novedosa entonces, Darwin ya había mostrado avances en está línea que plasmó en su libro *The Expressions of the Emotions in Man and Animals, 1872*. Uno de los hechos que despertó su interés fue observar como sus hijos expresaban sus emociones con gestos faciales, mientras jugaban, de la misma manera que él lo había observado en los monos.

[1] Para este proceso se utiliza una extensión del algoritmo de Viterbi [7].

La segunda teoría, que reivindica la idea de que las emociones son socialmente aprendidas, halla un fundamento diferente en cada cultura. Al expresar una emoción cada individuo tiene en cuenta sus planes, sus memorias, antecedentes, consecuencias de su acción, etc.

Con el pasar de los años los investigadores han encontrado que las descripciones que se presentan en ambas líneas en realidad son muy compatibles [19, 21]. La idea de que las emociones son producto de la influencia natural y social ha llevado a conciliar las definiciones de las emociones. Una idea común es que las manifestaciones de las emociones están diseñadas para comunicarse con otras personas o animales, o bien representan reacciones a determinados eventos [21]. Un descriptor importante de las emociones es la historia pasada, y la definición unificada considera que refiere a que ha sido adapatada durante la historia de la especie y que es adaptada en la vida del individuo en una sociedad [21].

Por otra parte, el modelo que define a las emociones de forma discreta ha perdido auge, entre otras cosas porque sostiene que estás emociones básicas no tienen influencia entre ellas y no pueden ocurrir simultáneamente. Entonces ganan relevancia las conceptualizaciones psicológicas de las emociones, que por medio de modelos bi-dimensionales y tri-dimensionales, caracterizan a las emociones [22, 23, 24] y serán comentados a continuación.

1.3.2. Modelos de emociones

Como se mencionó en la sección anterior, existen emociones que presentan características particulares y para ellas existe el concepto de emociones básicas (primitivas, primarias o fundamentales) que las agrupa. A pesar de que muchos coinciden en el concepto, existen diferencias a la hora de determinar cuantas, cuales y porque son básicas [25]. Desde una perspectiva biológica las emociones básicas son que pueden hallarse en culturas diferentes e incluso entre diferentes especies. Desde la psicología se sostiene la existencia de unas pocas primitivas y que a partir de éstas se pueden formar las demás, se podría decir que es análoga a la clasificación de los colores en primarios y secundarios. Más allá de los distintos fundamentos, desde un punto de vista práctico, es útil definir un pequeño conjunto de emociones para lograr que el problema sea manejable.

Como se ve, es posible que cada investigador considere como central determinados factores que motiven la clasificación y por otra parte hasta el mismo lenguaje condiciona la existencia de los pares *emoción-etiqueta*. Para uniformizar ésto se han definido teorías de "emociones dimensionales" basadas en pocos conceptos bien definidos. En estos modelos se define un espacio bi-dimensional donde los ejes representan la *valencia* y la *excitación*. En la Fig. 1.13 puede observarse la distribución de las "grandes seis" en ese espacio bi-dimensional. Según este modelo, todas las emociones pueden ser

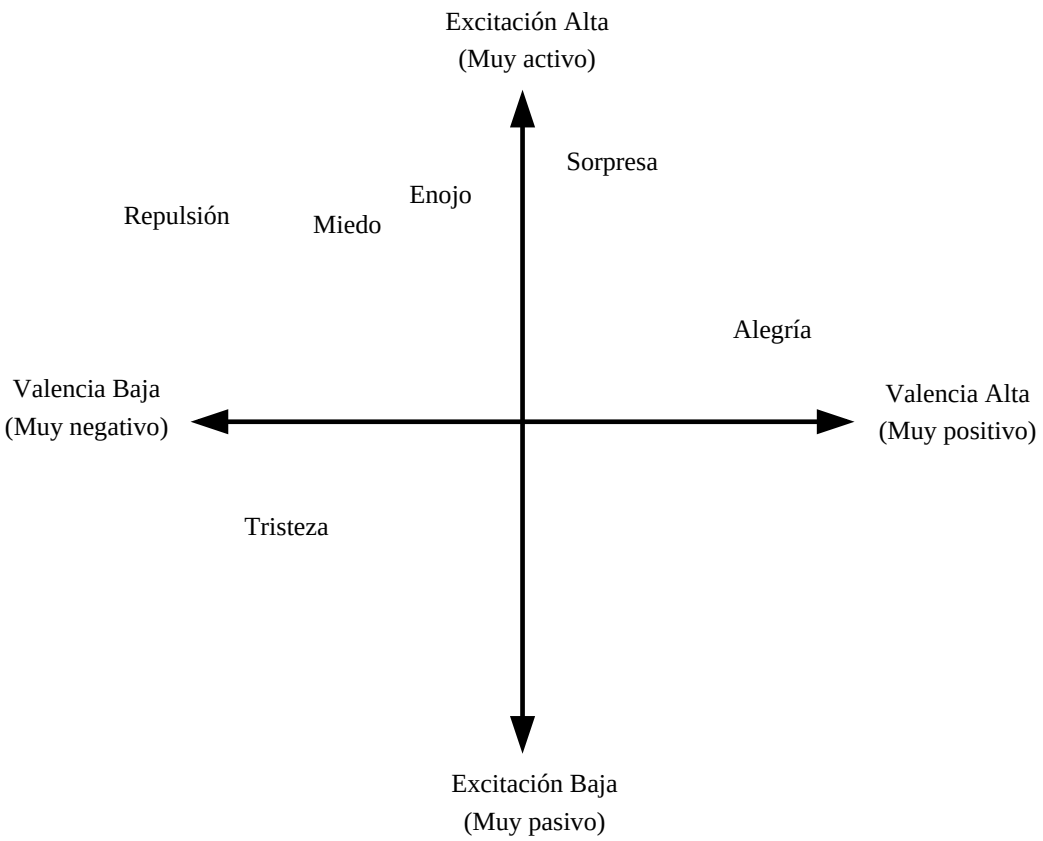

Figura 1.13: Diagrama del modelo bidimensional de emociones.

caracterizadas en función de estas dos variables. La *valencia* se mide respecto de la evaluación positiva o negativa de personas, cosas o eventos. Mientras que la *excitación* mide el temperamento o la fuerza con la que se involucra la persona para tomar alguna acción [26].

En lo que respecta a la tarea de reconocer emociones, se pueden mencionar muchos trabajos que basan sus análisis en las características prosódicas de la voz y en la información espectral que presenta ésta [27, 28, 29, 30, 31, 32]. Para la etapa de clasificación se han empleado muchas técnicas estándar entre las que se encuentran los clasificadores basados en GMM, HMM y MLP [33, 34, 35, 36, 37]. Si bien se han obtenido buenos resultados con este tipo de clasificadores estándar, las mejoras en el rendimiento de estas ténicas podrían haber encontrado un límite. La fusión, la combinación y el ensamble de clasificadores podría representar un nuevo paso en la búsqueda de mejores sistemas de reconocimiento de emociones.

En este capítulo se ha presentado el reconocimiento de patrones y algunas de las

técnicas que le son propias, y que han sido utilizadas en el desarrollo de la presente Tesis. Así también, se han introducido las aplicaciones prácticas sobre las que se han realizado las investigaciones. En el siguiente capítulo se expondrán los trabajos realizados en el reconocimiento automático del habla utilizando la información prosódica y los resultados obtenidos. Por otra parte, en el Capítulo 3 son presentados los avances realizados en el campo del reconocimiento de emociones utilizando señales de voz emocional.

Capítulo 2

Reconocimiento automático del habla con prosodia

Esta investigación tiene como objetivo general encontrar, dentro de las manifestaciones físicas de la prosodia, la información necesaria para mejorar el rendimiento de los sistemas de reconocimiento automático del habla. Se propone un método para caracterizar a las palabras según sus estructuras prosódicas y, aplicando modelos de lenguaje variables en el tiempo, se utiliza esta información para desambiguar hipótesis en el proceso de reconocimiento mediante modelos ocultos de Markov.

Para esta Tesis se parte de los fundamentos básicos de [38], donde se utiliza la información acentual del idioma en combinación con los rasgos prosódicos, pero se opta por la idea de que en realidad los rasgos prosódicos que presentan las señales de voz no están tan íntimamente relacionados con la acentuación definida en las reglas ortográficas. A partir de ésto se espera lograr una nueva forma de clasificar las prominencias acentuales del idioma. En las Figuras 2.1 y 2.2 se observa la información espectral y la información prosódica de dos palabras distintas (*papá* y *capa*). Si bien estas palabras sólo se diferencian en una consonante, la información prosódica es muy diferente. Esta información es muy importante para poder distinguir a las palabras entre sí y podría ser la clave para mejorar el rendimiento de un reconocedor automático del habla. Por otra parte, se puede ver que esta información está estrechamente relacionada con la acentuación ortográfica de las palabras. Sin embargo, ésto sólo es visible si consideramos a las palabras de forma aislada, mientras que en el discurso continuo y aún más para diferentes acentos del idioma este vínculo tiende a perderse [38] y es allí donde gana relevancia un análisis más detallado de la prosodia.

Esta información prosódica será integrada al sistema de RAH a través de los modelos de lenguaje [40], desambiguando las hipótesis del reconocedor para reducir así el error de reconocimiento.

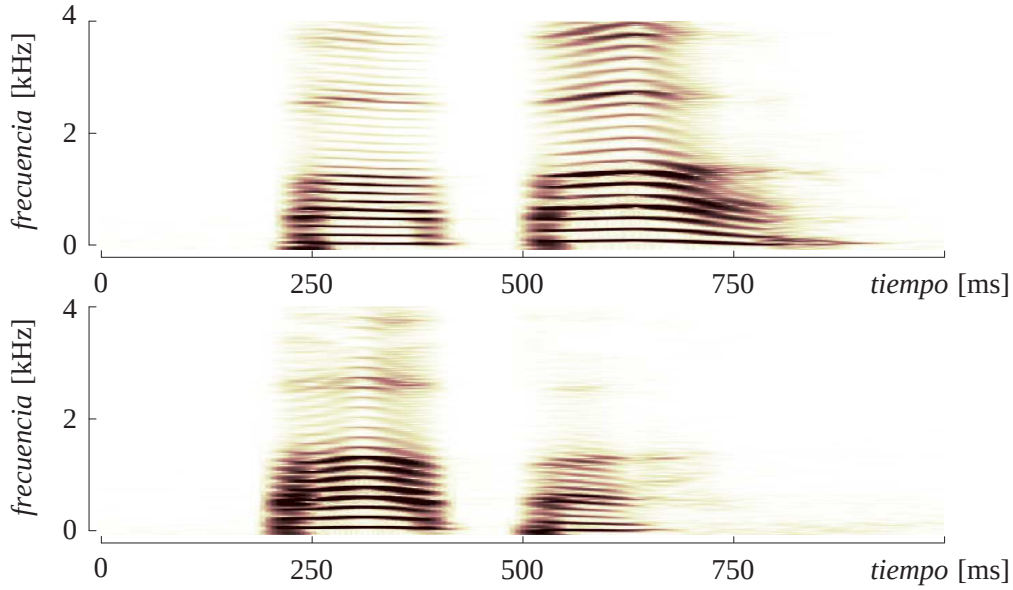

FIGURA 2.1: Información espectral de la palabras *papá* y *capa*. Adaptada con permiso de [39].

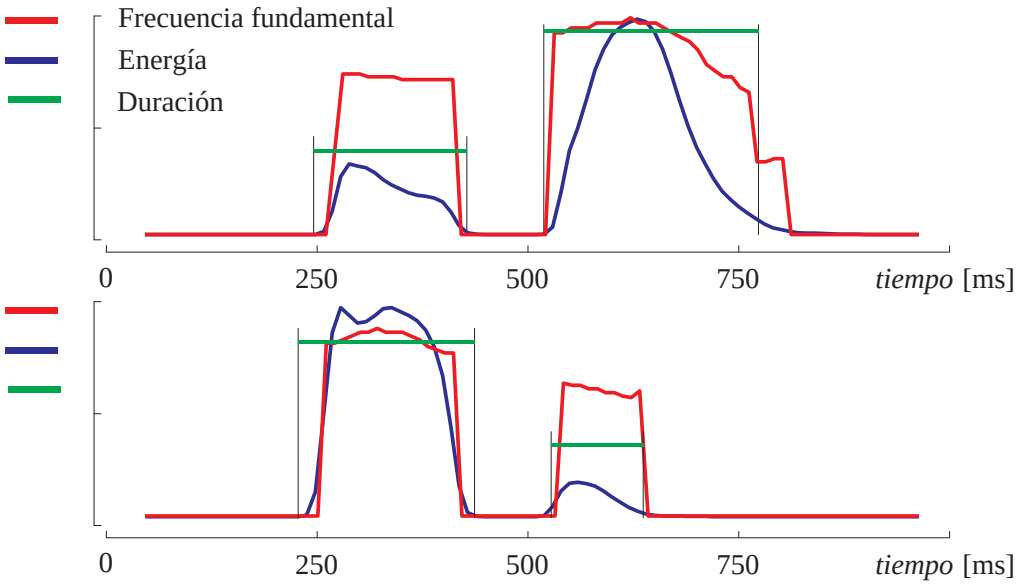

FIGURA 2.2: Información prosódica de la palabras *papá* y *capa*. Adaptada con permiso de [39].

2.1. Antecedentes

Se ha avanzado mucho en el reconocimiento automático del habla (RAH) y se ha incorporado información importante a los distintos niveles de análisis de los sistemas de RAH, desde el fonético hasta el gramatical. Por otra parte, los rasgos prosódicos se encuentran en uno de los niveles de análisis que aún no están completamente integrados al RAH.

Existen varias investigaciones donde se usa de la prosodia para mejorar el rendimiento del RAH. En [41] se define un clasificador binario de acentuación silábico en base a la prosodia, para luego comparar las secuencias acentuales de una palabra con las propuestas según la acentuación ToBI. Allí no se modifica el RAH original y se utiliza la información prosódica para refinar las N hipótesis más pobables (N-best). Chen et al. [42] proponen una extensión de los HMM para modelar explícitamente la duración. El método consiste en reetiquetar a las palabras que estén al inicio y al final de frase así como las que son una frase en sí mismas, lo que genera un nuevo modelo de lenguaje. Luego, se extiende la dependencia prosódica a los modelos de fonemas obteniendo así un modelo acústico diferente según sean vocales o consonantes, sean del inicio o el final de palabra y para la combinación de éstas. En [43] se utiliza un modelo de lenguaje enriquecido con prosodia (modelo de lenguaje prosódico más modelo acústico-prosódico) para reconocer sílabas. Ellos definen el modelo acústico-prosódico como un clasificador binario cuyas entradas son los rasgos prosódicos de la sílaba y que determina si la sílaba está acentuada. Paralelamente, los n-gramas tienen en cuenta también la historia de las etiquetas prosódicas. En [44], un pequeño conjunto de HMM es entrenado con rasgos prosódicos y es utilizado para segmentar unidades prosódicas definidas en idioma húngaro. El sistema de RAH utiliza una red de palabras ajustada en base a la salida del segmentador prosódico y a la red de palabras obtenida del sistema de RAH en una primera etapa. En el trabajo de Huang y Renals [45], se categorizan los rasgos prosódicos a nivel de sílabas en 16 clases con cuantización vectorial. Luego, las características prosódicas de las palabras se representan como una concatenación de estas clases. Estas definiciones se utilizan en un modelo de lenguaje factorizado compuesto de una interpolación entre un n-grama estándar y otro que incluye la prosodia; y en un modelo bayesiano jerárquico donde se asume un modelo generativo de palabras en base a la representación prosódica a través de una variable latente (entre palabra y prosodia). En [46] se propuso un método para incorporar información adicional a un sistema de RAH mediante la penalización adaptativa del modelo de lenguaje. En un trabajo posterior [38] se propuso una nueva definición de la acentuación prosódica, utilizando los rasgos prosódicos (F_0, energía, duración temporal del núcleo vocálico, etc.) combinados con la acentuación establecida en las reglas ortográficas del español. Allí se utilizó con éxito la misma técnica 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.

También se podrían mencionar algunas investigaciones donde han utilizado rasgos prosódicos en sistemas de traducción automática [47, 48], o para detectar eventos espurios y fines de frases o palabras [49, 50, 51, 52].

2.2. Modelado de estructuras prosódicas

En esta sección se presenta el método denominado *histogramas prosódicos* con el que se obtienen modelos prosódicos que caracterizan a las palabras. Se explica cómo son obtenidas las características prosódicas de las señales y cómo son creados los modelos de palabras según la separación silábica definida en las reglas ortográficas españolas. Finalmente, se realiza una valoración de los modelos obtenidos.

2.2.1. Cálculo de rasgos prosódicos

Los rasgos prosódicos utilizados aquí son F_0 y energía. Para el cálculo de la energía se realiza el producto interno de cada tramo de señal consigo mismo, lo que es igual al cuadrado de su norma-2 [16]:

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

Para la extracción de la F_0 de la señal se utiliza un algoritmo basado en CC similar al propuesto por Noll [53]. En éste se plantea como base el procesamiento homomórfico. Se puede considerar que 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 $h(t)$ [53]

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

Esto en el dominio frecuencial puede expresarse como

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

siendo $F(\omega) = \mathrm{FT}[f(t)]$, $H(\omega) = \mathrm{FT}[h(t)]$ y $S(\omega) = \mathrm{FT}[s(t)]$. Luego, aplicando de forma sucesiva las operaciones valor absoluto y logaritmo se tiene que

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

donde $C_y(\omega) = \log(|Y(\omega)|)$, $C_s(\omega) = \log(|S(\omega)|)$ y $C_h(\omega) = \log(|H(\omega)|)$. Finalmente, aplicando la FT inversa se obtiene

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

donde el cepstrum de la secuencia de salida $c_y(n)$ (señal de voz) 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 [54]. La notable diferencia en la localización de las secuencias cepstrales ($c_s(n)$ y $c_h(n)$) permite que se puedan separar las dos componentes. Esto se debe a que 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. A partir de estas consideraciones se puede ver que toda la información sobre el tracto vocal esté acumulada en los primeros coeficientes del cepstrum. Generalmente, mediante un proceso de *liftrado*[1] o simplemente tomando los primeros 30 coeficientes se puede obtener la información relativa al tracto vocal, debido a que sus componentes principales se encuentran en torno a valores pequeños de n. Para los sistemas de RAH se suele descartar la información relativa a los pulsos. En el cepstrum se pueden observar picos localizados en el periodo fundamental de la señal (y en múltiplos de éste) que van decayendo en amplitud con n. Entonces, obteniendo la posición del pico máximo dentro de los primeros 15 ms se puede identificar el período fundamental y así determinar la F_0 (Figura 2.3).

Finalmente, las características prosódicas son calculadas a lo largo de todas las frases y se obtiene la información que se presenta a modo de ejemplo en la Figura 2.4.

2.2.2. Método de modelado prosódico de palabras

En el español, la acentuación definida por las reglas ortográficas guarda una débil relación con las manifestaciones prosódicas del habla [38]. La idea principal en esta investigación es pasar la información de la acentuación definida según las reglas ortográficas a un segundo plano 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.

[1] denominación del filtrado en el dominio del cepstrum.

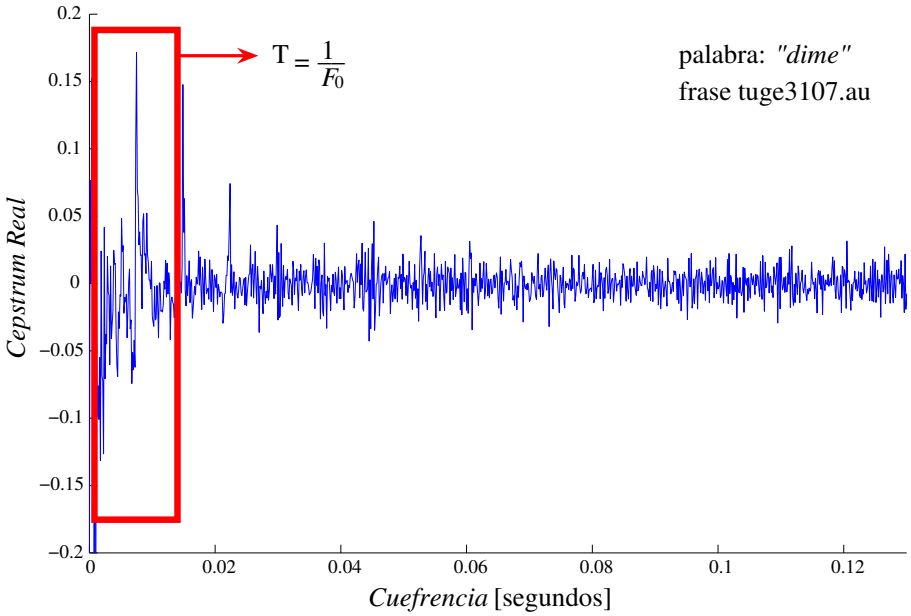

FIGURA 2.3: Cálculo de la F_0 a partir de los coeficientes cepstrales.

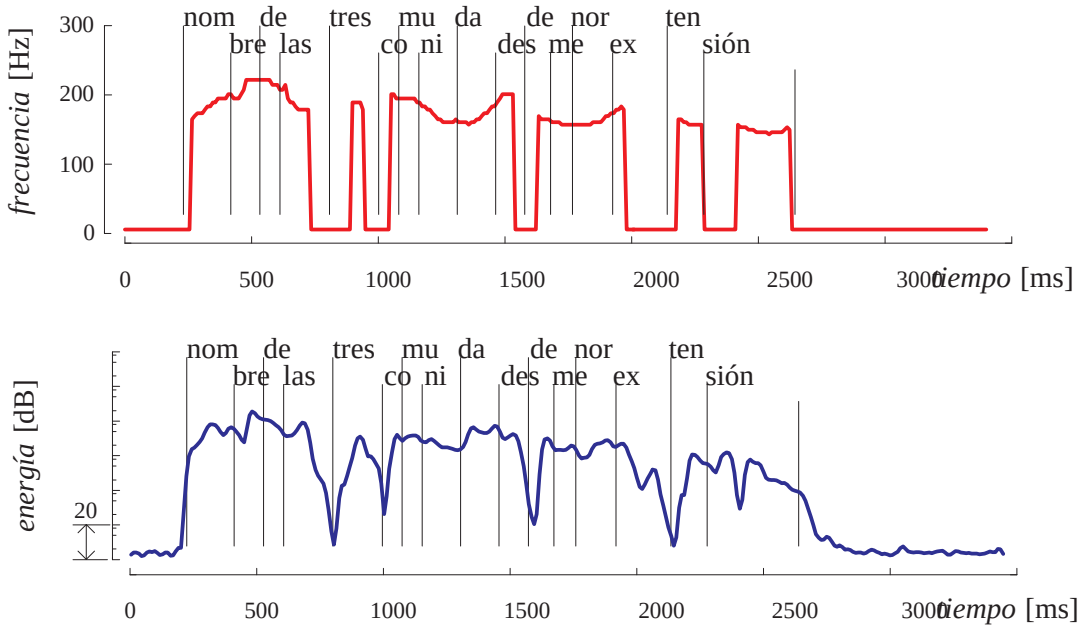

FIGURA 2.4: Cálculo de rasgos prosódicos de una frase. Adaptada con permiso de [39].

Tabla 2.1: Características de la miniGeo 2.

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

Corpus de habla

Las frases utilizadas fueron extraidas de la base de datos Albayzin [55], 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. El corpus se completo tiene 15600 elocuciones pronunciadas por 152 hombres y 152 mujeres. Las frases están distribuidas en un Corpus fonético, un Corpus geográfico y un corpus *Lombard*.

En esta Tesis se utilizó un subconjunto denominado MiniGeo 2, para los primeros avances sobre histogramas prosódicos [56]. En la Tabla 2.1 se presentan sus características principales.

Para comprobar los resultados y para los trabajos posteriores se utilizó el corpus geográfico. Este corpus se compone de 6800 elocuciones de frases que se extrajeron de la utilización de una aplicación de consulta a una base de datos de geografía española. Con restricciones semánticas y sintácticas se buscó reflejar la forma natural del habla en lengua castellana. Todas las frases se clasificaron según criterios lingüísticos, semánticos y de complejidad estructural. Se divide en un conjunto de entrenamiento 4400 frases de 88 locutores y un conjunto de prueba de 2400 frases de 48 locutores.

Método de histogramas

Para modelar las estructuras prosódicas de una palabra se ha propuesto una técnica de clasificación basada en histogramas que consiste en una serie de pasos, detallados en el Algoritmo 1. El método permite una caracterización de las palabras, asociadas por su cantidad de sílabas, en base a histogramas prosódicos [57, 58]. Los análisis se realizaron utilizando la base de datos de habla Albayzin [55] en idioma español, comentada en la sección previa.

1. Identificar exactamente las posiciones de los fonemas correspondientes a la palabra dentro de las frases.
2. Extraer rasgos prosódicos (F_0, energía, etc.) para el segmento (aplicando los métodos clásicos de análisis por tramos de señales [8]).
3. Identificar las sílabas en el segmento y asociarle sus valores prosódicos.
4. Calcular los mínimos, medias y máximos prosódicos para cada sílaba.
5. Definir los modelos prosódicos para la palabra:
   a. Comparar el valor de cada rasgo entre sílabas.
   b. Asignar una codificación de clase.
   c. Contabilizar los sucesos de cada clase.
6. Generar los histogramas y clasificar las palabras según sus clases prosódicas más características.

ALGORITMO 1: Generación de histogramas prosódicos.

SISTEMA PARA LA SEGMENTACIÓN DE PALABRAS

Una de las etapas del algoritmo comprende la segmentación de las palabras en sílabas. Para llevar adelante esta tarea, se ha utilizado un sistema de RAH estándar desarrollado utilizando un conjunto de herramientas denominado *Hidden Markov Model Toolkit* (HTK)[2] [17].

Dentro de los muchos pasos que son más bien comunes a todos los desarrollos de sistemas RAH, a continuación se comentan los más destacados en relación a esta Tesis. Las señales se analizaron con ventanas de 25 ms de ancho y con un paso de 10 ms. Los parámetros que se extrajeron fueron 12 coeficientes MFCC y la energía. También se utilizaron coeficientes delta y aceleración [17]. Para la creación y diseño del prototipo HMM se seleccionaron, como es habitual para modelar fonemas, modelos de 5 estados de los cuales el primero y el último no pueden observar. Los estados fueron modelados con mezcla de gaussianas esféricas en $\mathbb{R}^{39}$. Mientras que para el diseño del modelo de lenguaje se consideró una estructura de bigramas, como es habitual.

Finalmente, con el sistema entrenado y utilizando las transcripciones de las frases se puede obtener la segmentación (en fonemas) de los segmentos de audio. Ésto se logra utilizando el método de alineación forzada del algortimo de Viterbi [59]. Luego, sólo es necesario utilizar la información ortográfica de las palabras para obtener su separación en sílabas segmentada en el tiempo. Se decidió utilizar este sistema para segmentar las frases porque es el mismo que se utiliza para reconocer las frases. De esta manera se puede decir que se utiliza la misma información de alineación del sistema tanto para

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

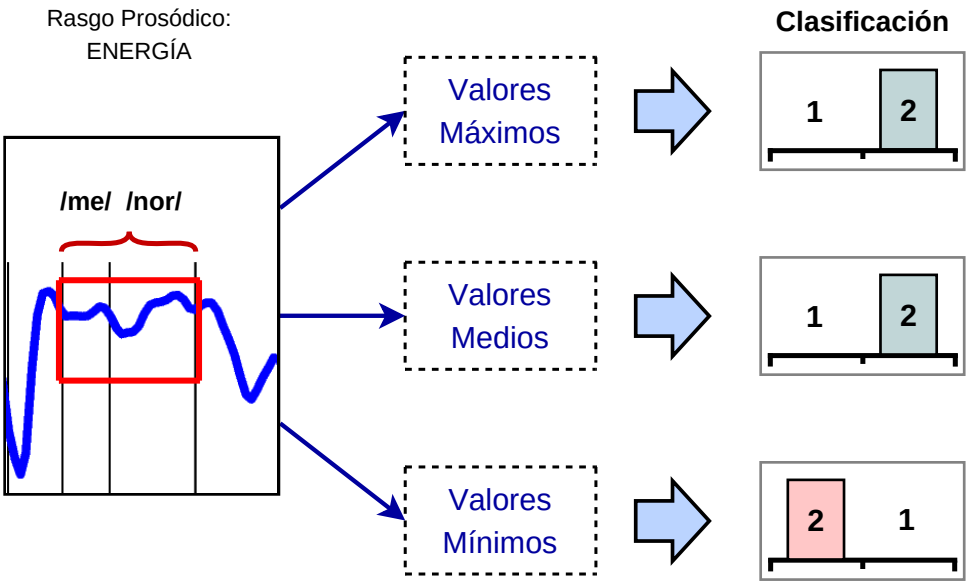

Figura 2.5: Ejemplo de obtención de los histogramas para la energía (palabra *dime*).

modelar acústicamente los fonemas como para generar los modelos prosódicos. Es decir, si hay algún sesgo en la segmentación, éste será el mismo en todos los casos.

Modelado de palabras basado en histogramas

El primer paso es obtener las sílabas de cada palabra bien delimitadas, utilizando el diccionario fonético y la segmentación comentada anteriormente. Luego, se asocian estas sílabas a la información prosódica calculada para dichos segmentos. Finalmente, con la palabra asociada a su información prosódica, se calculan para éstas los valores máximos, mínimos y medios en cada sílaba (puede verse un ejemplo en la Figura 2.5). Ésto se hace con todos los sucesos de todas las palabras en la base de datos[3]. Para cada suceso de una palabra, se considera un rasgo particular (por ejemplo: mínimo(F_0)) y se comparan los valores obtenidos en cada sílaba (por ejemplo para bisílabos serían 2 valores). Luego de la comparación se asigna a la sílaba con menor valor asociado un *1*, a la sílaba que tiene un valor inmediatamente mayor un *2* y así sucesivamente. Finalmente, se obtienen clases (asociadas a determinados rasgos prosódicos) codificadas en n dígitos, siendo n el número de sílabas de la palabra (Figura 2.5). El código indica en forma relativa la magnitud medida (como ser el máximo de F_0) para cada sílaba. A modo de ejemplo, la clase *321* (para palabras de tres sílabas) indica que la primer

[3]Las palabras fueron preseleccionadas eligiendo aquellas que tenían al menos 28 ocurrencias en la base de datos considerada.

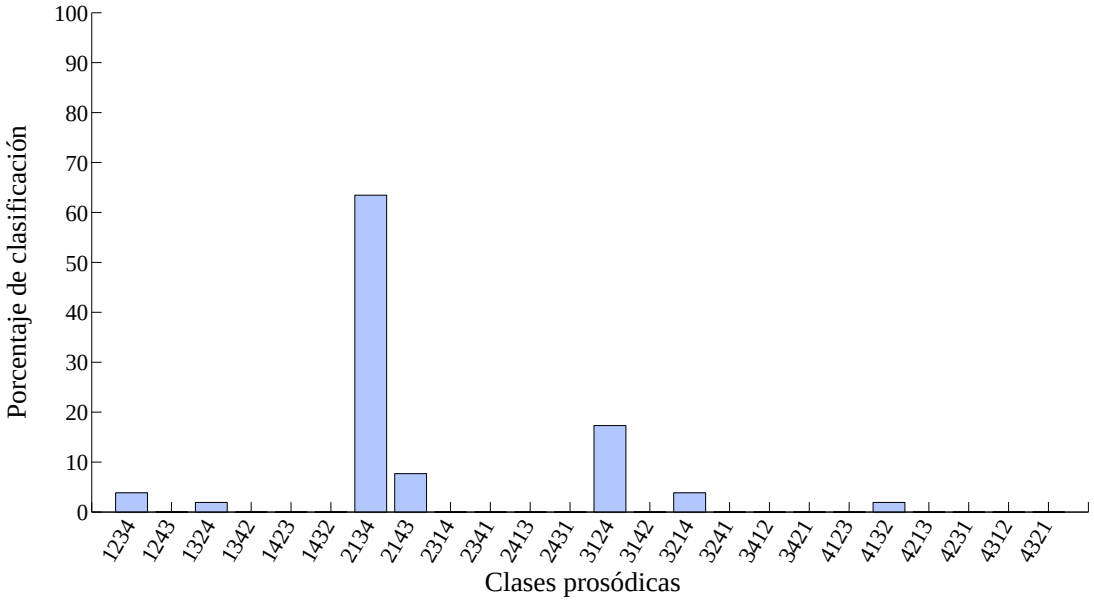

FIGURA 2.6: Histograma prosódico completo para *desemboca*. Rasgo: máximo de energía.

sílaba tiene valor máximo y la última valor mínimo; así la clase *213* indica que el valor máximo está en la última sílaba y mínimo en la segunda. En resumen, el método de modelización con histogramas prosódicos brinda la siguiente información:

- Clases codificadas en **n** dígitos.
- El código indica en forma relativa la magnitud medida para cada sílaba.
- Cada palabra, para sus distintos sucesos, puede pertenecer a alguna de las **n! clases**.

El objetivo final es obtener que "clases prosódicas" caracterizan a las palabras y ésto se obtiene contabilizando cuantos sucesos de cada clase poseen. En primera instancia se puede ver que cada palabra, para sus distintos sucesos, puede pertenecer a alguna de las $n!$ clases que se forman al intercambiar las distribuciones de las cantidades dentro de la palabra. Afortunadamente para este método, casi todas las palabras pertenecen a unas pocas clases y están caracterizadas en su mayoría por una única clase. Para ejemplificar ésto, se presentan algunas gráficas con valores relativos (normalizados) que son el resultado de la caracterización para distintos rasgos prosódicos. En la Figura 2.6 se puede ver la cantidad de sucesos que presenta la palabra *desemboca* para las distintas clases posibles (considerando el rasgo máximo de energía). Como se puede

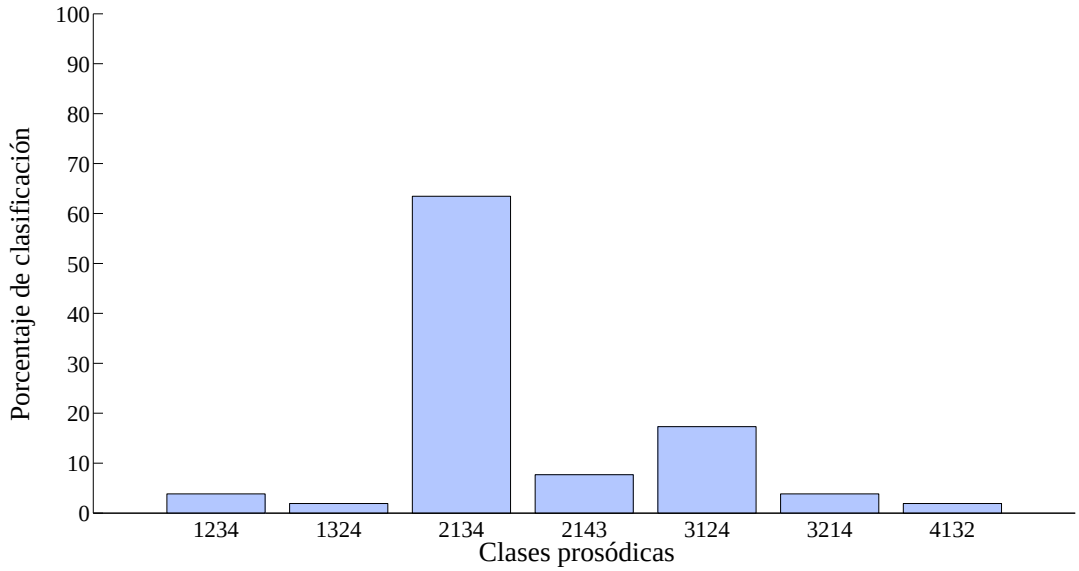

Figura 2.7: Histograma prosódico de la palabra *desemboca*. Rasgo: máximo de energía.

apreciar en la figura existen muchas clases que no presentan sucesos por lo que, para simplificar las gráficas posteriores, se han eliminado las clases para las que una palabra tiene cero sucesos. Así, se puede apreciar mejor en la Figura 2.7 como la clase *2134* caracteriza a *desemboca* para este rasgo prosódico. En la Figura 2.8 se observan las clases que caracterizan la palabra *dime* para la media de energía. Aquí se caracteriza completamente a la palabra con la clase *12* (para 256 palabras computadas) y ésto se interpreta como: *la palabra dime se caracteriza por tener un valor mayor de energía media en la segunda sílaba*. En la Figura 2.9 se ven las clases que definen la palabra *longitud* para el rasgo prosódico mínimo de energía. Con 134 palabras computadas, se ve claramente que la clase *321* define completamente esta palabra. En la Figura 2.10 se observan las clases para la media de F_0 de la palabra *valenciana*, que queda caracterizada por la clase *4321* (para 34 palabras computadas).

Por otra parte, es posible identificar palabras que no pueden ser clasificadas con este método, para ninguno de los rasgos prosódicos propuestos. Por ejemplo, para la palabra *cúbicos* (con 62 sucesos computados) no se encontró una clase, de al menos un rasgo prosódico, que la caracterice. En la Fig. 2.11 se puede observar una distribución que consideramos insuficiente para caracterizar a esta palabra (rasgo mínimo de energía). Un caso similar se presenta con la palabra *comunidad*. En la Fig. 2.12 se ven las clases para el rasgo mínimo de energía, con 256 sucesos computados de la palabra. En la Fig. 2.13 se observa que para la palabra *valencia* (con 29 sucesos computados) no se

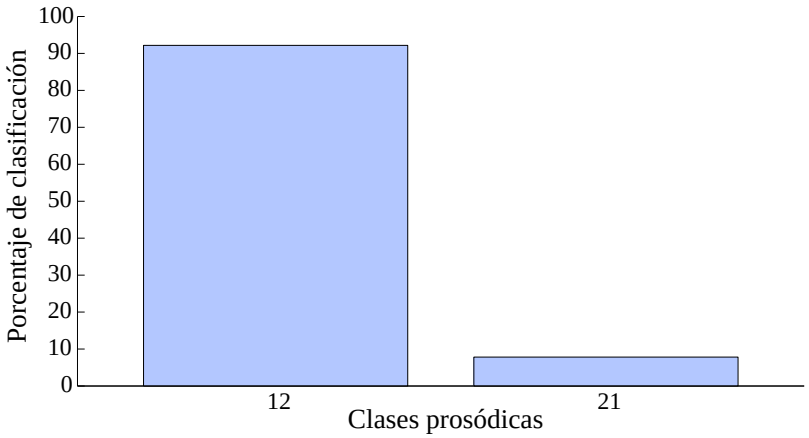

FIGURA 2.8: Histograma prosódico de la palabra *dime*. Rasgo: media de energía.

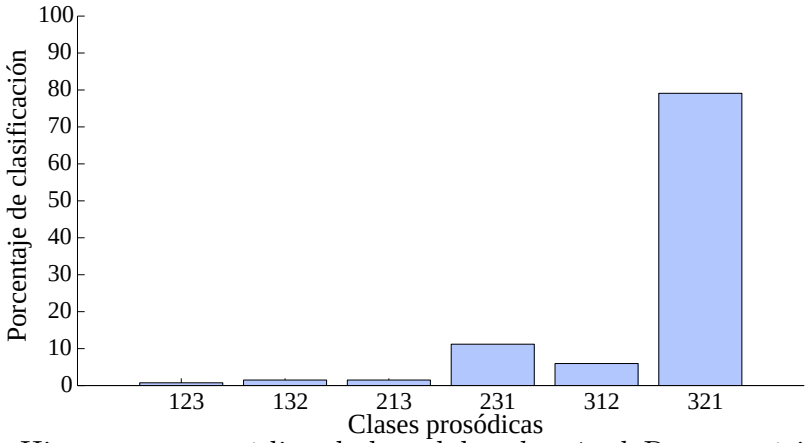

FIGURA 2.9: Histograma prosódico de la palabra *longitud*. Rasgo: mínimo de energía.

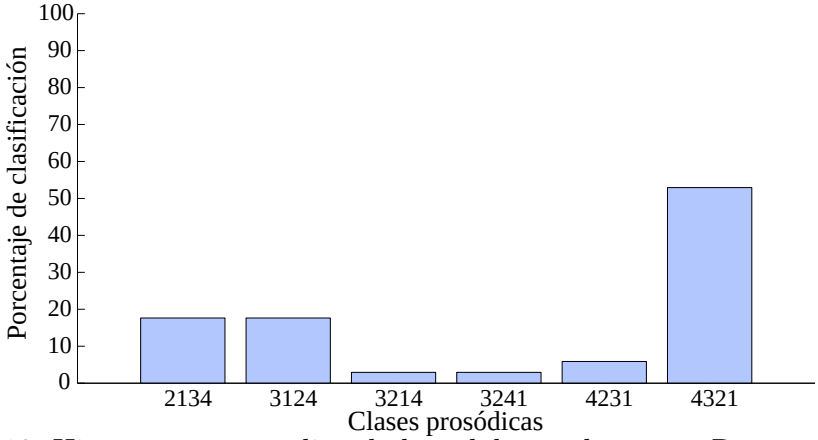

FIGURA 2.10: Histograma prosódico de la palabra *valenciana*. Rasgo: media de F_0.

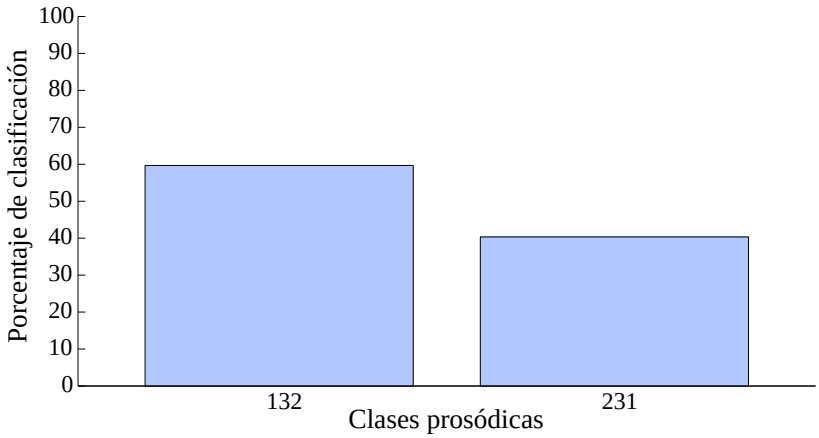

Figura 2.11: Histograma prosódico de la palabra *cúbicos*. Rasgo: mínimo de energía.

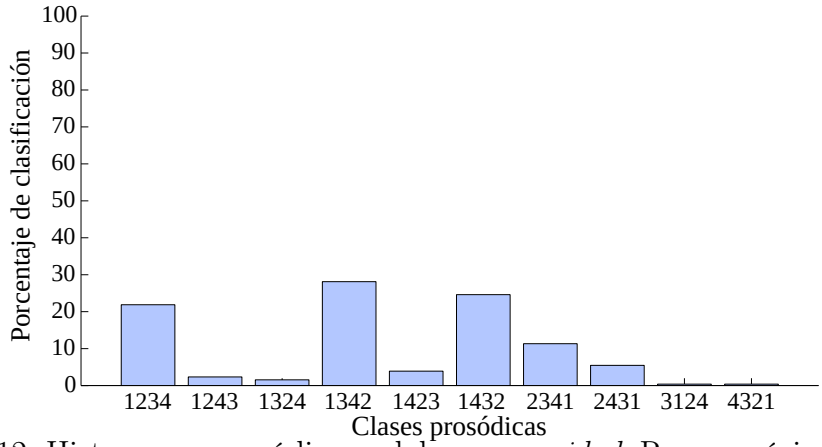

Figura 2.12: Histograma prosódico, palabra *comunidad*. Rasgo: mínimo de energía.

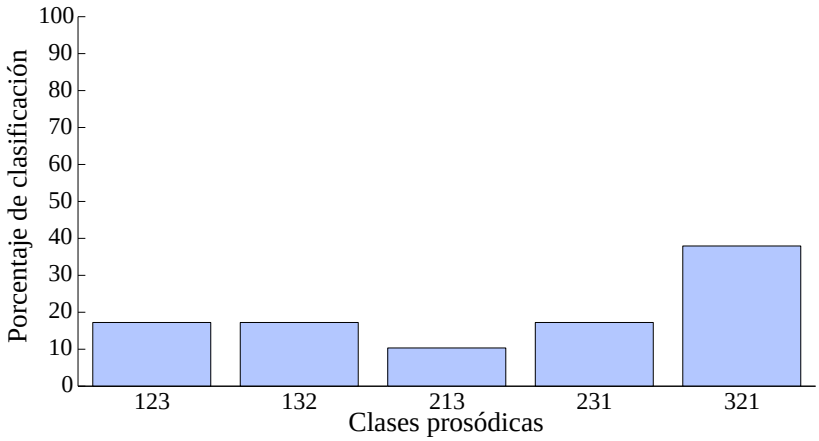

Figura 2.13: Histograma prosódico de la palabra *valencia*. Rasgo: media de F_0.

TABLA 2.2: Resultados de la caracterización de las palabras por el método de histogramas prosódicos.

Número de sílabas	Palabras evaluadas	Palabras consideradas	Nivel de caracterización [%]	Diferencia exigida
2 sílabas:	1646	25	96.00	$\geq$80 %
3 sílabas:	398	6	83.33	$\geq$40 %
4 sílabas:	792	9	77.78	$\geq$30 %
5 sílabas:	196	1	100.00	$\geq$20 %

encontró una clase, para la media de F_0, que la caracterice. En estos casos los rasgos prosódicos no aportarían ninguna información importante al sistema de RAH, al menos considerandolos independientes entre sí.

En la Tabla 2.2 se expone un resumen de los resultados. La columna *Palabras evaluadas* informa la cantidad total de palabras diferentes que existen respecto de la cantidad de sílabas; la columna *Palabras consideradas* muestra cuántas palabras cumplen con la cantidad mínima de sucesos exigida y fueron evaluadas; en la columna *Nivel de caracterización* se muestra el número de palabras que este método puede clasificar correctamente (se considera que al menos debe tener una clase bien definida para uno de los rasgos). La columna *Diferencia exigida* indica la diferencia relativa que debe presentar la clase que caracteriza a una palabra para ser considerada distinguible por el método. Se puede ver que la capacidad de discriminación del método decrece marcadamente con la cantidad de sílabas. Ésto puede ser atribuido a dos motivos, uno es que no se tienen palabras de muchas sílabas que sean representables por sus rasgos prosódicos y otro es que el método no es eficiente para determinadas cantidades de sílabas. Como ya se mencionó, si se incrementa el número de sílabas de una palabra aumenta la cantidad de clases a la que ésta puede pertenecer y por lo tanto se debe reducir la tolerancia en las diferencias exigidas. Una característica interesante que ha sido detectada en la etapa de análisis de los resultados, y que se da más frecuentemente en las palabras de tres o más sílabas, es la existencia de 2 o 3 clases dominantes. Ésto significa que para un parámetro prosódico determinado, existen 2 o 3 clases que claramente se destacan sobre las $n!$ clases posibles. Esta observación resulta interesante porque permite restringir la caracterización de una palabra a unas pocas clases prosódicas bien definidas. Esta discusión será reanudada en la siguiente sección.

VALORACIÓN DE LOS HISTOGRAMAS PROSÓDICOS

Con el fin de lograr una valoración menos subjetiva de los histogramas prosódicos, se proponen algunas medidas cuantitativas para determinar cuáles son más significati-

vos. A partir de estas definiciones y de un umbral para éstas, que se define a partir de la cantidad de sílabas y de algunas consideraciones sobre resultados preliminares, se ha logrado formalizar de alguna manera la clasificación de los histogramas. Incluso permiten considerar aquellos histogramas donde existen más de una clase que caracteriza a la palabra, y que hasta el momento se habían descartado. Las medidas propuestas se definen como:

- Media del histograma:

  $M_h = \dfrac{N_s}{n!}$, siendo n el número de sílabas y N_s la cantidad total de sucesos de la palabra h.

- Medida de pico:

  $\rho = \dfrac{P_{max}}{M_h}$, donde P_{max} es el número de sucesos de la clase más representativa.

- Medida relativa de picos:

  $\gamma = \dfrac{\rho - 1}{n! - 1}$

Se analizan todos los rasgos prosódicos de cada palabra y sólo se consideran *significativos* aquellos histogramas cuyos γ superen determinado umbral. Un umbral alto descarta mayor cantidad de histogramas, dejando aquellos bien determinados por una o pocas clases, mientras que, un umbral bajo acepta histogramas de menor calidad. Por lo tanto, se hace menester considerar el compromiso entre considerar buenos histogramas y no descartar demasiadas palabras. Con el fin de facilitar la comprensión de la forma en que estos parámetros permiten medir la significancia de los histogramas, se presentan a continuación algunos ejemplos ilustrativos. Del análisis del histograma de la palabra *valencia* para el rasgo media de F_0 (Figura 2.13) se obtienen los siguientes valores:

$n! = 6$	$N_s = 29$
$M_h = 4{,}83$	$P_{max} = 11$
$\rho = 11/4{,}83 = 2{,}28$	$\gamma = \mathbf{1{,}28/5 = 0{,}256}$

Por otra parte, si se calculan estos valores para el histograma de la palabra *desemboca* para el rasgo *máximo de energía* (Figura 2.7) se obtienen:

$n! = 24$	$N_s = 52$
$M_h = 2{,}17$	$P_{max} = 33$
$\rho = 33/2{,}17 = 15{,}21$	$\gamma = \mathbf{14{,}21/23 = 0{,}62}$

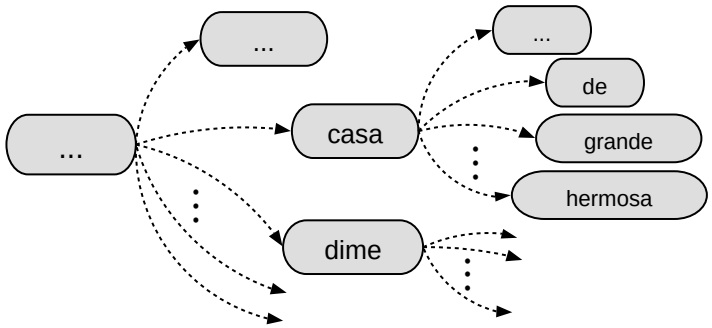

FIGURA 2.14: Red de palabras (instancia de un modelo de lenguaje).

En este segundo caso se obtiene un valor de γ alto que indica que el histograma correspondiente está bien definido, a diferencia del primer caso. Estas medidas son utilizadas para determinar qué histogramas pueden ser utilizados en la etapa de penalización que se detalla en la siguiente sección.

2.3. INCORPORACIÓN AL RECONOCEDOR DE HABLA

En esta sección se presenta un método para incorporar la información prosódica en un sistema de RAH. En primer lugar se propone un análisis de los segmentos de voz, para los que el reconocedor provee hipótesis de palabras, utilizando los histogramas prosódicos. Luego, se detalla el procedimiento que se realiza en los casos donde las hipótesis de palabras presentan una estructura prosódica que no coincide con el modelo que clasifica a dicha palabra.

2.3.1. INCORPORACIÓN EN REDES DE PALABRAS SIMPLES

Como ya se mencionó en la sección 1.2.5, un modelo de lenguaje (o red de gramática) está conformado por una lista de nodos que representan a las palabras y una lista de arcos que los une entre sí. Durante la etapa de entrenamiento, las probabilidades de transición asociadas a los arcos son computadas y normalizadas sobre todos los arcos que tienen el mismo destino. Cuando se está reconociendo una frase el reconocedor utiliza el modelo de lenguaje junto a la información acústica para generar una red de palabras (Figura 2.14). Ésta es una instancia del modelo de lenguaje conformada por un conjunto de hipótesis de palabras para diferentes segmentos de la frase y arcos de transición que recalculan sus probabilidades para la frase analizada.

El método de penalización propuesto se basa en obtener las redes de hipótesis del reconocedor y evaluar cada nodo de la red que representa una hipótesis de palabra

1. Búsqueda de las palabras (nodos) preclasificadas en la red de palabras.
2. Extracción de la información de los tiempos en que está contenida. Se contemplan todas las hipótesis.
3. Sobre el segmento de audio original:
   a. Se parametriza y se segmenta.
   b. Se evalúan los rasgos prosódicos y se asocian a las distintas sílabas.
   c. Se clasifica la palabra para cada rasgo (codificación de histogramas).
4. Comparación con las clases generales que modelan a la palabra. Y sólo en caso de **no existir coincidencia** se **penaliza**.

Algoritmo 2: Método de penalización prosódica.

localizada temporalmente. Entonces, son penalizadas aquellas hipótesis de la red que presenten estructuras prosódicas (temporalmente determinadas) diferentes a las que definen a la palabra en cuestión, según los histogramas obtenidos previamente. Los pasos básicos de este método pueden apreciarse en el Algoritmo 2.

Para dar un ejemplo ilustrativo del funcionamiento del método de penalización se propone la Figura 2.15. Aquí suponemos que se analiza una frase y que el reconocedor ha propuesto una red de hipótesis de palabras. Para cada hipótesis (en el ejemplo *dime*), se obtienen sus tiempos de inicio y fin y se extrae el segmento de audio. Para este segmento se calculan los rasgos prosódicos, se obtienen las segmentaciones de los fonemas que forman la palabra y se componen las sílabas. En el paso 3 se realiza la clasificación de las estructuras prosódicas según lo propuesto en la Sección 2.2.2. En el paso siguiente, las clases prosódicas obtenidas para el segmento analizado se comparan con las clases prosódicas que fueron calculadas previamente y definen a la palabra *dime*. En el paso 5, sólo si no existe coincidencia, se penaliza el arco de la red que conduce a la hipótesis analizada. Aquí se utilizan las redes de hipótesis generadas por el HVite [17] y la penalización consiste en disminuir la probabilidad logarítmica del modelo de lenguaje asociada a ese nodo. Luego de evaluar todas las hipótesis de la red se obtiene la red de hipótesis penalizada que será utilizada por el reconocedor.

La utilidad de este método fue explorada y se comprobó que esta información prosódica permite mejorar el desempeño de un sistema de RAH [56].

Penalización ponderada

Del análisis de los resultados preliminares del método de penalización surge que la utilización de valores fijos de penalización no permite cuantificar el grado del error cometido por cada hipótesis. Con el fin de proporcionar una mayor precisión en las penalizaciones, se definió un factor de penalización relativo a la clase que contempla la

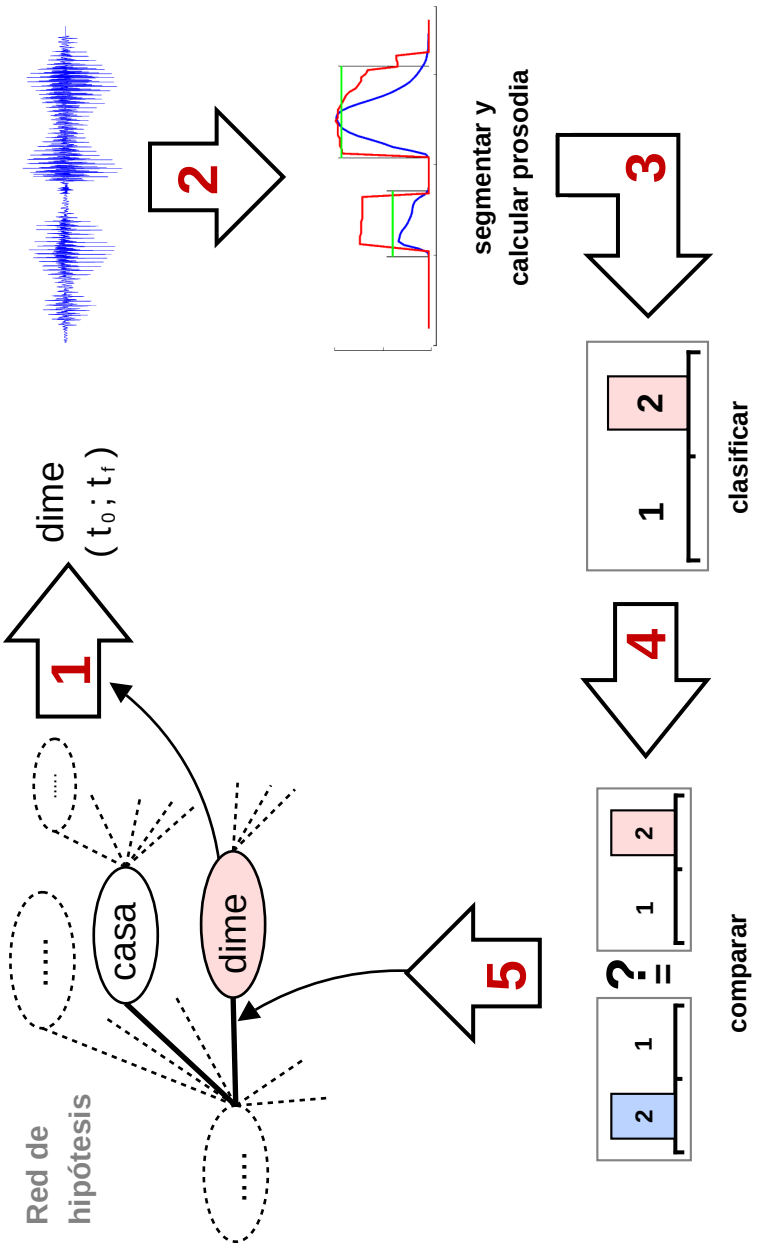

FIGURA 2.15: Diagrama general del método propuesto para la incorporación de la prosodia al sistema de RAH.

importancia de la clase obtenida en el histograma del rasgo analizado

$$F_p = 1 - P_R, \qquad \text{con } \mathrm{P_R} = \frac{\mathrm{C}}{\mathrm{C_{Max}}}$$

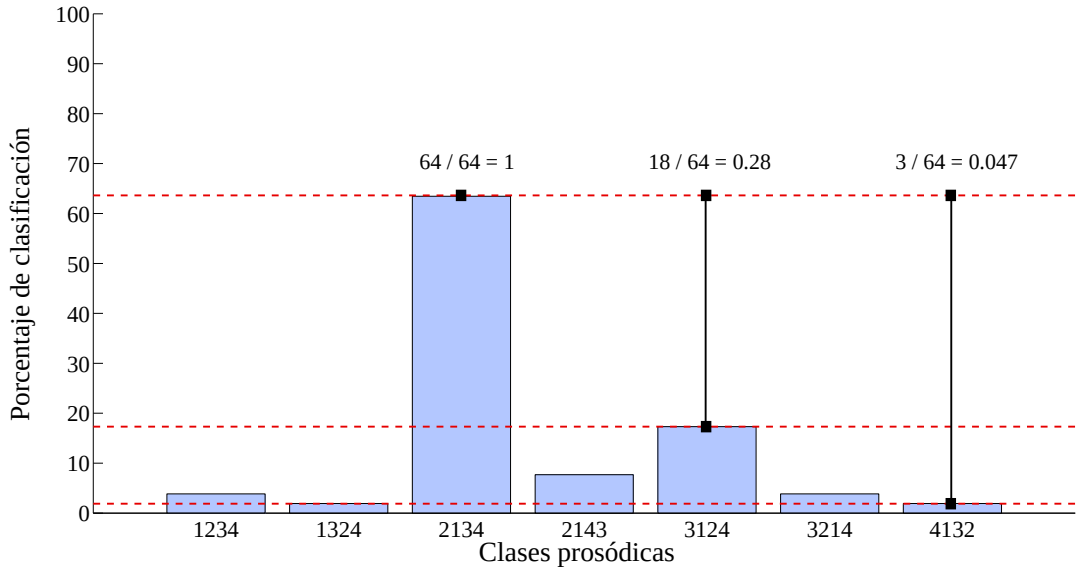

Figura 2.16: Cálculo de los valores P_R para las clases en un histograma.

donde P_R representa la importancia relativa de la clase evaluada (C) con respecto a la clase con mayor número de sucesos (C_{Max}) del histograma para el rasgo evaluado.

Finalmente, la utilización de F_p permite aplicar una mayor penalización en aquellas hipótesis de palabras cuya clase no se corresponde con la clase o las clases que caracterizan a la palabra. En los experimentos se utilizó el valor que representa el mayor error de clase (sobre todos los rasgos prosódicos), tratando a cada rasgo de forma independiente. En la Figura 2.16 se presentan los valores de importancia relativa (P_R) calculados para un histograma particular.

2.3.2. Experimentos y resultados

Un reconocedor del habla continua, con idénticas características al descripto en la Sección 2.2.2, fue utilizado como sistema de referencia para los experimentos. Para el cálculo de las características prosódicas de las señales se utilizó una ventana de 25 ms de ancho. Se utilizaron rutinas del ToFy[4] para extraer la información de la energía y la F_0 de las señales, esta última calculada con un algoritmo similar al de Noll [53] basado en coeficientes cepstrales. Las rutinas para realizar cálculos y estadísticas fueron

[4]Desarrollado en el Laboratorio de Cibernética de la Universidad Nacional de Entre Ríos (Argentina).

implementadas en *GNU C++*[5] y *Free Pascal*[6].

Con el fin de ilustrar las mejoras obtenidas con el método de penalización ponderada de la red de hipótesis, con los rasgos prosódicos propuestos en este trabajo, se presentan y analizan a continuación algunas de las frases reconocidas con penalización y sin penalización. En los resultados del reconocedor con prosodia se pueden identificar tres tipos de mejoras:

- Tipo 1 "mejoras totales": dan una solución completa y permiten un reconocimiento 100 % correcto de la frase.

- Tipo 2 "mejoras parciales": dan una solución en cuanto indican qué hipótesis de palabra (elegida por el reconocedor) no es correcta, aunque no corrigen todos los errores de reconocimiento en la frase.

- Tipo 3 "mejoras indirectas": dan una solución que indirectamente beneficia al resto de la red de palabras, lo que resulta en una frase con menos errores. Aquí se incluyen situaciones en que, si bien no se corrige una palabra prosódicamente incorrecta, se selecciona la mejor de varias hipótesis (de la misma palabra) que poseen diferentes localizaciones temporales.

RESULTADOS CON MEJORAS TOTALES

En este caso el método da una solución completa y permite un reconocimiento 100 % correcto de la frase. A continuación se presentan algunos ejemplos:

- 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.

[5]Disponible en `http://gcc.gnu.org`.
[6]Disponible en `http://www.freepascal.org`.

- 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.`

  penalizando a *nombre* y reduciendo 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. Algunos ejemplos de estos casos son:

- 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

En estos casos se incluyen situaciones donde se selecciona correctamente una de varias hipótesis (aquella con la mejor segmentación temporal), lo que indirectamente beneficia al resto de la red de palabras. A continuación se comentan algunos de ellos:

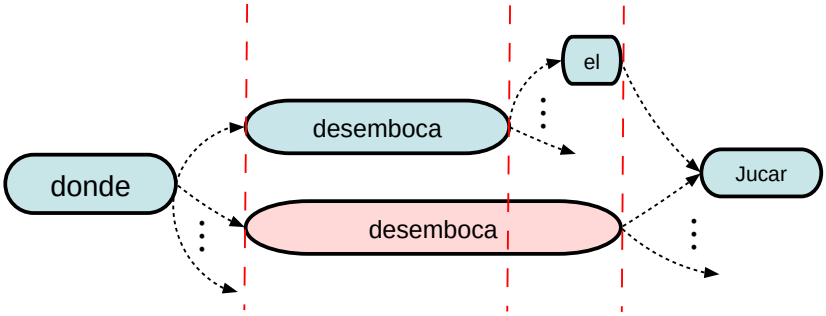

Figura 2.17: Elección de la hipótesis correcta en la red de palabras.

- 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.`

  Aquí se penaliza la hipótesis que presenta una localización temporal errónea (*desemboca*), y de esta manera se permite insertar la palabra *el* (que no está clasificada por el método pero se corrige de forma indirecta). En la Figura 2.17 se ilustra cómo se han presentado las hipótesis en la red de palabras para este caso.

- 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, la penalización de la hipótesis de *desemboca* mal ubicada en el tiempo permite corregir la 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.

DISCUSIÓN

Evaluando los errores generados por el reconocedor estándar se pudo definir de forma precisa el campo de acción del método propuesto. Considerar su desempeño en términos del error de reconocimiento de palabras (EP) no determina directamente su éxito, debido a que los errores computados en el EP no guardan relación directa con los que el método puede corregir. Se debe considerar que éste no actua sobre las palabras mal reconocidas sino sobre las probabilidades de las hipótesis posibles para una palabra mal reconocida. Ésta es una política poco invasiva y tiene una dependencia muy fuerte con las redes de hipótesis que el reconocedor otorga. Esto último implica un condicionamiento muy fuerte y es que la hipótesis correcta debe existir en la red de palabras. De otra manera, la aplicación de este método no se reflejará en una mejora del EP debido a que cualquier camino alternativo, obtenido luego de la penalización, seguirá siendo incorrecto por más buenos que sean los modelos de prosodia que se utilicen para penalizar.

Por otra parte, si la hipótesis de una palabra tiene una cantidad de sílabas incorrecta, tampoco se compara este tramo de la frase con los histogramas adecuados. Para ilustrar ésto se comenta un ejemplo de palabras entre las que se ha detectado mucha confusión: *nombre* y *número*. El método propuesto requiere conocer la cantidad de sílabas de las palabras para poder definir sus clases prosódicas y luego, en base a esta clasificación, se realiza una comparación. Si la hipótesis evaluada es *número* y esta asociada a un segmento cuya transcripción real es *nombre*, entonces se está en presencia de un error insalvable por el método propuesto.

Se ha comprobado también que tratar a los rasgos prosódicos de forma independiente en la definición del valor de la penalización no es lo más efectivo a la hora de corregir errores del reconocedor. La clasificación propuesta define a cada palabra de forma independiente para cada rasgo prosódico, por lo tanto aquellas palabras que pertenecen a una clase de un rasgo particular, pueden no compartir la misma pertenencia de clase para otro rasgo. Como al penalizar se comparan las clases prosódicas de forma independiente, y se extrae el mayor error como tasa de penalización, es comprensible que el método penalice de forma excesiva. Una alternativa para ésto podría ser la definición de un factor de penalización que contemple a todas las clases prosódicas que definen a la palabra de forma simultánea y las combine con pesos proporcionales al factor γ obtenido del histograma.

Si bien los histogramas prosódicos propuestos definen a las palabras, la pérdida de información realizada en la codificación de clases es un inconveniente cuando se trata de diferenciar a las palabras. Si se comparan de forma independiente las clases prosódicas de las palabras que habitualmente se confunden, se puede observar que éstas no son tan diferentes como para permitir una clasificación adecuada. En este marco, una redefinición de las clases donde se contemple a los rasgos de forma combinada permitiría una definición más precisa de las palabras y podría mejorar la clasificación.

Se han realizado algunas pruebas preliminares con el fin de analizar el comportamiento del método propuesto en presencia de ruido. En los experimentos realizados con ruido blanco aditivo se han observado altas tasas de error y muy pocas mejoras con la penalización prosódica. Sin embargo, ésto no es directamente atribuible al método de penalización con redes de hipótesis. En este método se utiliza la red de palabras de hipótesis, lo que supone que entre las posibilidades existe el camino correcto que buscamos. Sin embargo, si se tiene en cuenta que la parametrización con MFCC no es robusta en presencia de ruido, se puede ver que el modelado acústico no será bueno y las redes de hipótesis generadas serán malas. Luego, la penalización sobre estas redes no tendrán ninguna chance de mejorar el reconocimiento. Una alternativa para poder aplicar este método es considerar el uso de parametrizaciones y métodos de cálculo de la prosodia robustos al ruido, como los propuestos en [60].

El rendimiento del método de incorporación de la información prosódica depende fuertemente de la obtención de una red de hipótesis que contemple las palabras correctas. En casos donde no está contemplada la secuencia correcta como hipótesis, el método propuesto no presentará mejoras en el EP. Se ha verificado que aún al incrementar del número de ramificaciones en cada nodo de la red de hipótesis, no se incrementa la probabilidad de obtener la palabra buscada en algún tiempo. Una alternativa es utilizar las ideas previamente publicadas con redes expandidas [61] y así poder evaluar todas las palabras para los distintos tiempos. Se ha formulado una implementación para este esquema mediante un algoritmo incremental. En éste se propone avanzar en

la penalización palabra a palabra, de forma que para cada nueva clase la secuencia final de estructuras prosódicas (H_3, H_2, H_2, H_1, etc) se reestime en base a las mejores segmentaciones de las palabras anteriores. Las últimas ideas comentadas refieren a parte de los trabajos futuros derivados de esta Tesis.

En la siguiente sección se presenta un método que encuentra su motivación en los análisis profundos realizados sobre las redes de hipótesis. De éstos surge que las confusiones del sistema se producen generalmente entre determinadas palabras, que no necesariamente tienen el mismo número de sílabas y que presentarían una morfología acústica similar. Por lo tanto es necesario un modelo específico para cada palabra que pueda determinar que segmentos la representan y cuáles no.

2.4. CLASIFICADORES PROSÓDICOS DE PALABRAS

En esta sección se presenta un método de clasificación de hipótesis acústicas de palabras con el cual se pretende direccionar el análisis hacia aquellos segmentos en los cuales el sistema de RAH tiene problemas. El propósito de estos modelos es evaluar, para un determinado segmento acústico, la veracidad de las hipótesis presentadas por el clasificador en una red de hipótesis. En una primera parte se presenta un método novedoso de remuestreo del corpus de habla con el fin de obtener ejemplos de hipótesis ciertas y falsas de las palabras. Luego, se propone un método de clasificación para estas hipótesis.

Por lo general, el primer paso de un clasificador de palabras sería extraer segmentos acústicos y etiquetarlos de acuerdo con la palabra que le corresponde en la etiqueta real de la expresión. A partir de estos segmentos, se calculan diferentes características y se seleccionan algunas para componer las entradas de un clasificador. Este clasificador, después de la fase de entrenamiento, debe ser capaz de predecir una palabra para un conjunto de características que nunca ha visto antes. Los sistemas RAH del estado del arte (basados en HMM) [6] tienen un buen rendimiento y cometen pocos errores en condiciones de laboratorio. Por lo tanto, incorporar un clasificador general de palabras en éstos puede hacer que todo el problema aparezca nuevamente. Aquí, la atención se centra en el análisis de segmentos acústicos particulares en los que los sistemas de RAH tienen problemas.

La metodología propuesta se basa en el análisis del espacio de hipótesis de reconocimiento generado por el sistema de RAH cuando se reconoce una frase. Así, la primera etapa es el entrenamiento de un sistema de RAH basado en HMM y la generación redes N-Best para todas las frases de entrenamiento. Luego, estas redes de palabras se utilizan para hacer un muestreo del corpus y con esta información se desarrollan los clasificadores.

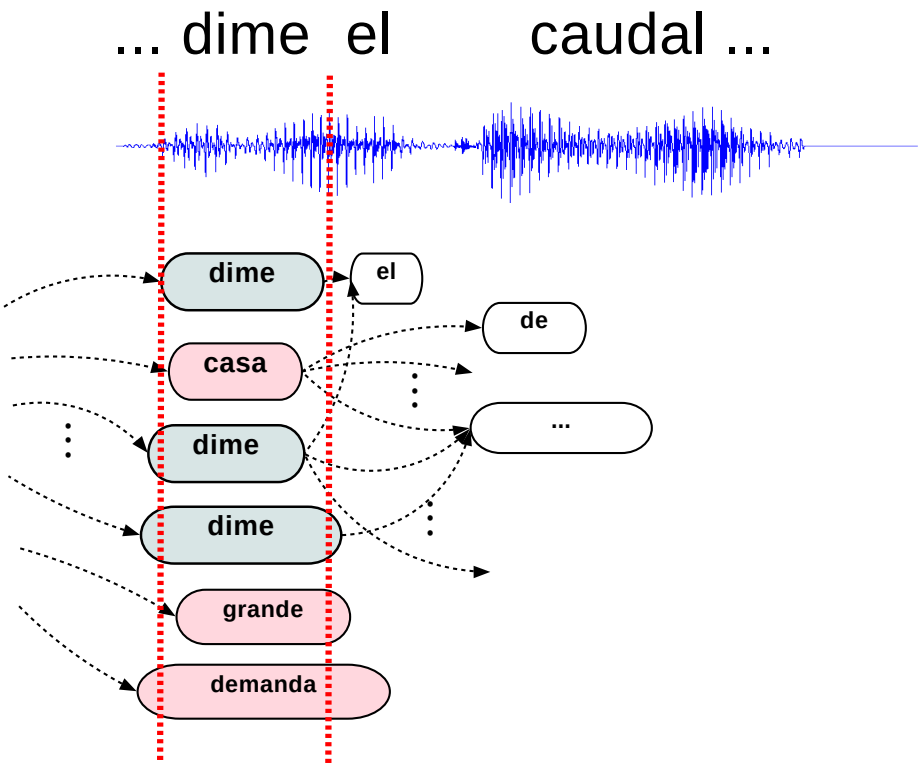

Figura 2.18: Interpretación gráfica de las hipótesis de palabras.

2.4.1. Desarrollo de un corpus de errores de reconocimiento

Las señales de voz utilizadas fueron tomadas del corpus geográfico de Albayzin [55] y se utilizaron 4.400 elocuciones correspondientes al conjunto de entrenamiento.

Para desarrollar el sistema de RAH estándar basado en HMM se utilizó nuevamente el *Hidden Markov Toolkit* [17]. La parametrización MFCC se calculó mediante ventanas de Hamming de 25 ms con un desplazamiento de 10 ms. Para modelar los fonemas se extrajeron los primeros 12 MFCC y la energía, además de sus primeras y segundas derivadas, y se generó un modelo de lenguaje basado en bigramas.

A continuación, se utilizó el sistema de RAH para obtener la lista de las N-best hipótesis para cada frase. Los segmentos acústicos fueron extraídos de las elocuciones con la información de las alineaciones de las hipótesis de las palabras. En la Figura 2.18 puede verse un ejemplo gráfico de este proceso. Para cada palabra, las hipótesis se clasifican en grupos de Verdadero o Falso en función de la correspondencia entre la hipótesis y la transcripción real de la frase.

Debido a que los reconocedores actuales cometen pocos errores en condiciones de laboratorio como las propuestas, los grupos mencionados están naturalmente desbalanceados y contienen mayor cantidad de hipótesis verdaderas. También es cierto que es muy útil contar con clases balanceadas para el entrenamiento de la mayoría de los clasificadores estándar. Por estos motivos, se definieron algunas reglas con el fin de equilibrar el nuevo corpus:

- Todas las hipótesis verdaderas, exactamente repetidas, son eliminadas debido a la gran redundancia encontrada.
- Todas las hipótesis falsas se han considerado tal cual, dado que se encontró poca redundancia en éstas.

luego, estos datos se vuelven a muestrear con el fin de balancear los conjuntos de verdaderos y falsos según:

1. Si ($cantidad(Verdadero) > cantidad(Falso)$) $\Rightarrow$ se hace un muestreo aleatorio simple sin reemplazo de los datos verdaderos.
2. Si ($cantidad(Verdadero) < cantidad(falso)$) $\Rightarrow$ el conjunto Falsos se define considerando: los datos falsos sin replicar más un subconjunto de éstos (obtenidos por un muestreo aleatorio simple sin reemplazo) con el que se balancea el conjunto Falsos respecto del conjunto Verdaderos.

2.4.2. CLASIFICADORES PROSÓDICOS

Para este análisis, fueron seleccionadas doce de las palabras que el reconocedor confundía con más frecuencia. El uso de los rasgos prosódicos en el reconocimiento de voz ha sido discutido ampliamente [41, 44, 45] y revisado en una sección previa de esta Tesis. La conocida herramienta denominada Praat[7] fue utilizada para extraer de las hipótesis los siguientes parámetros prosódicos: F_0, Energía, F_1, ancho de banda de F_1, F_2 y ancho de banda de F_2. Los valores mínimos, medios, máximos, desviación estándar, asimetría y de curtosis se calcularon para conformar el vector de características (FV, del inglés *Features Vector*). Finalmente, el FV tiene 42 características: las características mencionadas (36) más la distancia mínima y máxima entre F_1 y F_2, el error cuadrático medio entre F_1 y F_2, y las pendientes de F_0, F_1 y F_2.

Como el problema planteado requiere de un clasificador binario y una de las clases es muy variada, un modelo generativo no sería apropiado. Por lo tanto se utilizaron máquinas de soporte vectorial (SVM, del inglés *support vector machines*)[8]. SVM es

[7]Disponible en `http://www.praat.org` [62].

[8]La biblioteca *LIBSVM* [63] fue utilizada para implementarlos.

un método de aprendizaje supervisado ampliamente utilizado para la clasificación de patrones y su objetivo es encontrar un hiperplano capaz de separar los patrones de entrada en un espacio de dimensiones lo suficientemente altas [2].

El método consiste entonces en generar un clasificador binario para cada palabra. La entrada para éstos será un segmento particular con una hipótesis de palabra asociada. Estos clasificadores deben ser capaces de distinguir, en base a la información prosódica, si la hipótesis propuesta por el reconocedor es verdadera o es falsa.

2.4.3. Experimentos y resultados

En esta sección se describen en detalle los experimentos realizados. En primer lugar se comenta cómo se ha elegido el vector de características y el mejor modelo SVM para cada palabra. A continuación, se explican las pruebas de validación realizadas con estos modelos, utilizando una partición de datos nunca antes vista.

Para cada palabra se generó una partición de entrenamiento y prueba, utilizando el corpus balanceado. En éstas el 80 % de los datos fue seleccionado al azar para el entrenamiento y el restante 20 % quedó para la validación. Las etapas comentadas a continuación se realizaron con los datos de entrenamiento con las características prosódicas antes listadas en crudo y normalizadas. Con ésto se pretende evaluar la importancia de la etapa de normalización. El proceso de normalización se ha realizado sobre todos los datos de entrenamiento considerando cada dimensión del FV de forma individual. Los valores mínimos y máximos utilizados en cada dimensión se almacenan para poder normalizar los datos en la etapa de validación de forma correcta.

La técnica de *F-Score* fue utilizada para calificar las características en función de su capacidad discriminativa [63]. Dados los vectores de características FV_k, F-Score mide discriminación respecto de los dos conjuntos (aquí N_T son las instancias de Verdadero y N_F los casos de Falso) como

$$F(i) = \frac{\left(\bar{x}_i^{(T)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(F)} - \bar{x}_i\right)^2}{\frac{1}{N_T-1}\sum_{j=1}^{N_T}\left(x_{j,i}^{(T)} - \bar{x}_i^{(T)}\right)^2 + \frac{1}{N_F-1}\sum_{j=1}^{N_F}\left(x_{j,i}^{(F)} - \bar{x}_i^{(F)}\right)^2}, \tag{2.6}$$

donde $\bar{x}_i$ es el promedio de la i-ésima característica, $\bar{x}_i{}^{(F)}$ and $\bar{x}_i{}^{(T)}$ son la media para las instancias de Falso y Verdadero respectivamente, y $x_{j,i}$ es la i-ésima característica en la instancia j-ésima.

Se evaluaron cada una de las 42 características utilizando el *F-Score* y fueron categorizadas según su capacidad discriminativa. Luego, se definieron para cada palabra 12 patrones de entrada diferentes (considerando las 2, 4, 6, 8, 10, 12, 14, 16, 21, 26,

32 y 42 características más discriminatorias, respectivamente). Por cada subconjunto de características, se exploró un espacio definido de parámetros con el fin de lograr el mejor modelo SVM para la palabra. Para todos los modelos SVM se utiliza un núcleo de funciones de base radial (que en pruebas preliminares arrojó el mejor desempeño para este problema) y su rendimiento se calcula en un esquema de validación cruzada de cinco particiones:

1. Cinco conjuntos de datos disjuntos son determinados al azar de los datos de entrenamiento.

2. Cada conjunto es utilizado para validar un modelo entrenado con el resto de datos de entrenamiento.

3. El rendimiento del modelo es calculado como el promedio del rendimiento sobre todos los conjuntos.

Finalmente, se obtuvieron los mejores parámetros para cada subconjunto de características y sus resultados se muestran en las Tablas 2.3 y 2.4, para los datos de entrenamiento crudos y normalizados respectivamente. Las filas de estas Tablas representan las palabras consideradas y cada columna exhibe el mejor desempeño para cada subconjunto de características utilizado. Se resaltan (en negritas) los mejores resultados obtenidos para cada palabra. Se puede notar que con los datos normalizados siempre se logran mejores resultados, salvo para la palabra *DESEMBOCAN* donde se pierde desempeño y para el caso de *NUMERO* donde el porcentaje alcanzado se mantiene. También es importante destacar que, para el caso de los datos normalizados, se requiere un mayor número de características para alcanzar ese buen desempeño. Finalmente, para cada grupo de datos (crudos y normalizados) y por cada palabra se almacenó información acerca del subconjunto de características que logró el mejor desempeño y cuales fueron los parámetros del modelo considerado.

En la etapa de validación se entrenó un nuevo modelo de SVM con todos los datos de entrenamiento para cada palabra, con la configuración que logró el mejor rendimiento. Todos los modelos SVM fueron testeados con las particiones de validación antes mencionadas y los resultados se pueden observar en las Tablas 2.5 y 2.6, para los datos en crudo y normalizados respectivamente. En la primer columna de las tablas están listadas las palabras consideradas, en la segunda se informa el tamaño del subconjunto de datos utilizado y en la tercer columna se muestra el desempeño obtenido. Es posible observar en estas tablas que en ambas pruebas se logran buenos resultados para la clasificación de hipótesis de palabras. Como se puede observar en las últimas filas de las tablas, el desempeño promedio logra una mejora de $4,5\,\%$ cuando se utilizan los datos normalizados. Sin embargo, debe notarse que el proceso de normalización no ha resultado útil para todas las palabras. Por ejemplo, *MENOR* alcanzó un $77,19\,\%$ con

Tabla 2.3: Mejores resultados para datos de entrenamiento y validación cruzada, para diferentes cantidades de características en crudo (%). (En negritas se informan los mejores modelos.)

	Subconjuntos de características											
	42	*32*	*26*	*21*	*16*	*14*	*12*	*10*	*8*	*6*	*4*	*2*
CABO	61,11	63,89	63,89	63,89	63,89	63,89	70,83	76,39	77,78	**79,17**	76,39	62,50
CAUDAL	76,52	80,30	80,30	80,68	80,68	79,55	80,68	**85,61**	84,85	84,47	80,30	77,65
DESEMBOCA	80,38	80,38	80,38	80,38	80,38	80,38	80,38	80,38	80,38	**84,21**	80,38	74,64
DESEMBOCAN	79,79	79,79	79,79	83,94	84,72	84,72	84,72	84,72	84,72	**85,49**	**85,49**	63,73
MENOR	**75,76**	**75,76**	**75,76**	**75,76**	**75,76**	**75,76**	**75,76**	**75,76**	**75,76**	**75,76**	**75,76**	74,89
MENOS	**81,63**	**81,63**	**81,63**	**81,63**	**81,63**	**81,63**	**81,63**	**81,63**	**81,63**	**81,63**	**81,63**	**81,63**
NOMBRE	86,75	86,75	87,73	87,83	**87,93**	87,73	87,54	87,63	87,63	86,95	86,75	79,49
NUMERO	84,85	84,85	84,85	84,85	84,85	84,85	84,85	84,85	84,85	84,85	**89,39**	87,88
PASA	79,17	79,17	80,11	80,11	80,30	80,11	79,36	79,36	79,36	79,55	**80,49**	73,48
PASAN	56,22	56,22	56,22	56,22	56,22	56,22	56,22	56,22	57,25	59,33	61,66	**65,80**
TIENE	52,66	52,66	52,66	52,66	52,66	52,66	52,66	**76,86**	75,27	73,27	71,94	65,82
TIENEN	69,08	69,08	69,08	71,60	72,27	72,10	**73,95**	73,45	73,61	70,92	68,57	64,71

Tabla 2.4: Mejores resultados para datos de entrenamiento y validación cruzada, para diferentes cantidades de características normalizadas (%). (En negritas se informan los mejores modelos.)

	Subconjunto de características											
	42	*32*	*26*	*21*	*16*	*14*	*12*	*10*	*8*	*6*	*4*	*2*
CABO	77,78	81,94	77,78	81,94	86,11	86,11	**90,28**	73,61	75,00	75,00	75,00	66,67
CAUDAL	**89,77**	89,02	87,50	85,61	83,33	82,95	84,47	86,36	85,23	81,82	76,14	74,62
DESEMBOCA	84,93	**85,65**	85,41	82,78	84,21	82,78	83,49	81,34	81,82	71,05	65,79	61,24
DESEMBOCAN	80,83	80,05	**81,87**	79,53	79,02	78,50	77,72	75,13	78,24	69,95	62,69	58,55
MENOR	88,31	88,31	**89,18**	85,71	87,01	86,58	85,28	86,15	84,85	83,12	75,32	73,16
MENOS	86,39	**87,07**	85,71	86,39	86,39	85,71	85,71	85,03	82,99	84,35	82,31	73,47
NOMBRE	**88,32**	**88,32**	88,22	87,34	85,87	86,46	83,91	80,77	79,39	73,80	72,42	71,64
NUMERO	**89,39**	86,36	84,85	86,36	80,30	78,79	83,33	84,85	81,82	81,82	81,82	75,76
PASA	84,28	83,90	**84,47**	83,14	81,82	79,73	79,17	77,46	74,43	74,05	69,89	69,32
PASAN	74,61	**76,42**	74,35	75,13	72,80	74,09	75,13	74,61	72,54	68,13	69,43	65,54
TIENE	**78,19**	77,79	75,93	76,46	73,01	73,54	73,40	71,68	69,81	68,35	66,22	63,16
TIENEN	**75,97**	74,62	72,61	72,94	72,61	70,42	71,26	66,22	67,56	64,37	63,53	64,20

TABLA 2.5: Resultados de la clasificación de palabras con datos en crudo.

Palabras	Subconjunto seleccionado	Clasificación[%]
CABO	6	66,67
CAUDAL	10	74,24
DESEMBOCA	6	89,42
DESEMBOCAN	6	85,42
MENOR	42	77,19
MENOS	42	67,57
NOMBRE	16	90,20
NUMERO	4	75,00
PASA	4	81,06
PASAN	2	56,25
TIENE	10	82,98
TIENEN	12	85,91
Clasificación promedio		**77,66**

TABLA 2.6: Resultados de la clasificación de palabras con datos normalizados.

Palabras	Subconjunto seleccionado	Clasificación[%]
CABO	12	66,67
CAUDAL	42	84,85
DESEMBOCA	32	89,42
DESEMBOCAN	26	82,29
MENOR	26	91,23
MENOS	32	83,78
NOMBRE	42	85,49
NUMERO	42	81,25
PASA	26	81,82
PASAN	32	77,08
TIENE	42	75,00
TIENEN	42	86,57
Clasificación promedio		**82,12**

las características en crudo y un $91,23\,\%$ utilizando las características normalizadas, mientras que para *NOMBRE* se obtuvo un $90,20\,\%$ con las características en crudo y un $85,49\,\%$ con las normalizadas. Los resultados sugieren que para cada palabra se debe ajustar la configuración del modelo así como el proceso de normalización con el fin de mejorar el rendimiento promedio de los reconocedores.

El método propuesto se puede interpretar como un esquema de clasificación *uno-contra-todos*, donde *uno* representa las hipótesis verdaderas y *todos* a las hipótesis falsas (que son muchas y diversas). Por lo tanto, el clasificador debería ajustarse a la región fronteriza para la clase verdadera y el resto del espacio debería corresponder a las clases falsas. Siguiendo este razonamiento, el clasificador podría considerarse una primer aproximación para manejar hipótesis de palabras nunca vistas y palabras fuera de vocabulario.

Por otra parte es importante destacar que, si bien los experimentos se realizaron con un corpus en Español, ésto no implica una restricción ya que la propuesta no incluye ninguna regla dependiente del idioma y sería potencialmente aplicable en cualquier otro.

El método que se ha presentado aquí permite la clasificación de las hipótesis de palabras utilizando rasgos prosódicos y logra buenos resultados. Como este método se ha diseñado con la idea de complementar el sistema de RAH tradicional, permite encarar las confusiones reales de los sistemas de una forma más eficiente. El paso siguiente, dentro de los trabajos futuros, es la incorporación de este método dentro del sistema de RAH actual utilizando el método presentado de penalización de las redes de hipótesis.

En este capítulo se han presentado los avances realizados en el modelado de estructuras prosódicas para el RAH. Se desarrollaron dos métodos orientados a la modelización prosódica de las palabras. El método de los histogramas prosódicos ha demostrado ser útil para modelar las palabras según su separación silábica y su incorporación al sistema de RAH, utilizando el método de penalización de redes de hipótesis, mostró mejoras en el desempeño. Sin embargo, como el modelo es dependiente de los resultados parciales del sistema tradicional ha sido necesario un cambio de perspectiva en el tratamiento de esta información. El tratamiento de las hipótesis de segmentos conflictivos para el sistema de RAH revela un punto de vista muy interesante y a partir de los resultados obtenidos es posible estimar buenos resultados en su futura incorporación al RAH.

Capítulo 3

Reconocimiento de emociones en el habla

El reconocimiento del estado emocional de un hablante es un área multidisciplinar de investigación que ha recibido mucha atención en los últimos años. Uno de sus objetivos es el de mejorar las interacciones entre humanos y máquinas. Muchos trabajos recientes de investigación han focalizado sus estudios en las características prosódicas y espectrales de las señales de voz. Para el reconocimiento de emociones se utilizan una amplia variedad de técnicas estándar y no tan estándar de clasificación de patrones y otras comúnmente utilizadas en el reconocimiento automático del habla. En este capítulo se presentará inicialemente una revisión de los antecedentes en éste área, y el conjunto de datos con el que se trabajó para luego introducir la tarea de clasificación o reconocimiento de emociones. En una primera parte se presentan dos clasificadores estadísticos estándar para los que se evaluan y discuten sus configuraciones y desempeños [34]. En la sección siguiente se propone un análisis acústico detallado de las características prosódico-espectrales de las emociones y su utilización en mapas autorganizativos para identificar agrupamientos. En la última sección se presenta un novedoso método jerárquico de clasificación basado en los resultados de los agrupamientos, se presentan los resultados del método y se comparan con los clasificadores estándar [64].

3.1. Antecedentes

Como fue mencionado anteriormente, existen muchas aplicaciones de la vida real que motivan los estudios en este campo. Por ejemplo, en [65] se estudió un sistema de detección de emociones de la vida real con un corpus de diálogos entre agentes y clientes en un centro de llamadas para emergencias médicas. Para aplicaciones de seguridad, se

investigó la detección de la manifestación emocional del miedo en situaciones anormales [66]. En Tacconi et al. [67] se presenta el desarrollo de un sistema semiautomático para el diagnóstico de enfermedades psiquiátricas. También se ha utilizado en la detección de actitudes emocionales en diálogos espontáneos de interacciones de chicos con personajes de computadora [68]. En [69] se realizó un estudio para detectar cuando una persona está diciendo una mentira.

Con respecto a qué información es relevante para capturar los estados emocionales, está claro que la utilización de bio-señales (tales como ECG, EEG, temperatura corporal, etc.), imágenes del rostro y del cuerpo son alternativas muy interesantes [70, 71, 72]. Sin embargo, resulta bastante evidente que los métodos utilizados para obtener y utilizar estas señales son usualmente demasiado invasivos, complejos y/o imposibles de utilizar en ciertas aplicaciones reales. En estas condiciones, está claro que la utilización de las señales de voz se convierte en una de las opciones más viables.

Para la etapa de clasificación propiamente dicha, se han explorado una vasta variedad de métodos tradicionales de clasificación de patrones que al parecer han alcanzado su mejor desempeño. En los últimos años, se han presentado avances en el uso de combinaciones de métodos tradicionales. Un esquema de fusión de métodos se propuso en [22], donde se combinan los resultados en la etapa de decisión a partir de las salidas de clasificadores individuales (entrenados con diferentes tipos de caracteríticas). En [73], se propuso una idea similar orinetada a la discriminación entre risas y voz. Los autores presentan dos formas de combinar las salidas de los clasificadores: como una combinación lineal de las salidas de los clasificadores independientes y con un clasificador de segundo nivel entrenado con las salidas de un conjunto fijo de clasificadores independientes. Morrison et al. [74] utilizaron dos métodos de clasificación aplicados al reconocimiento de emociones: generalización de clasificadores apilados (del inglés, *stacked generalization*) y votación no ponderada (del inglés, *unweighted vote*). Estos clasificadores mejoran modestamente el rendimiento de los métodos de clasificación tradicionales. En [75], se presentó un clasificador de múltiples etapas basado en máquinas de soporte vectorial. En cada etapa se hace una clasificación entre dos clases y ésto se repite hasta que sólo queda una clase. Las clases más dificiles de separar son divididas al final. Las particiones de clases están basadas en conocimiento experto o derivado de las matrices de confusión de un clasificador de una etapa para múltiples clases basado en SVM. En el trabajo realizado por Fu et al. [76] se presentó un clasificador de dos etapas para clasificar cinco emociones. Los autores utilizaron un clasificador SVM para separar las cinco emociones en dos grupos. Luego, se utilizan HMMs para clasificar las emociones dentro de cada grupo. En [77], se utilizaron clasificadores de regresión logística bayesiana y SVM en un árbol de decisión binario. El orden de clasificación en cada capa de la decisión binaria está motivada por la teoría de valoración de las emociones [78]. También se ha propuesto un clasificador binario de múltiples etapas

Tabla 3.1: Elocuciones del corpus agrupadas por el tipo de emoción.

Clase de emoción	Anger	Boredom	Disgust	Fear	Joy	Sadness	Neutral
Nº de frases	127	81	46	69	71	62	79

conducido por el modelo dimensional de emociones [79].

En todos estos trabajos se puede notar que, a pesar de utilizar métodos tradicionales para la reducción del espacio de características, no se han realizado análisis acústicos profundos. En [80] se realizó un análisis simple de la distribución de probabilidad de los estados del modelo HMM obtenidos para diferentes tipos de emociones. Sin embargo, las razones de los aciertos y fallas en una matriz de confusión no son analizados usualmente. Por ejemplo, en [75] y [76], el agrupamiento se realiza por medio de las matrices de confusión de clasificadores estándar, conocimiento experto o las bondades del método de SVM. Si bien pueden verse que algunas formas de agrupamiento de emociones tienen base en las características acústicas, la mayor parte de los trabajos que utilizan grupos de emociones han adoptado modelos importados de la psicología.

3.2. Corpus de emociones

Como las manifestaciones emocionales se presentan de forma muy cambiante en las conversaciones naturales (no actuadas), el objetivo actual es lograr el reconocimiento de emociones utilizando habla espontánea. Por otra parte, el desarrollo de conjuntos de datos de habla espontánea es muy costoso y generalmente las bases de datos existentes tienen acceso restringido. A pesar de que las expresiones emocionales actuadas pueden no sonar exactamente como las expresiones reales, éstas se presentan como una aproximación interesante. Además, es posible advertir que estos conjuntos de datos se vuelven más apropiados si las personas que expresan las emociones no son actores y la naturalidad de las expresiones es oportunamente juzgada por expertos. En este contexto, las señales de voz emocional que se utilizaron aquí fueron tomadas de una base de datos de habla emocional en Alemán, desarrollada por el Instituto de Ciencias de la Comunicación de la Universidad Técnica de Berlín [81]. Este corpus actuado es muy reconocido y ha sido utilizado en una gran cantidad de estudios previos [30, 36, 82, 83] [1]. El corpus consiste en 535 elocuciones, donde se incluyen frases que expresan 6 emociones muy utilizadas y frases expresadas en un estado emocional neutral. Este corpus prácticamente cubre el conjunto de emociones "Grandes seis" [20] excepto porque tiene *aburrimiento* (en inglés: *boredom*) en vez de *sorpresa*. En la Tabla 3.1 se muestra la distribución de las frases en el corpus respecto de los tipos de emociones manifestadas.

[1] El corpus está disponible (libre) en http://pascal.kgw.tu-berlin.de/emodb/.

Los mismos textos fueron actuados y producidos en idioma alemán por diez personas, 5 mujeres y 5 hombres. Ésto permite hacer estudios sobre todo el grupo, comparaciones entre emociones y comparaciones entre hablantes. El corpus fue definido en base a 10 elocuciones de cada tipo de emoción, 5 frases cortas y 5 frases largas, de 1 a 7 segundos. Para lograr una mejor calidad en los registros de audio, las frases fueron grabadas en una cámara anecoica con una frecuencia de muestreo de 48kHz (luego fueron submuestreadas a 16kHz) y fueron cuantizados a 16 bits por muestra. Finalmente, fue realizado un test de percepción con 20 personas para asegurar la calidad emocional y la naturalidad de las elocuciones, y aquellas más confusas fueron eliminadas[2].

3.3. CLASIFICACIÓN DE EMOCIONES CON MÉTODOS ESTÁNDAR

Como fue introducido en la Sección 3.1, se han explorado muchos métodos para el reconocimiento de emociones, algunos estándar y otros no tanto para el área. La mayor parte de los trabajos citados utilizan los métodos pero no exploran detalladamente sus posibles configuraciones. En esta sección se presenta un estudio de las configuraciones y rendimientos de dos modelos estadísticos, uno estático y otro dinámico: los modelos de mezclas de Gaussianas y lo modelos ocultos de Markov [34]. Para los modelos ocultos de Markov se utilizan variadas configuraciones, que incluyen un análisis del número óptimo de estados para esta tarea. Los resultados muestran la influencia del número de componentes Gaussianas y de estados. El desempeño de los clasificadores ha sido evaluado para distintas cantidades de emociones, desde 3 hasta 7 emociones, en un marco que permita obtener resultados independientes del hablante.

3.3.1. DEFINICIÓN DEL SISTEMA DE RECONOCIMIENTO

Para el reconocimiento de las emociones se ha utilizado el esquema de clasificación tradicional, donde primero existe una etapa de extracción de características y luego una etapa de clasificación propiamente dicha. En la Figura 3.1 se presenta, a modo de ejemplo, un esquema de clasificación para 7 emociones.

En la primera etapa se realiza la parametrización de las señales de voz, donde se procesan las señales por tramos utilizando una ventana de Hamming de 25 ms con un paso de 10 ms. Se han obtenido los primeros 12 coeficientes MFCC y se calcularon sus primeras y segundas derivadas [17]. Se realizaron experimentos donde se comprobó la utilidad de los coeficientes de derivadas y aceleraciones.

[2]Las frases fueron eliminadas cuando los errores de reconocimiento fueron mayores al 20 % y cuando fueron juzgadas como no naturales por más del 40 % de los escuchas.

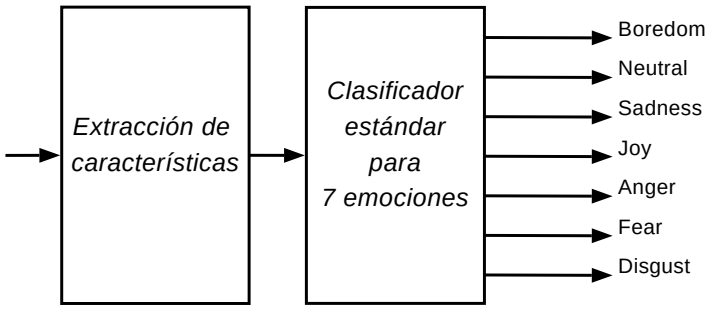

Figura 3.1: Esquema de un clasificador estándar de un nivel.

Para la segunda etapa se implementaron, por un lado clasificadores basados en GMM y por el otro clasificadores basados en HMM. Para ambos modelos se probaron diversas configuraciones que se comentan en detalle en la siguiente sección. Para la implementación de los modelos se utilizó la biblioteca llamada *Hidden Markov Toolkit* (HTK) [17], que ha sido citada en el capítulo anterior.

3.3.2. Experimentos y evaluación

El objetivo del sistema es identificar la emoción que se ha expresado en cada elocución. Por lo tanto, las transcripciones de las frases han sido ignoradas y cada una ha sido asociada a una etiqueta que indica la emoción expresada. De esta forma, cada frase representa un único patrón, de entrenamiento o test según sea el caso.

La estimación de la tasa de reconocimiento de un sistema puede estar sesgada, a favor o en contra, si se utiliza una sóla partición de entrenamiento y prueba. Para evitar una posible tendencia, aquí se ha utilizado validación cruzada (del inglés, cross validation) con el método de dejar afuera k patrones [84]. Se generaron 10 particiones y para cada una se seleccionó de forma aleatoria un 80 % de las frases para entrenamiento y el 20 % restante se conservó para validación.

Se desarrollaron varios modelos GMM con diferentes números de Gaussianas en la mezcla. A partir de un prototipo inicial definido con una Gaussiana, cada modelo se generó adiccionando Gaussianas al modelo precedente. Entonces, se definió un modelo con 1 Gaussiana, luego con 2 Gaussianas y los posteriores se desarrollaron adiccionando dos Gaussianas cada vez.

Por otro lado, para evaluar el comportamiento de los modelos HMM se realizaron modelos con diferentes configuraciones. En primer lugar se definió un modelo HMM con 2 estados y con una Gaussiana en las mezclas (prototipo inicial), y a partir de éste se generaron otros modelos, incrementando el número de Gaussianas en las mezclas de los estados de forma similar a lo descripto para GMM. Luego, al prototipo inicial se le

incorporó un estado (modelo de 3 estados) y de forma similar a lo comentado antes se definieron otros modelos incrementando el número de Gaussianas. El desarrollo de los modelos se continuo hasta alcanzar un modelo con 7 estados.

Cada modelo desarrollado fue evaluado para diferentes cantidades de emociones, hasta alcanzar las 7 emociones. La evaluación de los modelos se comenzó con 3 emociones (neutral, joy y anger) y se realizaron las demás pruebas a medida que se adiccionaban, una a una, las emociones restantes.

3.3.3. RESULTADOS Y DISCUSIÓN

En esta sección se pesentan, además de las pruebas de los modelos con las diferentes configuraciones, las comparaciones de los mejores sistemas obtenidos para GMM y HMM.

Para los modelos GMM se evaluan las cantidades óptimas de Gaussianas para reconocer las distintas cantidades de emociones. Por otra parte, el primer paso en la evaluación de los modelos HMMs es determinar el número de estados óptimos para el modelo. A continuación se presentan gráficamente los resultados del análisis de la cantidad óptima de estados para el modelo HMM. En la Figura 3.2 se exhiben los resultados de reconocimiento del modelo HMM en función del número de estados del mismo. Cada barra representa, en promedio, la tasa de reconocimiento de los 17 modelos con K Gaussianas en las mezclas de los estados ($K \in [1, 2, 4, \ldots, 32]$). Como se puede observar en la figura, no resulta necesario incrementar el número de estados del modelo HMM más allá de 2. Este comportamiento podría ser consecuencia de considerar a cada frase completa como un único patrón de entrenamiento. Probablemente existen demasiados segmentos de análisis para la misma frase con una variabilidad muy alta entre ellos. De esta forma, el modelo de 2 estados estaría capturando las aproximaciones a estas variabilidades y la incorporación de unos pocos estados sólo estaría incrementando el número de parámetros a entrenar sin incorporar información relevante. En la misma línea de pensamiento, para poder aproximar mejor la variabilidad a lo largo de las frases completas se deberían considerar modelos formados por una cantidad de estados de alrededor de 2 órdenes de magnitud mayor a las propuestas, sin embargo, la cantidad de parámetros a entrenar sería incrementada enormemente.

Considerando el mejor modelo HMM como aquel con 2 estados, a continuación se presenta un análisis de su desempeño en relación al número de componentes Gaussianas presentes en cada estado. En la Figura 3.3 puede observarse el comportamiento de un modelo HMM de 2 estados en función del número de Gaussianas, clasificando distintas cantidades de emociones. Se puede notar que, en el rango comprendido entre 14 y 22 Gaussianas, el rendimiento deja de incrementarse notablemente para mantenerse casi monótono. Finalmente, es posible obtener el número óptimo de Gaussianas según sea

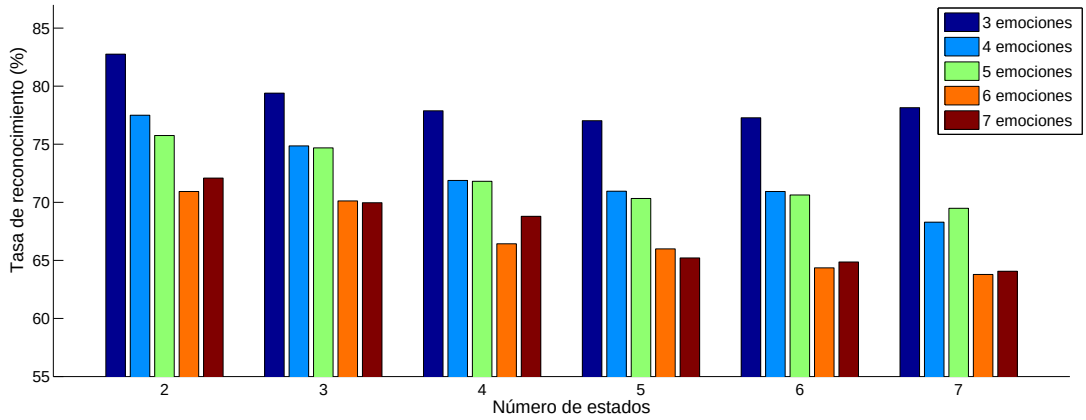

Figura 3.2: Tasa de reconocimiento promedio en función del número de estados del modelo HMM.

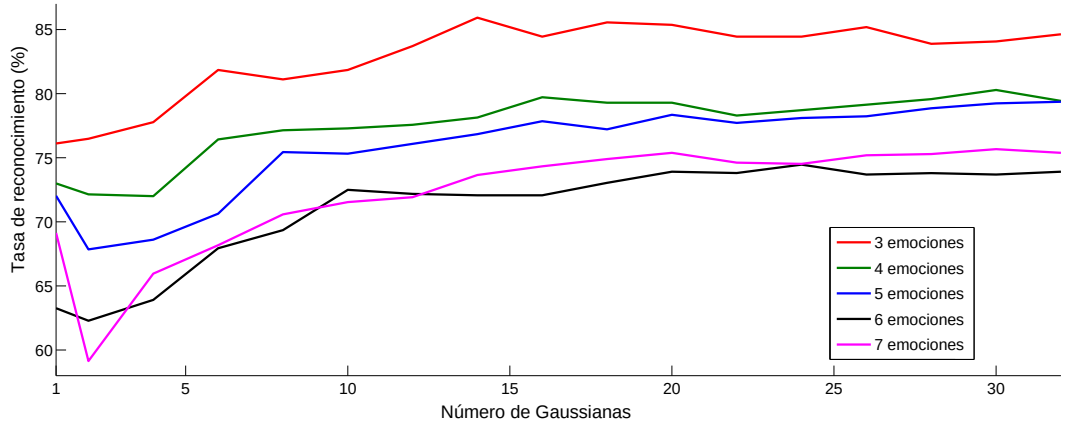

Figura 3.3: Tasa de reconocimiento promedio en función del número de Gaussianas para HMM de 2 estados.

el número de emociones que se clasifiquen.

Aunque se realizaron experimentos para todos los modelos GMM y para todas las combinaciones de estados y componentes Gaussianas en el análisis de los HMMs, se presentarán detalladamente los resultados de aquellos que se consideran más relevantes. Las matrices de confusión son una buena forma de presentar los datos pues permiten localizar los errores de clasificación de una forma precisa. Las columnas se corresponden con las entradas (etiquetas reales de las elocuciones) y las filas representan las salidas del clasificador. En la diagonal principal se pueden observar los aciertos, mientras que

TABLA 3.2: Matriz de confusión de un GMM con 22 Gaussianas para 3 emociones.

Emoción	Joy	Anger	Neutral	Aciertos
Joy	**99**	38	3	70.71 %
Anger	50	**192**	8	76.80 %
Neutral	11	4	**135**	90.00 %

TABLA 3.3: Matriz de confusión de un GMM con 32 Gaussianas para 7 emociones.

Emoción	Joy	Fear	Disgust	Sadness	Anger	Boredom	Neutral	Aciertos
Joy	**101**	9	2	0	25	0	0	73.72 %
Fear	21	**62**	2	16	17	4	4	49.21 %
Disgust	4	12	**67**	8	3	3	0	69.07 %
Sadness	0	0	1	**100**	0	14	5	83.33 %
Anger	23	6	1	0	**220**	0	0	88.00 %
Boredom	0	6	15	26	0	**63**	50	39.38 %
Neutral	2	4	6	21	3	30	**84**	56.00 %

en el resto de la matriz se encuentran los errores. De esta forma es posible identificar las sustituciones realizadas de forma precisa.

La Tabla 3.2 presenta la matriz de confusión obtenida al reconocer 3 emociones, con el mejor modelo GMM desarrollado para esta tarea. La tasa de reconocimiento alcanzó un 79 % y se puede observar una confusión importante entre las emociones *Joy* y *Anger*. En la Tabla 3.3 se expone la matriz de confusión resultante del reconocimiento de 7 emociones con el mejor modelo GMM (con 32 componentes Gaussianas en la mezcla). La tasa de aciertos es 67 %. Aquí las mayores confusiones se presentan entre *Joy* y *Anger*, *Boredom* y *Neutral*, *Boredom* y *Sadness*. Por otro lado, puede notarse que hay emociones que nunca se confunden.

En las Tablas 3.4 y 3.5 se presentan las matrices de confusión para 3 y 7 emociones, obtenidas por los modelos HMM de 2 estados. Se puede ver que para 3 y 7 emociones, las tasas de reconocimiento fueron de 86 % y 76 % respectivamente. Los resultados presentados en ambas tablas, también exhiben confusiones muy altas entre *Joy* y *Anger*. En la Tabla 3.5 puede observarse un decremento general en la cantidad de confusiones que comete el reconocedor respecto de los resultados con GMM. Sin embargo, para *Joy* y *Sadness* se han producido menos aciertos. Por otro lado, las confusiones siguen siendo altas entre *Joy* y *Anger*, *Boredom* y *Neutral*, *Boredom* y *Sadness*.

En los experimentos se pudo notar que es poco útil seguir incrementando el número de Gaussianas más allá de 32. Además, no fue posible determinar una cantidad óptima de Gaussianas para uso general, por lo que es necesario ajustar esta cantidad en cada

Tabla 3.4: Matriz de confusión de un HMM de 2 estados para 3 emociones (14 Gaussianas).

Emoción	Joy	Anger	Neutral	Aciertos
Joy	**101**	39	0	72.14 %
Anger	32	**216**	2	86.40 %
Neutral	2	1	**147**	98.00 %

Tabla 3.5: Matriz de confusión de un HMM de 2 estados para 7 emociones (30 Gaussianas).

Emoción	Joy	Fear	Disgust	Sadness	Anger	Boredom	Neutral	Aciertos
Joy	**93**	13	0	0	34	0	0	66.43 %
Fear	13	**93**	5	6	7	3	3	71.54 %
Disgust	4	7	**70**	0	4	3	2	77.78 %
Sadness	0	0	3	**93**	0	23	1	77.50 %
Anger	12	6	1	0	**231**	0	0	92.40 %
Boredom	0	3	7	14	0	**94**	42	58.75 %
Neutral	0	5	5	0	0	27	**113**	75.33 %

experimento. Los resultados presentandos exponen que los modelos HMM tienen un mejor desempeño que los GMM.

La utilización de la prosodia en las parametrizaciones es una alternativa muy interesante y existen trabajos que validan su uso [29, 31, 35]. En experimentos posteriores se ha encontrado una mejora interesante al incluir una característica prosódica en la parametrización propuesta. Se realizaron pruebas similares considerando la energía, su derivada y aceleración en la parametrización. En la Tabla 3.6 se presentan los rendimientos de los mejores modelos (HMM de dos estados) para distintas cantidades de emociones. En la segunda columna se listan los resultados de los experimentos que se comentaron previamente y en la tercer columna los resultados con la parametrización que incluye información de la energía. Se puede notar que la incorporación de estos valores en la parametrización permite mejorar el rendimiento del sistema.

Si bien los resultados obtenidos son muy interesantes, en el análisis se ha omitido, como en casi todas las investigaciones citadas, el desbalance existente entre las distintas clases de emociones (Tabla 3.1). Esta omisión no es menor, puesto que las clases más cuantiosas sesgan los resultados y debe considerarse un punto vital a la hora de comparar métodos. Este tema será abordado en la siguiente sección.

Es importante notar que la elección de los subgrupos de emociones que se clasifican fue un poco arbitraria. Aquí, el criterio para la selección inicial fue considerar aquellos

TABLA 3.6: Mejores resultados para HMM de 2 estados: parametrizaciones con y sin energía.

	(MFCC)+delta+aceleración	(MFCC+E)+delta+aceleración
3 Emociones:	14Gaus. → 85.93 %	20Gaus. → 87.41 %
4 Emociones:	30Gaus. → 80.29 %	12Gaus. → 81.86 %
5 Emociones:	32Gaus. → 79.37 %	26Gaus. → 81.01 %
6 Emociones:	24Gaus. → 74.46 %	32Gaus. → 76.30 %
7 Emociones:	30Gaus. → 75.67 %	28Gaus. → 78.08 %

estados más distantes en el sentir de las personas. Evidentemente esta definición condiciona los resultados de alguna forma, y es de interés descubrir *por qué* y *cómo* los condiciona para poder aprovecharlos en la clasificación. En las matrices de confusión (Tablas 3.3 y 3.5) pueden notarse algunas similaridades entre las clases de emociones. Tanto ésto es así que, considerando otras agrupaciones y realizando el mismo test, se han obtenido por ejemplo un 99,23 % clasificando 3 emociones (Neutral, Anger y Sadness) y un 88,24 % clasificando 5 emociones (Neutral, Anger, Sadness, Fear y Disgust). De las discusiones aquí expuestas se vuelve menester la realización de un análisis más profundo de la información prosódico-acústica de las señales.

3.4. ANÁLISIS PROSÓDICO-ACÚSTICO

Según lo expuesto en secciones previas, se puede ver que muchos trabajos en el área utilizan las características prosódicas y espectrales de las señales de voz para alimentar reconocedores basados en redes neuronales, mezclas de Gaussianas y otros clasificadores estándar. Generalmente, no existe una interpretación de los tipos de errores presentes en los resultados desde el punto de vista acústico [76, 77, 79].

Por otra parte, la conceptualización psicológica de los afectos, por medio de modelos bi-dimensionales o tri-dimensionales, es ampliamente conocida en la caracterización de emociones [22, 23, 24]. Usualmente estos modelos son utilizados para agrupar emociones con el fin de definir grupos o clases emocionales, tales como los relacionados con una baja excitación y un bajo placer o aquellos asociados con una alta excitación y un alto placer, entre otros [22, 24, 70]. A partir de estos grupos emocionales se confeccionan los clasificadores. Por ejemplo, en [70], se presenta un clasificador diádico de múltiples niveles basado en estos modelos de emociones bi-dimensionales. En esta sección, se realiza un análisis de las características prosódicas y espectrales orientado a la caracterización de emociones y a la definición de grupos [64]. Estas agrupaciones están basadas exclusivamente en la información acústica, dejando de lado las consideraciones

de nivel psicológico o la taxonomía tradicional de las emociones humanas.

Para comenzar el análisis de las emociones se definió, para cada elocución, un vector con las medias del logaritmo del espectro, calculado por tramos, para cada banda de frecuencias

$$MLS(k) = \frac{1}{N_i} \sum_{n=1}^{N_i} \log |v_i(n,k)|, \tag{3.1}$$

donde k es la banda de frecuencias, N_i es el número de tramos en la elocución i y $v_i(n,k)$ es la transformada de Fourier discreta de la señal i para el tramo n. Luego, se calcula el promedio de los MLS (AMLS) sobre todas las frases que expresan la misma emoción

$$S_\ell(k) = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} MLS(k) = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \frac{1}{N_i} \sum_{n=1}^{N_i} \log |v_{i\ell}(n,k)|, \tag{3.2}$$

siendo N_ℓ es el número de elocuciones que pertenecen al mismo tipo de emoción ℓ.

Como fue mencionado, el uso de características prosódicas en el reconocimiento de emociones ha sido ampliamente discutido [29, 31, 35]. Generalmente, en estos trabajos se utilizan los métodos clásicos para calcular la *Energía* y la F_0 a lo largo de las señales [8], tal y como se comentó en la Sección 1.2. Se pueden calcular muchos parámetros y medidas estadísticas para estas características prosódicas; en general se utilizan el mínimo, la media, el máximo y la desviación estándar sobre las frases completas. Este conjunto de parámetros ha sido bien estudiado y algunos trabajos reportan una importante ganancia de información para discriminar emociones [30, 75, 85]. Aquí se ha utilizado este conjunto de características prosódicas, aunque no se descarta la idea de utilizarlo en el futuro junto a otros parámetros como la perturbación de frecuencia (en inglés, *Jitter*), la perturbación de amplitud (en inglés, *Shimmer*) y estadísticos de más alto orden.

A partir de las características prosódicas y acústicas relevadas aquí, se pretende descubrir similitudes entre las distintas emociones en algún tipo de estructura subyacente de los datos. El objetivo principal es agrupar a las emociones de forma no supervisada, utilizando las características más relevantes de los datos de entrada. Entre los métodos más interesantes para llevar adelante esta tarea se encuentra la técnica de mapas auto-organizativos.

3.4.1. Agrupamiento mediante mapas auto-organizativos

Como se presentó en la Sección 1.1.2, un SOM es una red neuronal artificial que se entrena de forma no-supervisada y tiene la capacidad de preservar las propiedades topológicas del espacio de entradas.

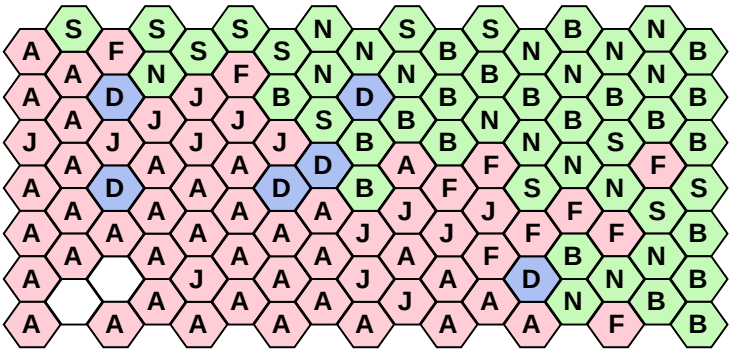

FIGURA 3.4: Agrupamiento de emociones utilizando un SOM (30 coeficientes MLS). Referencias: (A)nger; (B)oredom; (D)isgust; (F)ear; (J)oy; (S)adness; (N)eutral.

Para cada frase, se utilizaron los MLS como patrones de entrada al SOM. Como los archivos de audio tienen una frecuencia de muestreo de 16 kHz y las bandas frecuenciales consideradas en el análisis tienen una extensión de 40 Hz, los vectores MLS tienen 200 coeficientes. Con la idea de eliminar los coeficientes menos significativos, se realizaron pruebas simples con SOM para diferentes números de coeficientes MLS. Para éstos se fueron descartando los coeficientes de las bandas más altas, y entonces se realizaron entrenamientos con 200 $(0 - 8000\text{Hz})$, 50 $(0 - 2000\text{Hz})$, 40 $(0 - 1600\text{Hz})$, 30 $(0 - 1200\text{Hz})$ y 20 $(0 - 800\text{Hz})$ coeficientes. Los SOM entrenados se utilizaron para clasificar los patrones y se computó la tasa de aciertos. En estos resultados se pudo observar que, si se utilizan los primeros 30 coeficientes se obtiene el mejor compromiso entre la reducción dimensional de los vectores y la capacidad de discriminación entre clases de emociones.

Con el fin de explorar las similitudes entre las distintas emociones, se utilizó el SOM como un clasificador (luego de la etapa de entrenamiento) y cada celda fue etiquetada con la emoción que aparece más frecuentemente en esa posición. La proyección de datos obtenida muestra que ciertas clases de emociones podrían ser consideradas como grupos cuando usamos las características espectrales (Fig. 3.4). Los datos correspondientes a *Joy* y *Anger* están dispuestos desde la esquina inferior izquierda hacia el centro del mapa; mientras que *Boredom*, *Neutral* y *Sadness* aparecen propagados desde la esquina superior derecha hacia el centro. Por otra parte, los patrones correspondientes a *Fear* y *Disgust* están dispuestos de una forma más distribuida. Esta información visual provista por el mapa SOM puede ser considerada para agrupar emociones de diferentes formas. Entonces, por ejemplo, un grupo podría contener las emociones *Joy*, *Anger* y *Fear* (JAF) mientras que otro contendría las emociones *Boredom*, *Neutral* y *Sadness* (BNS) y finalmente, la emoción *Disgust* establecería sola un tercer grupo. Asimismo, un esquema de dos grupos podría ser propuesto, donde *Disgust* se incorpora

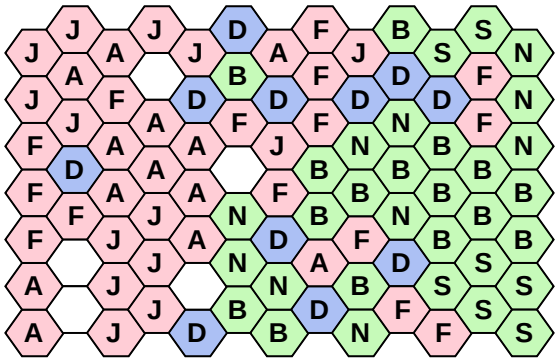

Figura 3.5: Agrupamiento de emociones utilizando un SOM (12 MFCCs promedio + 8 coeficientes prosódicos). Referencias: (A)nger; (B)oredom; (D)isgust; (F)ear; (J)oy; (S)adness; (N)eutral.

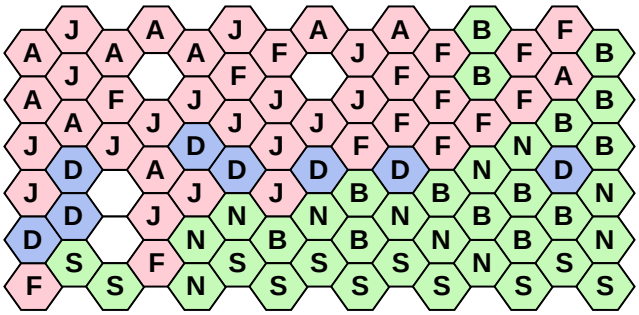

Figura 3.6: Agrupamiento de emociones utilizando un SOM (30 coeficientes MLS + 8 coeficientes prosódicos). Referencias: (A)nger; (B)oredom; (D)isgust; (F)ear; (J)oy; (S)adness; (N)eutral.

a un nuevo grupo JAFD (formado por JAF + Disgust) y BNS es el otro grupo que podría considerarse.

De la misma manera, se realizaron experimentos para un conjunto de vectores con información combinada de 30 MLS, 12 MFCCs promedio y 8 coeficientes con información prosódica (mínimos, máximos, medias y desvio estándar de la energía y la F_0). Los resultados obtenidos con estos vectores de características evidencian topologías de agrupamiento similares. Por ejemplo, en la Figura 3.5 se puede ver el agrupamiento realizado con un SOM cuyos vectores de entrada están formados por 12 coeficientes MFCC (promediados sobre todas las ventanas) más 8 características prosódicas. En la Figura 3.6 puede verse el agrupamiento obtenido para vectores que incluyen 30 coeficientes MLS y 8 prosódicos.

En esta primer etapa se han buscado similitudes entre las distintas clases de emo-

ciones. Este análisis preliminar es meramente exploratorio y su finalidad no es la clasificación, por lo tanto no se reportan resultados numéricos. Se puede ver que las características elegidas permiten realizar agrupamientos de las clases de emociones de forma no supervisada. En la siguiente sección se realiza un análisis más detallado de esta información.

3.4.2. Análisis espectral

Con el fin de validar las propuestas de agrupamiento presentadas previamente con SOM, se presenta en forma gráfica la información obtenida con AMLS. En la Figura 3.7 se pueden apreciar algunas características principales, similitudes y desigualdades entre las gráficas de las distintas clases de emociones.

Es notorio que, en las altas frecuencias, las curvas son similares y aparentemente no brindan información importante para la discriminación entre clases. La información más importante, para discriminar las distintas clases entre sí, se puede localizar entre 0 y 1200 Hz. Aquí, se ratifica de alguna manera que el rango de frecuencias para los AMLS, que presentan la información más importante, coincide con los coeficientes elegidos como óptimos en el análisis del SOM.

La Figura 3.8 muestra la información de AMLS para cada clase de emoción, las clases están agrupadas por la similaridad de sus curvas. Por ejemplo, se puede notar que *Joy*, *Anger* y *Fear* presentan una morfología similar y un máximo entre 240 y 280 Hz. Alrededor de 75 Hz existe un mínimo en las gráficas de *Joy*, *Anger*, *Fear* y *Disgust*. Por otro lado, *Boredom*, *Neutral* y *Sadness* tienen una forma similar y un pico entre 115 y 160 Hz. En las figuras es posible advertir que las emociones que se agrupan en el SOM son espectralmente similares en el análisis de AMLS. Las emociones que son acústicamente más similares coinciden con aquellas que son más díficiles de discriminar, tal como puede verse en las matrices de confusión presentadas previamente y en trabajos anteriores [28, 34, 30, 36].

En base a lo expuesto es posible aseverar que este análisis permite definir grupos utilizando la información prosódica y espectral, independientemente de las consideraciones psicológicas. Aquí, el enfoque heurístico presentado con AMLS, es usado para validar la información encontrada previamente con SOM. Esta información, relevante para el agrupamiento de emociones, será utilizada en la Sección 3.5 para definir un clasificador jerárquico.

A continuación se presentan algunos aspectos importantes directamente relacionados con el análisis espectral.

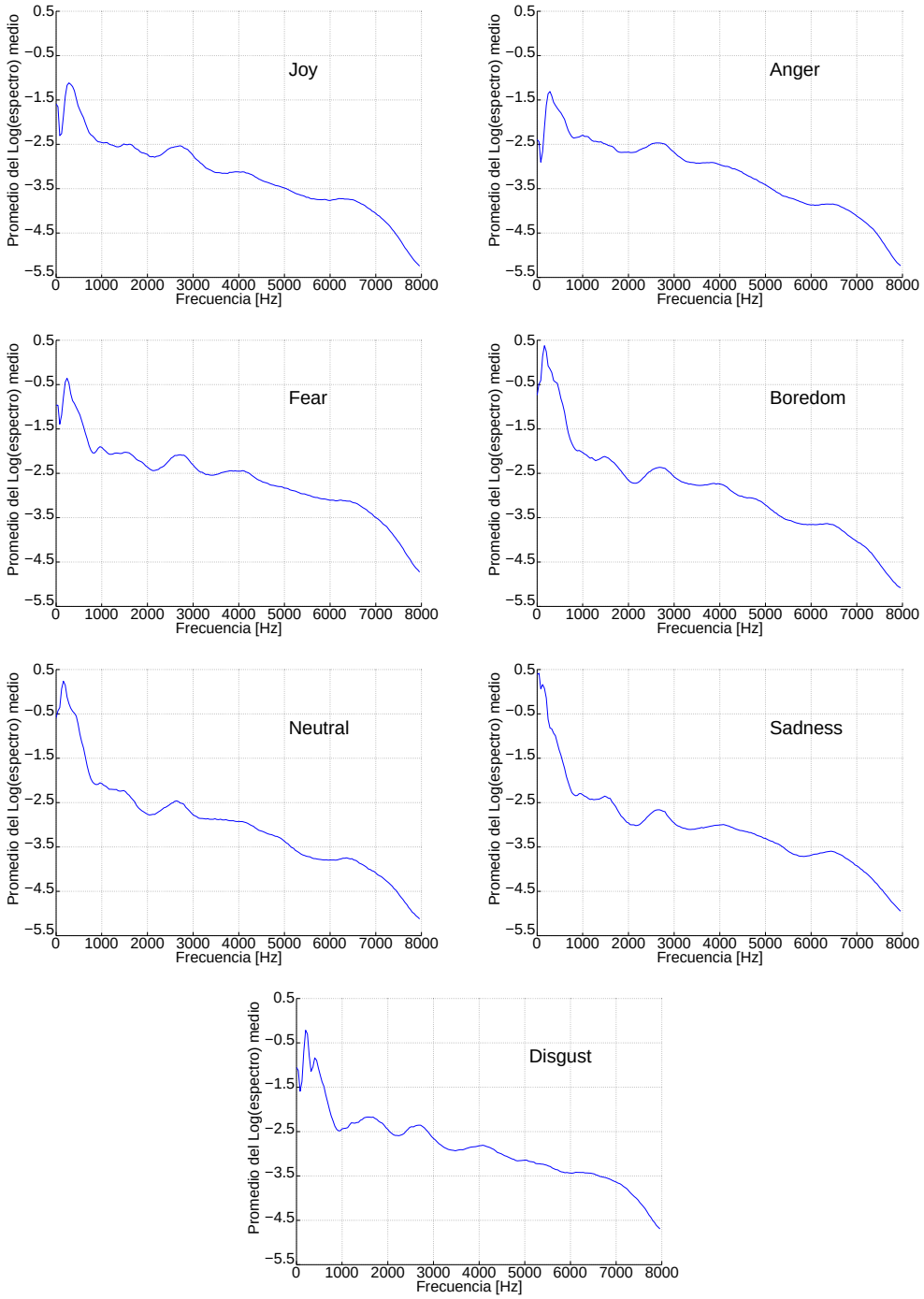

Figura 3.7: Promedio del log(espectro) medio (AMLS) para cada tipo de emoción.

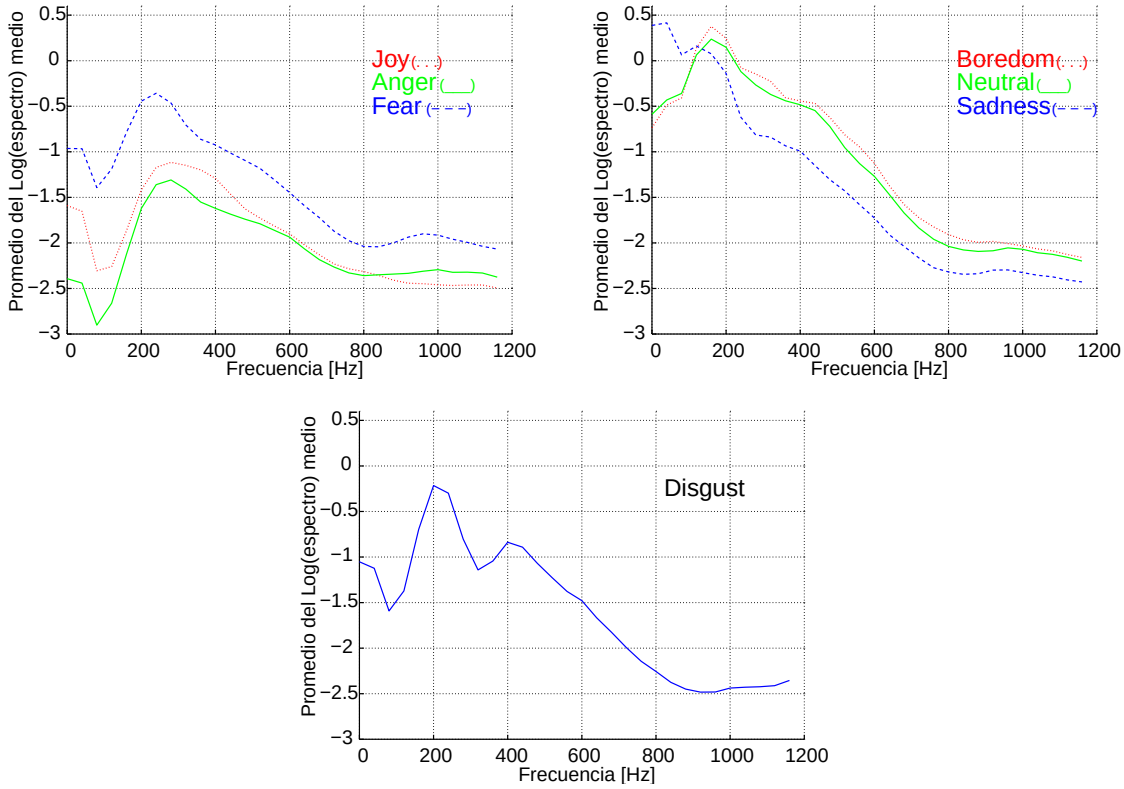

FIGURA 3.8: AMLS de todas las clases de emociones, agrupadas por sus similaridades espectrales.

FILTRADO DE PRE-ÉNFASIS

Como se puede observar en las gráficas de los AMLS y a partir de las discusiones anteriores, se ha considerado que la información relevante para discriminar las emociones se encuentra aproximadamente hasta los 1200 Hz. Sin embargo, en el proceso de generación de estas señales, las altas frecuencias son atenuadas y ésta podría ser la causa por la cual allí no se observa información útil para distinguir emociones. Para poder analizar más detalladamente esta hipótesis se ha realizado un filtrado de pre-énfasis estándar para realzar las altas frecuencias en estas señales. A modo de ejemplo, se presenta en la Figura 3.9 el oscilograma y espectrograma de la frase "*Der Lappen liegt auf dem Eisschrank.*" pronunciada por un hombre adulto. En la Figura 3.10 puede observarse el resultado de aplicar pre-énfasis de altas frecuencias a la misma frase.

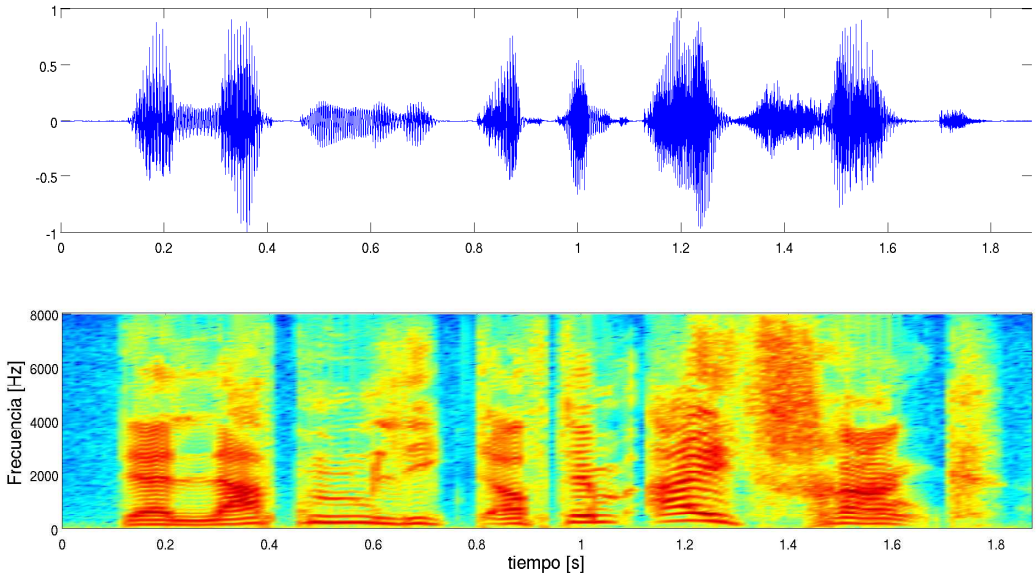

Figura 3.9: Oscilograma y espectrograma de la frase 03a01Wa del corpus.

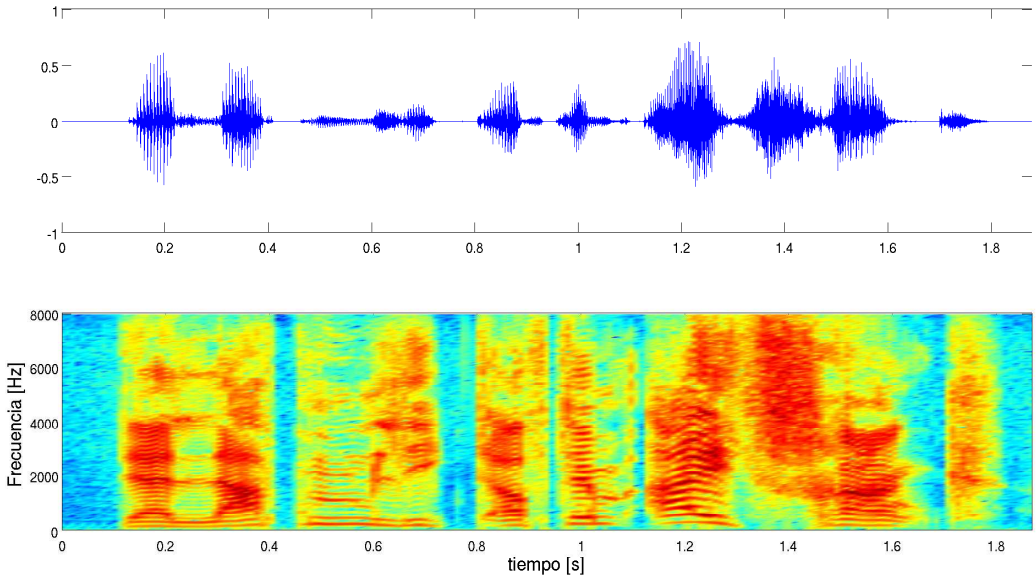

Figura 3.10: Oscilograma y espectrograma de la frase 03a01Wa del corpus, con preénfasis.

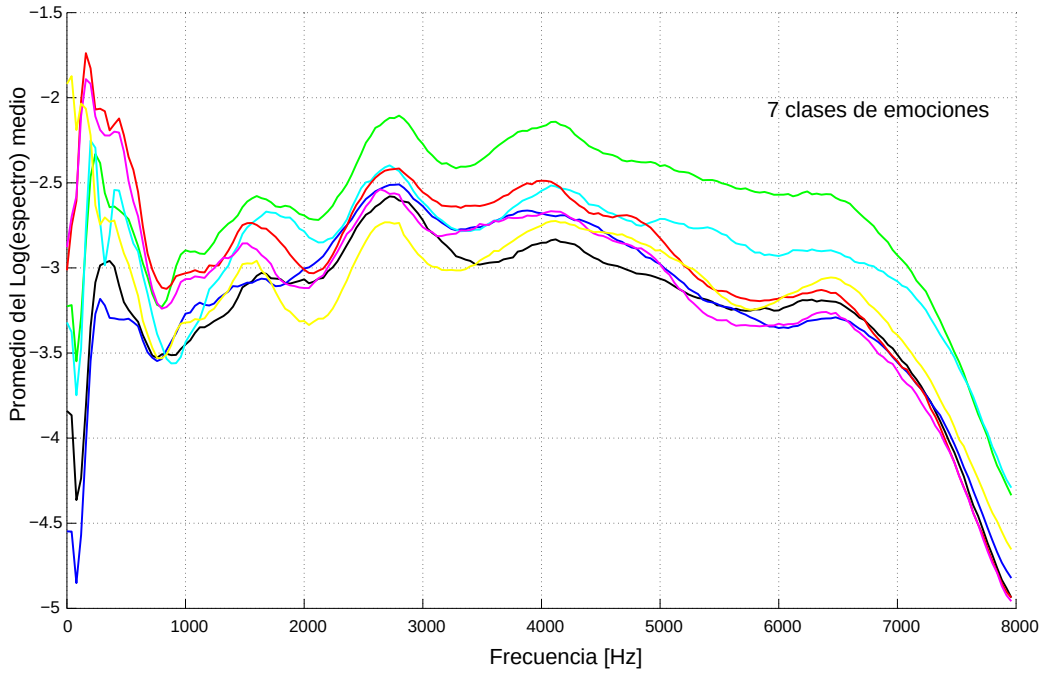

FIGURA 3.11: AMLS por cada clase de emoción, con pre-énfasis de altas frecuencias. Referencias: Joy, Anger, Fear, Disgust, Boredom, Neutral, Sadness.

En la Figura 3.11 se presentan las gráficas de los AMLS calculados sobre las señales realzadas, para cada tipo de emoción. Como en el caso presentado en la sección previa, no es posible visualizar información relevante para diferenciar las distintas clases de emociones más allá de los 1200 Hz, a pesar de haber realizado el pre-énfasis en las altas frecuencias.

ASPECTOS RELATIVOS AL GÉNERO

A priori es posible hablar de la existencia de cierta variabilidad dependiente del género en las señales de voz emocional. Algunos enfoques han considerado ésto en una etapa previa donde se realiza una clasificación de género. Sin embargo, en nuestra propuesta ésto no se contempla (como se verá luego en la Sección 3.5). Este trabajo, análisis acústico y desarrollo de un reconocedor, tiene como objetivo encontrar un clasificador que sea capaz de manejar las diferencias de género de forma implícita, es decir, no se incluyen bloques específicos para la discriminación de género.

A pesar de que se ha propuesto continuar con un análisis detallado en investigaciones futuras, parece apropiado incluir aquí algunas gráficas que permitan analizar visualmente la morfología espectral de las señales. En la Figura 3.12 se puede ver el análisis MLS realizado para 6 estados emocionales expresados en la frase **b09**: "*Ich will das eben wegbringen und dann mit Karl was trinken gehen*" (en español: Yo descartaría eso y entonces iría a tomar un trago con Karl.). El análisis se realizó para dos individuos, una mujer y un hombre cuyas denominaciones en la base de datos son 08 y 15 (hombre y mujer) respectivamente. En las gráficas puede verse que las envolventes presentan una morfología similar para cada una de las emociones analizadas. Si consideramos los primeros 30 coeficientes, que ya definimos anteriormente como los más representativos de las diferencias entre las emociones, se puede apreciar un pequeño desfasaje que posiblemente estaría relacionado de alguna forma a la F_0 de cada individuo (y dependiente de cada tipo de emoción). Para tener una idea cuantitativa de esto se calculó una correlación cruzada entre los hablantes para estos 30 coeficientes por cada tipo de emoción. Los valores máximos de la correlación cruzada entre 15 y 08 se obtuvieron cuando los coeficientes del hablante 08 se desplazaron una cantidad $0; -2; -2+2; -1; -3$ y 0 para *J; A; F; B; N* y *S* respectivamente. Los valores obtenidos muestran una tendencia de desplazamiento hacia las altas frecuencias en las curvas del hablante 15 (la mujer). Los análisis más detallados y sobre más ejemplos, inclusive considerando bandas frecuenciales más angostas en el cálculo de los MLS para obtener mejores detalles, están considerados para un trabajo futuro. Esta primer aproximación muy sencilla sobre las diferencias de géneros insinua la necesidad de un análisis y discriminación por género como parte del pre-procesamiento o como un módulo al inicio del clasificador [79]. Sin embargo, si la hipótesis planteada aquí acerca de las morfologías similares desfasadas es verificada, podría proponerse una función que tenga como fin alinear la información de las diferentes emociones a partir de la F_0 propia del hablante e independientemente de su género. Esta propuesta también podría ser considerada en trabajos futuros con el fin de lograr un pre-procesamiento más útil para reconocer emociones.

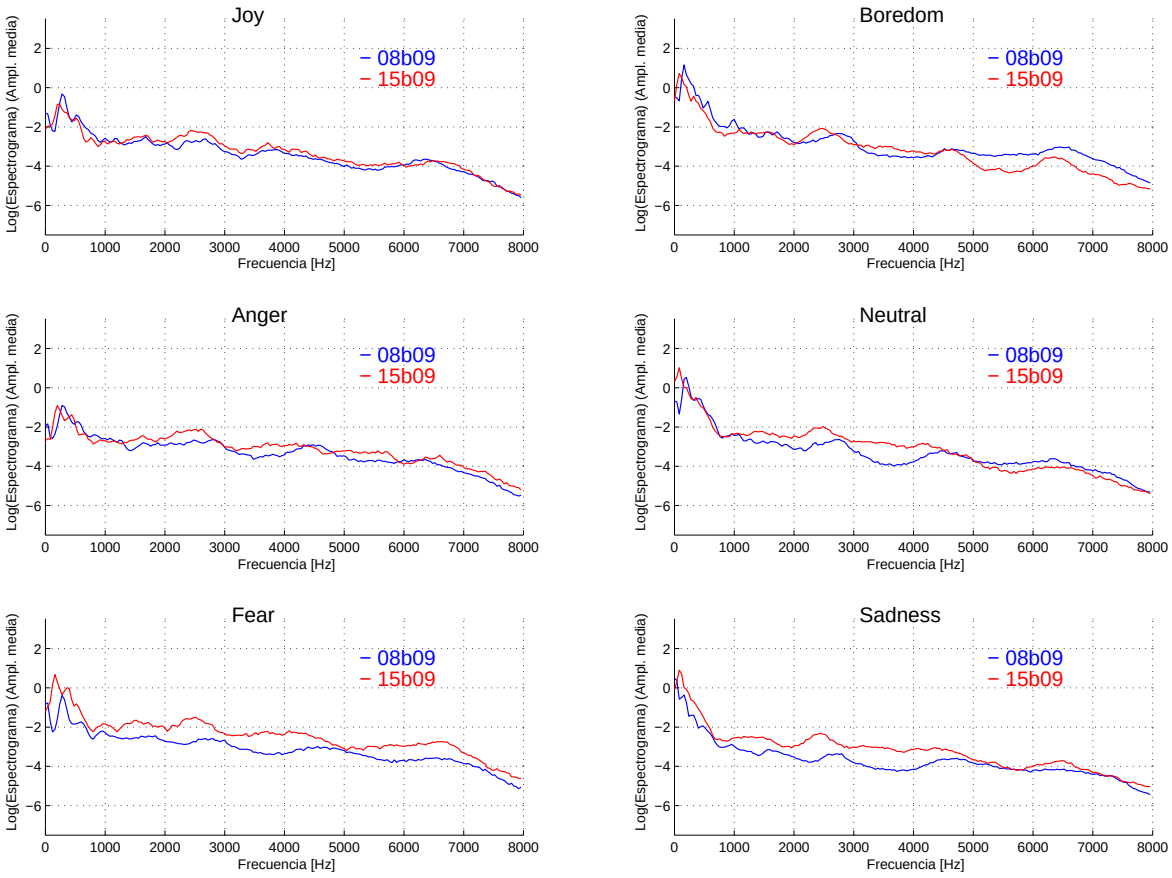

FIGURA 3.12: MLS de una frase expresada para distintas emociones, por una mujer y por un hombre.

3.5. CLASIFICADOR JERÁRQUICO DE MÚLTIPLES CARACTERÍSTICAS

En esta sección se presenta un novedoso método de clasificación jerárquico de múltiples características [86] basado en los análisis acústicos presentados en las secciones previas. La principal motivación para el desarrollo de un clasificador jerárquico es tomar ventaja de las similaridades prosódico-espectrales para mejorar la tasa de clasificación de emociones. También se hace uso del hecho de que, como se presentó anteriormente,

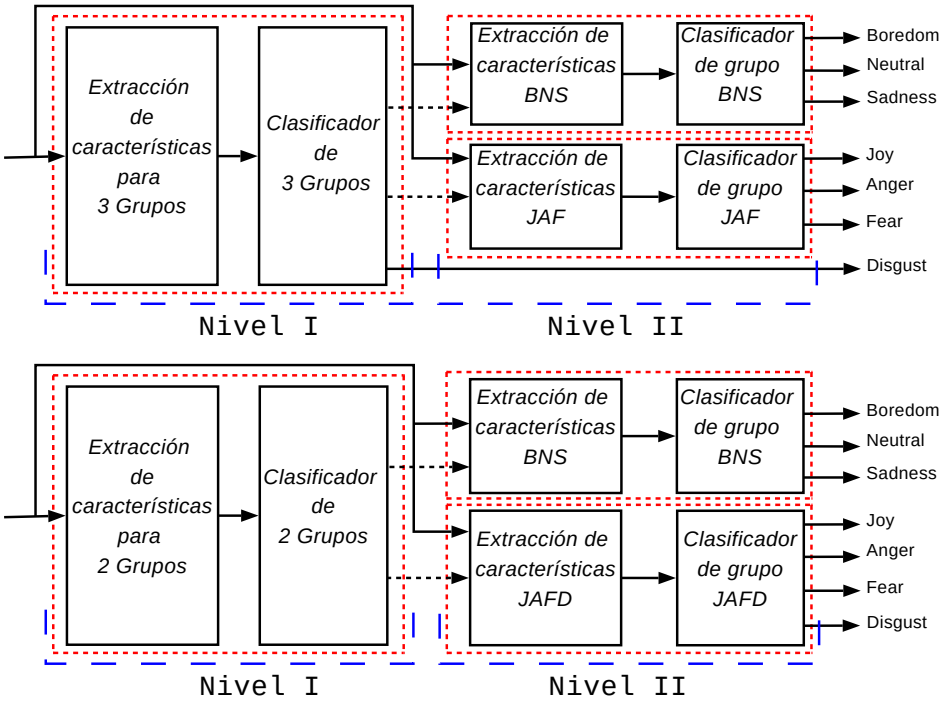

Figura 3.13: Esquema de dos posibles clasificadores jerárquicos de 2 niveles.

para un mismo clasificador estándar se obtienen mejores resultados cuando el número de emociones decrece. Además, se comprobó que las principales diferencias entre determinadas emociones son más evidentes cuando se utiliza un vector de características particular y que se obtiene una mejor clasificación cuando se utiliza un clasificador con una estructura peculiar adaptada a un grupo reducido de clases.

En esta sección también se presentan, como referencia, los resultados de los clasificadores estándar basados en GMM, HMM y MLP. Las estructuras generales de los modelos GMM y HMM son las que obtuvieron los mejores resultados en [34], mientras que el mejor MLP se encontró luego de evaluar diferentes configuraciones para cada uno de los distintos vectores de características de entrada. Los mejores modelos fueron tomados como base para una posterior comparación. Para implementar los sistemas con HMM y MLP, se utilizaron: el *Hidden Markov Toolkit* [17] y el *Stuttgart Neural Network Simulator* [87], respectivamente.

Como se puede ver en la Figura 3.13, se proponen dos estructuras de clasificadores jerárquicos en dos etapas o niveles. Cada etapa está formada por uno o dos bloques que contienen una sección de *extracción de características* y otra sección de *clasificación*. El clasificador que se ve en la parte superior de la Figura 3.13 tiene un Nivel I donde

la frase será clasificada en uno de los 3 grupos (BNS, JAF o Disgust). Luego ésta debe ser clasificada otra vez en el bloque del grupo que le corresponde (si no es clasificada como Disgust) y finalmente se obtiene la etiqueta de la emoción reconocida. El segundo modelo tiene un Nivel I para clasificar la frase en alguno de los 2 grupos (BNS o JAFD) y un Nivel II donde se obtiene la etiqueta de la emoción después de hacer la clasificación en el bloque correspondiente.

Para definir la estructura del modelo jerárquico en cada bloque se evaluaron diversas configuraciones de MLP y HMM con diferentes vectores de parámetros. Se estableció una fase de *diseño* donde, mediante una partición de datos especialmente extraída, se evalua la estructura del modelo jerárquico. La evaluación se realiza de forma independiente y aislada para cada bloque. Finalmente, los niveles del modelo son elegidos y ensamblados con aquellos clasificadores que obtuvieron los mejores resultados en las pruebas aisladas, realizando una validación final con otras particiones de datos diferentes.

3.5.1. SELECCIÓN DE CARACTERÍSTICAS

De la Tabla 3.1 está claro que la distribución de las emociones está desbalanceada, donde *Anger* es mayoritaria en el conjunto de datos (24 % del total). Como se ha mencionado, esta característica en el conjunto de datos puede generar un sesgo en los métodos de validación. Es interesante notar que en los trabajos previos practicamente no se aborda este tema, y se producen entonces resultados sesgados y no comparables. Además, es bien sabido que los conjuntos de datos de entrenamiento desbalanceados conducen a resultados inapropiados en clasificadores como el MLP. Para evitar estos problemas, el conjunto de datos fue balanceado igualando el tamaño de las clases. Ésto se llevo a cabo seleccionando de forma aleatoria el mismo número de muestras para todas las clases en cada partición (46 x 7 = 322 elocuciones). Como se comentó anteriormente, las transcripciones de las elocuciones no son consideradas y cada frase tiene una etiqueta que refiere a la emoción que expresa. Luego, cada elocución es un patrón de entrenamiento o prueba de acuerdo con la situación en que es usada.

Para cada elocución, se extrajeron 3 tipos de características: MLS, MFCCs y parámetros prosódicos. Los MLS se calcularon de la misma forma en que se definieron en la Sección 3.4. Los espectrogramas y parametrización MFCC fueron realizados utilizando una ventana de Hamming de 25 ms con un paso de 10 ms. También se extrajeron de las señales, los primeros 12 coeficientes MFCC más sus primeras y segundas derivadas [17]. Los coeficientes MFCC y Log-Spectrum fueron calculados por tramos, y sus medias fueron calculadas sobre todos los tramos. Como fue mencionado, la utilización de características prosódicas en el reconocimiento de emociones ha sido extensamente discutida. Aquí se utilizaron la *energía* y la F_0 para todas las elocuciones, y sobre éstas

Parámetros	FV12	FV14	FV16	FV18	FV20	FV30	FV32	FV34	FV36	FV38	FV42	FV44	FV46	FV48	FV50
12 MFCC medios	•	•	•	•	•						•	•	•	•	•
30 MLS						•	•	•	•	•	•	•	•	•	•
$\mu(F_0)$, $\mu(E)$		•	•	•	•		•	•	•	•		•	•	•	•
$\sigma(F_0)$, $\sigma(E)$			•		•			•		•			•		•
$Min(F_0)$, $Max(F_0)$				•	•				•	•				•	•
$Min(E)$, $Max(E)$				•	•				•	•				•	•

Tabla 3.7: Vectores de características para MLP (estáticos).

se calcularon los valores mínimos, máximos, promedio y desviación estándar de éstos. En algunos trabajos se han utilizado estos valores y han reportado una ganancia de información importante para discriminar emociones [30, 75, 85].

Distintas combinaciones de estas características (MLS, la media de cada MFCC y la información prosódica) fueron dispuestas en vectores. La Tabla 3.7 muestra el número de características para cada vector y los tipos de características que incluyen. Por ejemplo, el vector de características *FV14* incluye 12 MFCC promedio, la media de F_0 y la media de la energía. Para las pruebas con MLP, cada dimensión del vector fue normalizada de forma independiente con el máximo valor encontrado en el conjunto de vectores, para esa dimensión.

Otro factor que puede sesgar la estimación de la tasa de reconocimiento ya mencionado es utilizar sólo una partición de entrenamiento y una de validación. Nuevamente se utilizó una validación cruzada con el método dejar-k-afuera [84]. Después de generar la *partición de diseño*, se generaron 10 particiones de datos para la etapa de prueba. En los experimentos con MLP, el 60 % de los datos de cada partición se seleccionó de forma aleatoria para entrenamiento, un 20 % fue usado para el monitoreo de generalización y el restante 20 % se dejó para la etapa de validación. Cada entrenamiento de los MLP fue detenido cuando la red alcanzó el pico de generalización con los patrones de monitoreo [3]. En el caso de los HMM, el 20 % definido para monitoreo en MLP fue adicionado al conjunto de entrenamiento.

3.5.2. Diseño del clasificador jerárquico

Se eligió una partición de diseño para poder definir la estructura de los bloques, cada uno fue evaluado de forma aislada con diversas configuraciones. Los 15 vectores de características con 3 configuraciones de capa oculta diferentes (90, 120 y 150 per-

ceptrones) fueron evaluados en cada prueba realizada a un bloque con MLP. Por otro lado, se utilizaron los vectores de 36 coeficientes (12 MFCCs más delta y aceleración) para las pruebas de bloques con HMM, de la misma forma que en [34]. Para el caso de los HMM se debe entrenar un modelo de k estados por cada tipo de emoción o grupo de emociones, esto implica que se deben adaptar los parámetros de transición de estados, las probabilidades de observación de cada estado y las distribuciones de las observaciones. Para entrenar un modelo HMM para una emoción particular se cuenta con los parámetros calculados en cada tramo (ventanas de 25 ms y paso de 10 ms) de todas las frases de entrenamiento. Por ejemplo, para una señal de 1 segundo se tendrían 98 frames y en cada uno se obtienen las 39 características. Mientras que los parámetros a modelar de un modelo de 3 estados para una emoción son: la media y varianza de las 39 gaussianas por cada estado, los 4 coeficientes de transición entre estados y 9 coeficientes de probabilidades de observación. Por el lado de los MLP, se deben entrenar los pesos que asocian los vectores con características de entradas y las neuronas de la capa oculta y los que asocian estas neuronas y las de la capa de salida. Por ejemplo para una red con la configuración *12+90+3* se deberían entrenar los pesos que unen las 12 entradas con las 90 neuronas de la capa oculta y los pesos que enlazan las 90 neuronas con las 3 salidas. Aquí por cada señal completa se calcula único un conjunto de parámetros y éstos son presentados tantas veces como sean necesarias hasta que la red esté entrenada y se alcance el punto de generalización. En la Figura 3.13 es posible identificar 5 bloques diferentes que deben ser evaluados (2 en el Nivel I y 3 en el Nivel II). Se realizaron todos los experimentos posibles con MLP y HMM en cada bloque.

DEFINICIÓN DEL NIVEL I

Para esta etapa se considera el primer Nivel de forma aislada. Para el Nivel I del clasificador jerárquico se evaluaron 6 diferentes opciones (3 por cada posible agrupamiento):

a) reagrupar las salidas de un clasificador HMM estándar en 3 grupos (HMM^7g^3);

b) modelar cada uno de los 3 grupos con un HMM (HMM^3);

c) utilizar un MLP con 3 neuronas en la capa de salida (MLP^3);

d) reagrupar las salidas de un clasificador HMM estándar en 2 grupos (HMM^7g^2);

e) modelar cada uno de los 2 grupos con un HMM (HMM^2);

f) utilizar un MLP con 2 neuronas en la capa de salida (MLP^2).

Tabla 3.8: Desempeño del MLP para 3 grupos (*etapa de diseño*).

Entrada	Mejor red	Entrenamiento [%]	Validación [%]
FV12	12+90+3	98.98	85.71
FV14	14+90+3	95.92	87.30
FV16	16+90+3	97.96	87.30
FV18	18+150+3	98.47	79.37
FV20	20+90+3	100.00	77.78
FV30	30+90+3	100.00	87.30
FV32	32+90+3	99.49	85.71
FV34	34+120+3	**98.98**	**88.89**
FV36	36+90+3	99.49	84.13
FV38	38+120+3	100.00	82.54
FV42	42+120+3	92.86	87.30
FV44	44+150+3	96.94	84.13
FV46	46+150+3	94.39	85.71
FV48	48+90+3	100.00	80.95
FV50	50+150+3	100.00	82.54

Las opciones a) y d) fueron computadas a partir de modelos que implementan las configuraciones que arrojaron mejores resultados en [34]. A partir de las mejores configuraciones para los HMM, se evaluó la cantidad óptima de Gaussianas en las mezclas para encontrar el mejor modelo para las opciones b) y e).

Los mejores modelos resultaron tener 30 Gaussianas para 3 grupos y 8 Gaussianas para 2 grupos. Las Tablas 3.8 y 3.9 muestran los mejores resultados de MLP para cada FV sobre los datos de entrenamiento y validación, para 3 y 2 grupos respectivamente. Los mejores resultados obtenidos para el Nivel I se resumen en las Tablas 3.10 y 3.11. En la Tabla 3.10 se puede ver que con MLP se obtuvo el mejor resultado para 3 grupos, sin embargo, éste es el que presenta la peor clasificación para Disgust. Ésto puede deberse a que el MLP no es un buen clasificador cuando las clases están notablemente desbalanceadas, como lo son en este caso JAF, BNS y Disgust. Por otra parte, en la Tabla 3.11 se puede ver como el MLP alcanzó una tasa de clasificación del 100 % para 2 grupos casi balanceados (JAFD y BNS). Ambas configuraciones de HMM obtuvieron buenos resultados en el Nivel I.

TABLA 3.9: Desempeño del MLP para 2 grupos (*etapa de diseño*).

Entrada	Mejor red	Entrenamiento [%]	Validación [%]
FV12	12+90+3	98.47	98.41
FV14	14+90+3	92.35	98.41
FV16	16+90+3	93.88	98.41
FV18	18+90+3	95.41	92.06
FV20	20+90+3	93.37	92.06
FV30	30+150+3	100.00	95.24
FV32	32+90+3	100.00	95.24
FV34	34+90+3	97.45	95.24
FV36	36+120+3	93.37	90.48
FV38	38+90+3	98.98	93.65
FV42	42+120+3	100.00	98.41
FV44	44+90+3	100.00	96.83
FV46	46+90+3	**98.98**	**100.00**
FV48	48+120+3	96.43	98.41
FV50	50+150+3	100.00	96.83

TABLA 3.10: Desempeño de los modelos de clasificación para 3 grupos (*etapa de diseño*).

	HMM^7g^3 [%]	HMM^3 [%]	MLP^3 [%]
JAF	88.89	77.78	88.89
BNS	85.19	92.59	100.00
D	66.67	88.89	55.56
promedio	84.13	85.71	**88.89**

DEFINICIÓN DEL NIVEL II

En esta etapa se extraen los 2 bloques posibles para el Nivel II, dependientes del agrupamiento elegido en el Nivel I. Para cada bloque posible en el Nivel II se realizaron, de forma aislada, las evaluaciones con HMM y MLP utilizando la *partición de diseño*. Las pruebas con HMM se realizaron con el fin de encontrar la cantidad adecuada de Gaussianas en las mezclas. Los mejores resultados obtenidos para HMM fueron de un 74,07 % para clasificar individualmente las emociones incluidas en JAF con 26 Gaussianas en las mezclas, un 77,78 % para la clasificación dentro del grupo JAFD con 20 Gaussianas en las mezclas, mientras que sólo 4 Gaussianas alcanzaron un 77,78 % para el caso de BNS. Para el caso de MLP se realizaron experimentos con cada uno de los FV y para 3 diferentes configuraciones de la estructura de la red. Los mejores resultados para los bloques aislados del Nivel II se pueden apreciar en la Tabla 3.12.

Tabla 3.11: Desempeño de los modelos de clasificación para 2 grupos (*etapa de diseño*).

	HMM^7g^2 [%]	HMM^2 [%]	MLP^2 [%]
JAFD	94.44	88.89	**100.00**
BNS	85.19	96.30	**100.00**
promedio	90.48	92.06	**100.00**

Tabla 3.12: Mejores desempeños en la clasificación aislada para el Nivel II (*partición de diseño*).

Grupo	modelo del Nivel II	Rendimiento [%]
JAF	MLP (46+90+3)	**85.19**
	HMM (26 Gauss.)	74.07
JAFD	MLP (12+90+4)	66.67
	HMM (20 Gauss.)	**77.78**
BNS	MLP (44+150+3)	**81.48**
	HMM (4 Gauss.)	77.78

Estos resultados revelan la importancia de hacer frente a cada problema particular utilizando FV específicos y ésto se relaciona directamente con las mejoras en los desempeños. Una interpretación similar justifica la elección particular del clasificador y de su estructura para cada bloque. En la sección siguiente se utilizan las mejores configuraciones encontradas, para cada bloque, para construir el sistema completo y poder realizar su validación.

3.5.3. Evaluación y resultados

Para validar el sistema jerárquico de múltiples características fueron generadas 10 particiones de datos (aquí no se utiliza la partición de diseño). El modelo se ensambló con las características de entrada y las estructuras de los bloques de aquellas configuraciones que lograron los mejores resultados en la *fase de diseño*, de forma aislada. En la Tabla 3.13 se puede observar un resumen de los mejores resultados obtenidos para 3 y 2 grupos en el Nivel I. Estos resultados fueron obtenidos realizando validación cruzada con las 10 particiones.

El desempeño del sistema completo, utilizando 3 grupos en el Nivel I, es presentado en la columna "Mejor" de la Tabla 3.14. En la segunda columna, se muestra el resultado de Disgust para cada modelo en el Nivel I. Los patrones clasificados como JAF en el Nivel I son evaluados con ambos modelos en el Nivel II, y allí se pueden comparar sus resultados al identificar las emociones específicas de forma individual. Los promedios

TABLA 3.13: Mejores resultados del clasificador jerárquico en el Nivel I.

Nivel I	Modelo	Desempeño [%]
3 grupos	HMM^7g^3 (30 Gaussianas)	**89.84**
	HMM^3 (30 Gaussianas)	86.82
	MLP (34 + 120 + 3)	82.06
2 grupos	HMM^7g^2 (30 Gaussianas)	92.86
	HMM^2 (8 Gaussianas)	90.16
	MLP (46 + 90 + 2)	**93.02**

TABLA 3.14: Resultados finales para el modelo jerárquico de 3 grupos en el Nivel I.

Nivel I		Nivel II				
Modelo	Disgust [%]	JAF [%]		BNS [%]		Mejor [%]
		HMM	MLP	HMM	MLP	
HMM^7g^3	**80.00**	63.70	**71.48**	**69.26**	62.96	**71.75**
HMM^3	68.89	59.63	**67.78**	**71.48**	62.22	**69.52**
MLP	57.78	56.30	**64.07**	**68.89**	62.22	65.24

de reconocimiento para estas 3 emociones son mostrados en las columnas 3 y 4 de la tabla, para cada tipo de clasificador en el Nivel II. En las columnas 5 y 6 puede verse la misma información para el grupo BNS. En la columna *Mejor* se muestra la tasa de reconocimiento computada a partir de la combinación de los mejores modelos para los grupos JAF y BNS. La tasa de clasificación final se ha calculado de forma proporcional a la cantidad de patrones de validación presentes en cada grupo de emociones ($R = (R_D+3R_{JAF}+3R_{BNS})/7$). Como se ha mencionado previamente, el número de patrones de validación está balanceado (por emoción, no por grupo). En la tabla es posible apreciar que siempre los MLPs son mejores que los HMMs para los bloques de JAF, mientras que los HMMs han obtenido mejores resultados en los bloques de BNS.

En la Tabla 3.15 se muestra el rendimiento para los bloques JAFD y BNS con ambos modelos, para cada modelo implementado en el Nivel I. Los resultados obtenidos para las mejores combinaciones considerando cada modelo de clasificación de 2 grupos en el Nivel I son: 66,99 % para HMMs reagrupados (HMM^7g^2), 64,44 % para el modelo de 2 HMMs (HMM^2) y 66,03 % para MLP. Considerando 2 grupos en el Nivel I, el mejor modelo jerárquico de múltiples características está formado por un modelo de HMM reagrupados (HMM^7g^2) con 30 Gaussianas en las mezclas en el Nivel I; un HMM con 20 Gaussianas en las mezclas para el bloque JAFD y un HMM con 4 Gaussianas en las mezclas para el bloque BNS. En este modelo, nuevamente los HMMs han obtenido el mejor rendimiento en el bloque BNS, mientras que los HMMs han resultado mejores

Tabla 3.15: Resultados finales para el modelo jerárquico de 2 grupos en el Nivel I.

Nivel I	Nivel II				
Modelo	JAFD [%]		BNS [%]		Mejor [%]
	HMM	MLP	HMM	MLP	
HMM^7g^2	**65.28**	58.61	**69.26**	62.96	66.99
HMM^2	**57.78**	55.56	**73.33**	64.82	64.44
MLP	**63.33**	60.00	**69.63**	63.33	66.03

que MLP también para el bloque JAFD.

Análisis comparativo de resultados y discusión

Aquí se presentan los resultados de evaluar el clasificador propuesto, basado en la información prosódica y espectral, en las mismas condiciones experimentales que los clasificadores estándar con características fijas para 7 emociones.

En [34] se ha presentado un análisis comparativo entre GMM y HMM en la tarea de reconocer 7 emociones. Allí los mejores resultados para 7 emociones fueron obtenidos utilizando un modelo HMM de 2 estados con mezclas de 30 Gaussianas y un modelo GMM con 32 Gaussianas. En ambos casos la parametrización abarca 12 coeficientes MFCC junto a las primeras y segundas derivadas. Aquí se exponen los resultados de evaluar estos sistemas con las mismas 10 particiones con que se evalua el modelo jerárquico. También se ha utilizado validación cruzada para obtener resultados válidos para hacer comparaciones. Como se puede observar en la Tabla 3.16, la tasa de clasificación fue 63,49 % con GMM y 68,57 % con HMM. Aquí también se busco el mejor clasificador basado en MLP para 7 emociones con la partición de diseño. Luego, el mejor resultado fue de 66,83 % con validación cruzada para la red tiene 90 neuronas en la capa oculta y utiliza *FV46* como vector de entrada.

Con respecto a los clasificadores jerárquicos, y para el esquema que contempla 2 grupos en el Nivel I no se ha encontrado una configuración que mejore el desempeño del mejor de los clasificadores estándar. Mientras que, para el modelo jerárquico que considera 3 grupos en el Nivel I se obtiene el mejor resultado y su configuración es la que se presenta en la Figura 3.14. Éste está formado por un modelo de HMMs reagrupados (HMM^7g^3) con 30 Gaussianas en las mezclas para el Nivel I; un MLP con un vector de entrada *FV46* y 90 neuronas en la capa oculta para el bloque de JAF y HMMs con 4 Gaussianas en las mezclas para el bloque BNS.

La Tabla 3.16 contiene una comparación entre los clasificadores estándar y los clasificadores jerárquicos de múltiples características que se han propuesto en esta Tesis. Los resultados muestran que el método jerárquico mejora el rendimiento en 3,18 %

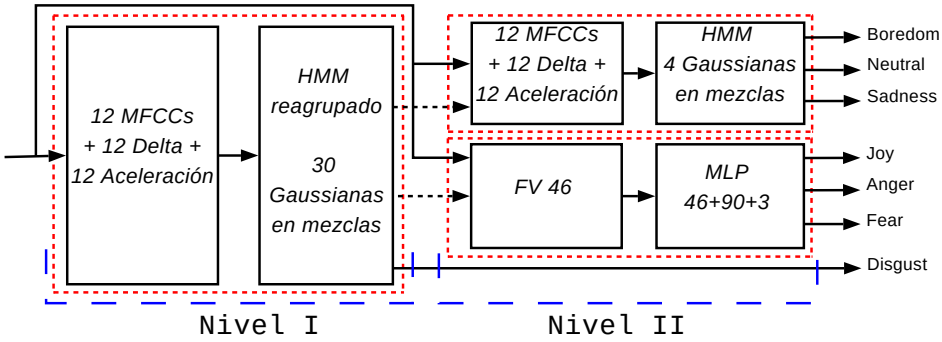

FIGURA 3.14: Mejor configuración obtenida para el modelo jerárquico.

TABLA 3.16: Tasas de reconocimiento para los distintos clasificadores.

Modelo	Rendimiento [%]
GMM	63.49
MLP	66.83
HMM	68.57
Jerárquico 2	66.99
Jerárquico 3	**71.75**

(absolutos) sobre el mejor clasificador estándar propuesto, con una validación cruzada de 10 particiones [86]. Las comparaciones directas con los resultados presentados en trabajos de otros autores no fueron posibles por diversos motivos entre los que se pueden mencionar la falta de detalles en las definiciones de los métodos y códigos de implementación, la utilización de diferentes bases de datos (naturales y/o actuadas)[88, 37, 74, 75, 76], etc. Inclusive para el caso de trabajos donde se utiliza la misma base de datos los criterios de selección de las frases de este corpus son diversos y algunos son fundamentados en medidas del test perceptual del mismo [3] [30, 80], generalmente no se balancean los datos [30, 80, 89, 90] y algunos adoptan el criterio de eliminar ciertas emociones o bien la emoción con menos sucesos (*disgust*) para evitar tratar este problema [79, 83, 36, 89, 91, 90]. La alternativa a este inconveniente fue proponer, bajo las mismas condiciones, algunos clasificadores estándar con los cuales se comparan habitualmente las nuevas propuestas [74, 36, 88, 35, 33, 92, 75, 31].

Como ya se mencionó, las matrices de confusión se utilizan para obtener una buena representación de los resultados y para poder hacer un análisis más detallado de los

[3] `http://pascal.kgw.tu-berlin.de/emodb/`

Tabla 3.17: Matriz de confusión del mejor clasificador estándar (HMM).

Emoción	Joy	Anger	Fear	Disgust	Boredom	Neutral	Sadness
Joy	60	16	9	5			
Anger	14	**71**	2	1		2	
Fear	11	5	58	3	5	8	
Disgust	1	3	6	72	5	3	
Boredom			2	6	**55**	21	6
Neutral	2		4	5	27	51	1
Sadness			1	2	11	11	65

Tabla 3.18: Matriz de confusión del clasificador jerárquico 3.

Emoción	Joy	Anger	Fear	Disgust	Boredom	Neutral	Sadness
Joy	**63**	14	8	5			
Anger	17	65	5	1		2	
Fear	6	3	**65**	3	5	8	
Disgust	1	3	6	72	5	3	
Boredom			2	6	53	17	12
Neutral	2		4	5	15	**64**	
Sadness			1	2	10	7	**70**

errores. Las matrices de confusión del modelo HMM estándar y del mejor modelo jerárquico son expuestas en las Tablas 3.17 y 3.18, respectivamente.

Como se puede observar en la Tabla 3.17, las confusiones más importantes suceden dentro de los grupos emocionales JAF y BNS. Por ejemplo, hay 48 errores entre Boredom y Neutral y existen 20 confusiones entre Joy y Fear. Como el modelo jerárquico se enfoca en cada grupo emocional de forma individual, éste es capaz de identificar de una mejor forma las emociones dentro de cada grupo. En la Tabla 3.18, los grupos emocionales pueden verse sombreados y es posible discernir una menor confusión dentro de éstos, a pesar de que permanecen ciertas dificultades para discriminarlos. Evidentemente las confusiones no están completamente resueltas, sin embargo, se han logrado reducir de manera significativa dentro de los grupos. Esto da la pauta de que es apropiado seguir trabajando con esta nueva propuesta de estructuras jerárquicas para continuar mejorando el rendimiento del clasificador.

Por otra parte, no se debería perder de vista que el objetivo de este tipo de sistemas es imitar las capacidades humanas de la mejor forma posible, en este caso reconocien-

TABLA 3.19: Matriz de confusión derivada del análisis perceptual.

Emoción	Joy	Anger	Fear	Disgust	Boredom	Neutral	Sadness
Joy	1324	31	17	9	1	41	4
Anger	8	2483	13	18	5	20	4
Fear	15	28	1270	37		16	18
Disgust	8	3	16	850	7	10	30
Boredom	2	9	4	4	1514	76	20
Neutral	9	10	2	1	40	1495	30
Sadness	1		41	8	45	33	1122

do emociones en la voz. Es interesante constrastar los resultados obtenidos aquí con resultados perceptivos para apreciar donde se presentan dificultades y que similitudes pueden encontrarse cuando se reconocen diferentes emociones. Como se comentó en la sección 3.2, el corpus fue sometido a un test de percepción[4] donde participaron 20 personas[5]. A partir del análisis de los resultados del test perceptual sobre todo el conjunto de datos que fue utilizado en este trabajo se confeccionó una matriz de confusión (Tabla 3.19). En esta matriz se pueden ver los grupos que hemos definido para el clasificador jerárquico (JAF y BNS) con un fondo gris y las mayores confusiones entre emociones subrayadas. Desestimando la alta tasa de reconocimiento propia de las personas y las diferencias de escalas en los valores, se observan similitudes en las emociones que se confunden tanto en el test perceptual como en los resultados de clasificación obtenidos en este trabajo. De la misma forma que en las Tablas 3.17 y 3.18 se encuentran muchas confusiones dentro del grupo BNS, confusiones dentro del grupo JAF y algunas confusiones en la percepción de emociones que no se observaron antes. Es apreciable cierta similitud entre la agrupación natural del sistema de percepción humano y los grupos prosódico-acústicos que se han utilizado. Esta discusión podría continuarse en un trabajo futuro considerando la realización de un análisis similar pero utilizando parametrizaciones basadas en modelos perceptuales humanos tales como los coeficientes de predicción lineal perceptuales (PLP, del inglés Perceptual Linear Prediction) [93] o las representaciones de espectros relativos (RASTA, del inglés RelAtive SpecTrA) [94].

Clasificar los grupos de emociones utilizando similitudes prosódico-acústicas permite atacar el problema de una manera más eficiente. La configuración que logra mejores resultados que los clasificadores de una etapa considera un esquema con 3 grupos emo-

[4]Disponible en `http://pascal.kgw.tu-berlin.de/emodb/`.

[5]Observación: en esta información existe inconsistencia pues, si bien la mayor parte de las frases fueron evaluadas por 20 personas, se encuentran frases evaluadas por 21 personas.

cionales en el Nivel I. Se puede ver la importancia de utilizar todas las características en el bloque JAF para obtener un buen resultado con el MLP. Mientras que, para el caso de BNS ha resultado muy útil considerar las dinámicas temporales mediante HMM.

Los resultados confirman que cada grupo de emociones debería ser tratado con un modelo específico para poder mejorar el rendimiento del reconocedor. Por ejemplo, HMM es mejor para discriminar entre B, N y S mientras que, MLP es mejor para clasificar J, A y F. De la misma forma, se ha mostrado que ciertas características son mejores para distinguir emociones específicas. Como ya se comento para el Nivel I, los coeficientes MFCC, delta y acceleración logran un mejor desempeño discriminando 3 grupos emocionales, mientras que el vector FV46 es mejor para distinguir entre 2 grupos emocionales.

Capítulo 4

Conclusiones

4.1. Aportes en el reconocimiento automático del habla con prosodia

Para esta línea de investigación se llevó adelante un estudio de la información prosódica presente en las palabras en Español. Los análisis se realizaron sobre las palabras de una base de datos de habla continua. La prosodia presente en éstas se manifiesta de forma diferente a la que se encuentra en las palabras que se pronuncian de forma aislada. En este trabajo se dejó de lado el enfoque que contempla la utilización de la información prosódica asociada a la acentuación definida según las reglas ortográficas del idioma español, aunque éste ha significado un punto de partida. En una primer etapa se decidió agrupar a las palabras según su separación silábica definida en las reglas ortográficas y se propuso un método de clasificación de palabras basado en su estuctura prosódica. En el análisis de cada uno de los parámetros prosódicos se cotejaron los valores obtenidos en cada una de las sílabas de una palabra y, conforme a su importancia relativa, se propuso un código para la palabra, asociado al rasgo prosódico analizado. La evaluación de cada uno de los sucesos de cada palabra permitió generar histogramas que dan cuenta de las ocurrencias de las distintas estructuras prosódicas. Este método de los histogramas prosódicos permitió clasificar a las palabras según una estructura de clase particular asociada a cada rasgo prosódico. Las pruebas sobre distintos subconjuntos del corpus permitieron verificar que la clasificación de las palabras da buenos resultados [56, 58], siempre y cuando la segmentación en sílabas sea correcta. La definición de las medidas $(\tilde{h}, \rho, \gamma)$ permitió una valoración más objetiva de los histogramas.

Se propuso un método para incluir la información prosódica en un sistema de reconocimiento automático del habla estándar, basado en la penalización de las probabili-

dades de transición de las hipótesis de las palabras presentes en las redes de hipótesis. Se observó que las posibilidades de influir en el rendimiento del reconocedor, tal y como se plantea la penalización, se ven limitadas a la existencia de hipótesis correctas en las redes. El factor de penalización relativo (F_p), que considera el grado de incoherencia de la hipótesis en relación a la clase prosódica que define a la palabra, permitió efectuar una penalización más justa. También se observó la necesidad de plantear una alternativa a la fuerte dependencia que presenta el método propuesto respecto de la cantidad de sílabas de las palabras en las hipótesis, pues el reconocedor brinda hipótesis con diferentes cantidades de sílabas para un mismo segmento acústico.

Se analizó profundamente la perplejidad de las redes de hipótesis de un sistema de RAH y se definió un método para atacar este problema de una forma más directa. En primer lugar se consideró la independencia de las hipótesis respecto de la cantidad de sílabas y se investigó la relación entre las hipótesis de palabras y los errores cometidos por el sistema de RAH estándar. En este caso, el tema crucial fue determinar que segmentos acústicos resultan confusos para el reconocedor y cómo atacar estas confusiones. Para ésto se propuso la definición de un nuevo corpus, remuestreando las frases a partir de las redes de hipótesis obtenidas del RAH. La nueva información ha sido agrupada por palabras y por cada una, se encuentran diferenciados los segmentos acústicos que dan origen a hipótesis correctas y aquellos que ocasionan hipótesis erróneas. Con este corpus fue posible crear clasificadores binarios de hipótesis verdaderas y falsas. Se definieron vectores de características prosódicas específicos que fueron determinados en base a una clasificación de la capacidad discriminativa de cada una de las características (utilizando la técnica *F-Score*). Por cada palabra se desarrolló un modelo SVM específico para el que se determinó el mejor conjunto de características y la mejor configuración del modelo utilizando validación cruzada durante el entrenamiento. Los resultados muestran que la información prosódica junto a los modelos SVM permiten clasificar los segmentos que están asociados a hipótesis de palabras, verdaderas y falsas, con una buena precisión. La metodología propuesta tiene como objetivo complementar cualquier sistema de reconocimiento del habla, atacando los segmentos que dificultan el modelado acústico. Es importante destacar que ha sido desarrollada para que pueda ser utilizada indistintamente del método de segmentación empleado y es potencialmente aplicable a cualquier idioma independientemente de sus características particulares.

4.2. Aportes en el reconocimiento de emociones

Para abordar el tema de reconocimiento de emociones, inicialmente se propusieron algunas técnicas muy utilizadas en el RAH como lo son los GMM y los HMM. Estos modelos se utilizaron para reconocer hasta 7 emociones diferentes. Se evaluaron las mejores configuraciones de ambos modelos y se analizaron sus resultados utilizando validación cruzada. Se comprobó que, si bien los sistemas basados en GMM obtienen resultados aceptables, los sistemas basados en HMM obtienen un mejor rendimiento y los coeficientes MFCC son apropiados para esta tarea [34].

Luego, se presentó una caracterización de las emociones y un análisis de sus semejanzas en base a las características acústicas y prosódicas. Utilizando estas características se propuso realizar un agrupamiento no supervisado empleando mapas autoorganizativos. En estos primeros análisis se halló una pista importante acerca de las similitudes de las emociones, y se comprobó que los agrupamientos eran similares utilizando diferentes combinaciones de las características extraidas. En un segundo paso se propuso un análisis espectral donde se observaron morfologías similares, en las curvas AMLS propuestas, para ciertos tipos de emociones. También se comprobó, al igual que en estudios anteriores, que la información más discriminativa de estas curvas se presenta en la zona de frecuencias más bajas, incluso considerando un filtro de pre-énfasis de altas frecuencias. Con estos indicios acerca de los agrupamientos prosódico-acústicos se propusieron nuevas clases o grupos emocionales, y se formuló un nuevo método jerárquico para clasificar emociones [64]. Se definieron dos esquemas basados en bloques de grupos acústico-prosódicos, en los que se consideró una primer etapa que debe distinguir entre 2 y 3 clases respectivamente y una segunda etapa que clasifica las emociones de forma particular según su clasificación en la etapa previa. En un esquema se propuso el grupo BNS (Boredom, Natural y Sadness) frente al grupo JAFD (Joy, Anger, Fear y Disgust), mientras que en el otro la clasificación debía hacerse entre BNS, JAF y Disgust. Para ensamblar el modelo final, se tuvo en cuenta el mejor funcionamiento de cada uno de los bloques de forma independiente y con una partición específica para la etapa de diseño. Se realizaron experimentos con diferentes vectores de características y estructuras internas de MLP, para cada bloque. También, se hicieron pruebas incrementando el número de Gaussianas en los modelos GMM y en las mezclas para los HMM. Además se evaluó el número óptimo de estados para el modelo HMM. Con los mejores modelos (características + clasificador) de cada bloque, se ensamblaron las dos propuestas de clasificadores jerárquicos y se evaluaron utilizando validación cruzada. Los resultados mostraron que la información espectral combinada con la prosodia permite el agrupamiento de emociones y puede direccionar el desarrollo de clasificadores jerárquicos. Estos modelos lograron mejorar las tasas de reconocimiento de los clasificadores estándar (de un nivel). Más aún, éstos muestran que la prosodia combinada con

las características espectrales mejoran los resultados en la tarea de clasificar emociones [86].

Los resultados obtenidos pueden ser generalizados a otros hablantes puesto que se ha utilizado un corpus con múltiples hablantes y el método de validación cruzada en los experimentos. Ésto representa un punto interesante para la comparación con trabajos previos donde se presentan resultados de reconocimiento de emociones dependiente del hablante o dependiente del género del hablante [79, 35, 75, 28].

4.3. Trabajos futuros

El próximo trabajo sobre el sistema de RAH es incluir los clasificadores de hipótesis en el sistema de RAH tradicional para mejorar las probabilidades de las hipótesis verdaderas y comparar su desempeño con el que se obtuvo a partir de los histogramas prosódicos. Además, se propone incorporar el esquema de redes expandidas [61] para poder contar con aquellas palabras que no están en el espacio de hipótesis de reconocimiento del RAH. En una etapa posterior se propone desarrollar un esquema on-line (de una pasada) en el sistema de RAH que utilice estos clasificadores puesto que, tal como se ha propuesto en esta Tesis, el método sólo requiere las alineaciones de la hipótesis.

Respecto a los trabajos futuros del área de reconocimiento de emociones se propone evaluar el clasificador jerárquico con señales ruidosas. También se prevé contrastar los resultados obtenidos con los de otro modelo que considere la variabilidad de género de forma explícita. Además, se pretende realizar una adaptación de la información de las señales emocionales considerando la variabilidad del tono, independientemente del género. Está planeado realizar análisis similares con corpus en otros idiomas.

Bibliografía

[1] Albert Mehrabian. Communication without words. *Psychology Today*, 2:53–56, Sep. 1968.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1 edition, Aug. 2006.

[3] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2nd edition, Jul. 1998.

[4] Leszek Rutkowski. *Computational intelligence: methods and techniques*. Springer, Berlin, english edition, Jul. 2008.

[5] T. Kohonen. *The Self-Organizing Map*. Springer-Verlag, 1995.

[6] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[7] D. Milone. Modelos ocultos de Markov para el reconocimiento automático del habla. Technical report, Universidad Nacional del Litoral, Santa Fe, Argentina, Grupo de investigación en señales e inteligencia computacional, Feb. 2004.

[8] John R. Deller, Jr., John G. Proakis, and John H. Hansen. *Discrete-Time Processing of Speech Signals*. Prentice Hall, Upper Saddle River, NJ, USA, 1993.

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

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

[11] Ana María Borzone Manrique. *Manual de Fonéica Acústica*. Hachette, Buenos Aires, 1980.

[12] Antonio Quilis. *Tratado de Fonología y Fonética Españolas*. Biblioteca Románica Hispánica. Ed. Gredos, Madrid, 1993.

[13] Emilio Alarcos Llorach. *Gramática de la Lengua Española*. Real Academia Española. Colección Nebrija y Bello. Editorial Espasa Calpe, Madrid, 1999.

[14] Juan María Garrido Almiñana. *Modelización de Patrones Melódicos del Español para la Síntesis y el Reconocimiento del Habla*. Servei de Publicacions de la Universitat Autònoma de Barcelona, Facultad de Filosofia i Lletres, Departament de Filologia Espanyola, Barcelona, 1991.

[15] Roman Kuc. *Introduction to digital signal processing*. McGraw-Hill Book Company, 1988.

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

[17] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. *The HTK Book (for HTK Version 3.1)*. Cambridge University Engineering Department., England, Dec. 2001.

[18] Hamid Aghajan, Juan Carlos Augusto, and Ramon Lopez-Cozar Delgado. *Human-Centric Interfaces for Ambient Intelligence*. Academic Press, 2009.

[19] Jesse J. Prinz. Which Emotions Are Basic? In D. Evans and Pierre Cruse, editors, *Emotion, Evolution, and Rationality*. Oxford University Press, 2004.

[20] P. Ekman, E. R. Sorenson, and W. V. Friesen. Pan-cultural elements in facial displays of emotions. *Science*, 164(3875):86–88, Apr. 1969.

[21] P. Ekman. *Handbook of Cognition and Emotion*, chapter Basic Emotions. John Wiley & Sons Ltd., 1999.

[22] Jonghwa Kim. *Robust Speech Recognition and Understanding*, chapter Bimodal Emotion Recognition using Speech and Physiological Changes, pages 265–280. I-Tech Education and Publishing, Vienna, Austria, 2007.

[23] Roddy Cowie and Randolph Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32, 2003.

[24] Klaus R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, Dec. 2005.

[25] A. Ortony and T. J. Turner. What's basic about basic emotions? *Psychol Rev*, 97(3):315–331, July 1990.

[26] R. Cowie. Describing the emotional states expressed in speech. In *ISCA Workshop on Speech and Emotion*, pages 11–18, Belfast, 2000.

[27] F. Dellaert, T. Polzin, and A. Waibel. Recognizing Emotions in Speech. In *Proceedings of International Conference on Spoken Language Processing - ICSLP '96*, volume 3, pages 1970–1973, Philadelphia, PA, 1996.

[28] A. Noguerias, A. Moreno, A. Bonafonte, and J. Mariño. Speech Emotion Recognition Using Hidden Markov Models. In *European Conference on Speech Communication and Technology - Eurospeech 2001*, pages 2679–2682, Aalborg, Denmark, Sep. 2001.

[29] Dong-Mei Yu and Jian-An Fang. Research on a methodology to model speech emotion. *Wavelet Analysis and Pattern Recognition, 2007. ICWAPR '07. International Conference on*, 2:825–830, Nov. 2007.

[30] M. Borchert and A. Dusterhoft. Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering. IEEE NLP-KE 2005.*, pages 147–151, Oct. 2005.

[31] Iker Luengo Gil, Eva Navas Cordón, Inmaculada Concepción Hernáez Rioja, and Jon Sánchez de la Fuente. Reconocimiento automático de emociones utilizando parámetros prosódicos. *Procesamiento del lenguaje natural*, (35):13–20, Sep. 2005.

[32] Anton Batliner, Stefan Steidl, Björn Schuller, Dino Seppi, Thurid Vogt, Johannes Wagner, Laurence Devillers, Laurence Vidrascu, Vered Aharonson, Loic Kessous, and Noam Amir. Whodunnit - searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language.*, 25:4–28, January 2011.

[33] Alexander I. Iliev, Michael S. Scordilis, João P. Papa, and Alexandre X. Falcão. Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech & Language.*, 24(3):445–460, July 2010.

[34] Enrique M. Albornoz, María B. Crolla, and Diego H. Milone. Recognition of emotions in speech. In *Proceedings of XXXIV Congreso Latinoamericano de Informática - CLEI*, pages 1120–1129, Santa Fe, Argentina, Sep. 2008.

[35] Yi-Lin Lin and Gang Wei. Speech emotion recognition based on HMM and SVM. In *Proceedings of International Conference on Machine Learning and Cybernetics, 2005.*, volume 8, pages 4898–4901, Aug. 2005.

[36] M.M.H. El Ayadi, M.S. Kamel, and F. Karray. Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2007*, volume 4, pages IV–957–IV–960, Apr. 2007.

[37] Jia Rong, Yi-Ping Phoebe Chen, Morshed Chowdhury, and Gang Li. Acoustic Features Extraction for Emotion Recognition. In *6th IEEE/ACIS International Conference on Computer and Information Science - ICIS 2007*, pages 419–424, Jul. 2007.

[38] D. H. Milone and A. J. Rubio. Prosodic and accentual information for automatic speech recognition. *IEEE Trans. on Speech and Audio Proc.*, 11(4):321–333, Julio 2003.

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

[40] 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, 2001. SOMIB.

[41] S. Ananthakrishnan and S. Narayanan. Improved Speech Recognition using Acoustic and Lexical Correlates of Pitch Accent in a N-Best Rescoring Framework. *Acoustics, Speech and Signal Processing. ICASSP 2007. IEEE International Conference on*, 4:873–876, Apr. 2007.

[42] 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.

[43] S. Ananthakrishnan and S. Narayanan. Prosody-enriched lattices for improved syllable recognition. In *INTERSPEECH-2007*, pages 1813–1816, 2007.

[44] György Szaszák and Klára Vicsi. *Verbal and Nonverbal Communication Behaviours*, volume 4775/2007 of *LNCS*, chapter Using Prosody in Fixed Stress Languages for Improvement of Speech Recognition, pages 138–149. Springer, Berlin, Heidelberg, 2007.

[45] Songfang Huang and Steve Renals. *Machine Learning for Multimodal Interaction*, volume 4892/2008 of *Lecture Notes in Computer Science*, chapter Using Prosodic Features in Language Models for Meetings, pages 192–203. Springer Berlin, 2008.

[46] 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, Oct. 2001. SOMIB.

[47] 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.

[48] 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*, 8(5):519–532, 2000.

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

[50] 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*, volume 1, pages 25–28, Seattle, 1998.

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

[52] 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*, volume 1, pages 311–314, 1999.

[53] A. M. Noll. Cepstrum pitch determination. *The Journal of the Acoustical Society of America*, 41(2):179–195, Feb. 1967.

[54] J. G. Proakis and D. G. Manolakis. *Tratamiento digital de señales: Principios, algoritmos y aplicaciones*. Prentice Hall, Madrid, 3ra edición edition, 1998.

[55] 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*, pages 175–178, Berlin, September 1993.

[56] Enrique M. Albornoz and Diego H. Milone. Construcción de patrones prosódicos para el reconocimiento automático del habla. In *34ta JAIIO*, pages 225–236, Rosario, Argentina, September 2005. Simposio ASAI.

[57] Enrique M. Albornoz and Diego H. Milone. La prosodia en el reconocimiento automático del habla. In *9no. Encuentro de Jóvenes Investigadores de la Univ. Nacional del Litoral*, Grupo SINC(i), Oct. 2005.

[58] Enrique M. Albornoz and Diego H. Milone. Modelado de estructuras acentuales a partir de rasgos prosódicos básicos con modelos ocultos de Markov y su incorporación a un sistema de reconocimiento automático del habla. Technical report, FICH (UNL), Grupo SINC(i), Jun. 2005. Cientibeca.

[59] H. Ney and S. Ortmanns. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, 16(5):64–83, 1999.

[60] Ning Chen and Ying Hu. Pitch Detection Algorithm Based on Teager Energy Operator and Spatial Correlation Function. *Machine Learning and Cybernetics, 2007 International Conference on*, 5:2456–2460, Aug. 2007.

[61] D. H. Milone and A. J. Rubio. Including prosodic cues in ASR systems. In *Proceedings of the 5th World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001) and the 7th International Conference on Information Systems Analysis and Synthesis (ISAS 2001)*, Orlando, Jul. 2001. Paper No. IS0051403.

[62] Paul Boersma and David Weenink. *Praat: doing phonetics by computer [Computer program]*, 2011. Version 5.2.11. Software available at `http://www.praat.org/`.

[63] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

[64] Enrique Albornoz, Diego Milone, and Hugo Rufiner. Multiple feature extraction and hierarchical classifiers for emotions recognition. In *Development of Multimodal Interfaces: Active Listening and Synchrony*, volume Volume 5967/2010 of *Lecture Notes in Computer Science*, pages 242–254. Springer Berlin / Heidelberg, 2010. issn 0302-9743 (Print) 1611-3349 (Online).

[65] Laurence Devillers and Laurence Vidrascu. *Speaker Classification II: Selected Projects*, volume 4441/2007 of *Lecture Notes in Computer Science*, chapter Real-Life Emotion Recognition in Speech, pages 34–42. Springer-Verlag, Berlin, Heidelberg, 2007.

[66] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, 2008.

[67] D. Tacconi, O. Mayora, Paul Lukowicz, Bert Arnrich, Cornelia Setz, Gerhard Tröster, and C. Haring. Activity and emotion recognition to support early diagnosis of psychiatric diseases. In *Proceedings of 2nd International Conference on on Pervasive Computing Technologies for Healthcare '08*, pages 100–102, Tampere, Finland, Feb. 2008.

[68] Serdar Yildirim, Shrikanth Narayanan, and Alexandros Potamianos. Detecting emotional state of a child in a conversational computer game. *Computer Speech & Language*, 25(1):29 – 44, 2011. Affective Speech in Real-Life Interactions.

[69] Julia Hirschberg, Stefan Benus, Jason M. Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Gir, Martin Graciarena, Andreas Kathol, and Laura Michaelis. Distinguishing Deceptive from Non-Deceptive Speech. In *In Proceedings of Interspeech'2005 - Eurospeech*, pages 1833–1836, 2005.

[70] Jonghwa Kim and Elisabeth André. Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(12):2067–2083, Dec. 2008.

[71] Konrad Schindler, Luc Van Gool, and Beatrice de Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Networks*, 21(9):1238–1246, 2008.

[72] Vasco Vinhas, Luís Paulo Reis, and Eugénio Oliveira. Dynamic Multimedia Content Delivery Based on Real-Time User Emotions. Multichannel Online Biosignals Towards Adaptative GUI and Content Delivery. In *International Conference on Bio-inspired Systems and Signal Processing - Biosignals 2009*, pages 299–304, Porto, Portugal, 2009.

[73] Khiet P. Truong and David A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007.

[74] Donn Morrison, Ruili Wang, and Liyanage C. De Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112, 2007.

[75] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief

network architecture. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP '04*, pages I–577–80, May 2004.

[76] Liqin Fu, Xia Mao, and Lijiang Chen. Speaker independent emotion recognition based on SVM/HMMs fusion system. *Audio, Language and Image Processing. ICALIP 2008. International Conf. on*, pages 61–65, Jul. 2008.

[77] C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan. Emotion recognition using a hierarchical binary decision tree approach. In *Interspeech 2009*, pages 320–323, Brighton, UK, 2009.

[78] R. Lazarus. *Appraisal Processes in Emotion: Theory, Methods, Research (Series in Affective Science)*, chapter Relational meaning and discrete emotions, pages 37–67. Oxford University Press, USA, Feb. 2001.

[79] Zhongzhe Xiao, Emmanuel Dellandréa, Weibei Dou, and Liming Chen. Recognition of emotions in speech by a hierarchical approach. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 312–319, Sep. 2009.

[80] Johannes Wagner, Thurid Vogt, and Elisabeth André. A Systematic Comparison of Different HMM Designs for Emotion Recognition from Acted and Spontaneous Speech. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction - ACII '07*, pages 114–125, Berlin, Heidelberg, 2007. Springer-Verlag.

[81] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. In *9th European Conference on Speech Communication and Technology - Interspeech'2005*, pages 1517–1520, Sep. 2005.

[82] B. Schuller, B. Vlasenko, D. Arsic, G. Rigoll, and A. Wendemuth. Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition. In *IEEE International Conference on Multimedia and Expo '08*, pages 1333–1336, Apr. 2008.

[83] B. Yang and M. Lugger. Emotion recognition from speech signals using new harmony features. *Signal Processing*, 90(5):1415 – 1423, 2010. Special Section on Statistical Signal & Array Processing.

[84] D. Michie, D.J. Spiegelhalter, and C.C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, University College, London, 1994.

[85] J. Adell Mercado, A. Bonafonte Cávez, and D. Escudero Mancebo. Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech. In *XXI Congreso de la Sociedad Española. Procesamiento del Lenguaje Natural - SEPLN 2005*, number 35, pages 277–284, Granada, España, Sep. 2005.

[86] Enrique M. Albornoz, Diego H. Milone, and Hugo L. Rufiner. Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language*, 25(3):556–570, 2011.

[87] Andreas Zell, Gunter Mamier, Michael Vogt, Niels Mache, Ralf Hubner, Sven Doring, Kai-Uwe Herrmann, Tobias Soyez, Michael Schmalzl, Tilman Sommer, Artemis Hatzigeorgiou, Dietmar Posselt, Tobias Schreiner, Bernward Kett, and Gianfranco Clemente. *SNNS (Stuttgart Neural Network Simulator) - user manual Version 4.2*. 70565 Stuttgart, Fed. Rep. of Germany, 1998. SNNS User Manual Version 4.0.

[88] Tsang-Long Pao, Wen-Yuan Liao, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng, and C.S. Chien. Comparison of Several Classifiers for Emotion Recognition from Noisy Mandarin Speech. *Intelligent Information Hiding and Multimedia Signal Processing. IIHMSP 2007. Third International Conference on*, 1:23–26, Nov. 2007.

[89] M. Gaurav. Performance analysis of spectral and prosodic features and their fusion for emotion recognition in speech. In *Spoken Language Technology Workshop, 2008. (SLT 2008). IEEE*, pages 313–316, dec. 2008.

[90] Yashpalsing Chavhan, M. L. Dhore, and Pallavi Yesaware. Speech emotion recognition using support vector machine. In *International Journal of Computer Applications*. Foundation of Computer Science (Sweden), 2010.

[91] M. Lugger, M. Janoir, and B. Yang. Combining classifiers with diverse feature sets for robust speaker independent emotion recognition. In *Proceedings of European Signal Processing Conference (EUSIPCO)*, pages 1225–1229, Glasgow, UK, Aug. 2009.

[92] B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 2*, ICME '03, pages 401–404, Washington, DC, USA, 2003. IEEE Computer Society.

[93] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[94] H. Hermansky and N. Morgan. Rasta processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2(4):578–589, October 1994.