UNIVERSIDAD POLITÉCNICA DE VALENCIA

UNIVERSIDAD
POLITECNICA
DE VALENCIA

Departamento de Sistemas Informáticos y Computación

Tesis de Máster:

Reconocimiento Automático del Habla Multilingüe[1]

Autora:

Míriam Luján Mares

Directores:

Carlos D. Martínez Hinarejos, Vicente Alabau Gonzalvo

Septiembre 2008

[1] Trabajo parcialmente subvencionado por VIDI-UPV bajo las becas FPI del programa PAID06 y por el EC (FEDER) y el proyecto subvencionado por el Ministerio de Educación y Ciencia Español TIN2006-15694-C02-01.

Resumen

El reconocimiento automático del habla es uno de los campos pioneros en el reconocimiento de formas y tecnologías del lenguaje. Desde su comienzo se han producido avances muy importantes en las tecnologías del habla. El estado del arte actual consigue unas tasas de acierto elevadas para tareas complejas [39]. Esto ha permitido que la investigación en el reconocimiento del habla haya dado un paso adelante y se centre en aspectos más avanzados como la tolerancia al ruido [8, 11], adaptación al locutor [21] y reconocimiento multilingüe [35, 17].

Este último aspecto es el que se va a tratar en profundidad en este trabajo. En concreto, se estudiará el caso de la Comunidad Valenciana, cuyos idiomas oficiales son el castellano y el valenciano. El castellano y el valenciano presentan ciertas peculiaridades que los hacen particularmente interesantes: su parecido fonético y gramatical.

En el capítulo 1 se presentará la aproximación al reconocimiento del habla tradicional para poder sentar las bases necesarias para entender el trabajo que se mostrará en el resto de la memoria.

En el capítulo 2 explicaremos las características principales de los reconocedores que hemos utilizado en este trabajo.

Seguidamente, en el capítulo 3 hablaremos de los sistemas multilingües y de técnicas de adaptación e identificación del lenguaje que hemos implementado y con las que vamos a llevar a cabo la experimentación.

En el capítulo 4 se explica la tarea que vamos a utilizar. En los capítulos 5 y 6 mostraremos los resultados que hemos obtenido con todas las técnicas que hemos empleado.

Finalmente en el capítulo 7 se comentarán las conclusiones y la línea a seguir para el trabajo futuro.

Abstract

Automatic speech recognition is a pioneering field in pattern recognition and language technology. Some important advances have been achieved in speech technologies in the last decade. The current state of the art achieves good performance in complex tasks [39]. For this reason, the research in speech recognition has moved towards more advanced areas, such as noise tolerance [8, 11], speaker adaptation [21] and multilingual recognition [35, 17].

Multilingual recognition is the topic of this work. We studied specifically the case of the Comunitat Valenciana, where the two official languages are Spanish and Valencian. These two languages share most of their phonemes, and their syntax and vocabulary are also quite similar since they have influenced each other for many years.

In the chapter 1 we will presente the theory about speech recognition to understand the complete work.

In the chapter 2 we will explain the main features of the recognisers that we used in this work.

Next, in the chapter 3 we will talk about multilingual systems and adaptation techniques and language identification techniques that we have implemented and we used in the experimentation.

In the chapter 4 we explain the task that we used.

In the chapters 5 and 6 we will present the results that we obtained with every technique that we used.

Finally in the chapter 7 we will summarize the conclusions and the future work.

Índice general

CAPÍTULO 1

INTRODUCCIÓN

1.1. Reconocimiento del habla

El habla es la materialización individual de los pensamientos de una persona, sirviéndose del modelo o sistema que facilita el lenguaje. Esto es uno de los motivos que marca la diferencia de un ser humano al resto de los animales. El lenguaje hablado es la forma más natural y eficiente de comunicación entre los seres humanos. Por lo tanto, existe un alto interés en esta área para conseguir una comunicación hombre-máquina tan natural e intuitiva como sea posible.

La comunicación por medio del habla con los ordenadores es posible porque en la actualidad disponemos de dispositivos (elementos electrónicos) que hacen posible la adquisición del sonido, el procesamiento de la señal y la salida de la respuesta tanto hacia el computador como hacia el ser humano. Así, el proceso de comunicación por medio del habla hombre-máquina puede dividirse en dos vías con distinto tratamiento:

- Proceso del ser humano al computador: el ordenador trata de entender lo que se le ha dicho por medio de dos procesos:
  - RAH: Reconocimiento Automático del Habla; trata de encontrar, a partir de la señal vocal, las palabras del lenguaje a la que corresponde [29].
  - PLN: Procesamiento del Lenguaje Natural; trata de encontrar el significado de una frase transcrita, verificando su corrección sintáctica y semántica [4].
- Proceso del computador al ser humano:
  - Síntesis del habla: Es el proceso en el que se crea voz a partir de texto; la calidad de la voz sintética vendrá dada por su inteligibilidad y su naturalidad [9].

En nuestro caso, nos centraremos en el Reconocimiento Automático del Habla (RAH). Las líneas de investigación que se han seguido en el RAH han sido dos vertientes muy diferentes:

- Basada en el conocimiento: *Knowledge-based*.

  Pretende que el reconocimiento se lleve a cabo a partir de reglas acústico-fonéticas que se basan en características de la forma de la onda de entrada. Los resultados conseguidos con esta línea de investigación podríamos decir que han sido muy pobres, debido a la incapacidad del ser humano para formalizar su conocimiento y poder implementar éste como reglas en un computador.

- Basada en los datos o en la estadística: *Data-based*.

  Busca la extracción del ”conocimiento” a partir de la propia forma de onda mediante un análisis estadístico. Esta línea ha obtenido mejores resultados ya que los análisis estadísticos son fácilmente describibles por un algoritmo.

Para llevar a cabo el reconocimiento del habla son necesarios algunos elementos que describiremos de forma muy breve a continuación.

1.2. Estructura de un reconocedor de habla

Un reconocedor de habla busca hallar a partir de una onda sonora las palabras del lenguaje a la que corresponde, siguiendo un esquema como el de la Figura 1.1.

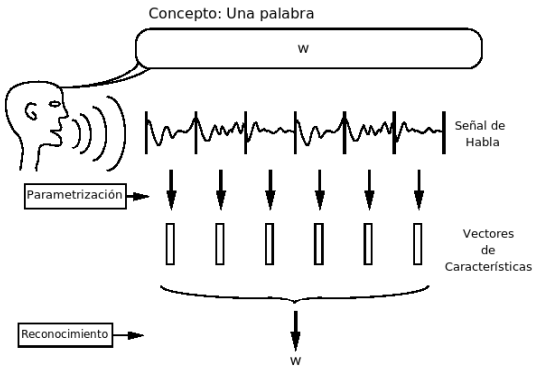

Figura 1.1: Esquema de un sistema de R.A.H.

Un reconocedor de habla actual (basado en datos) necesita una serie de modelos para funcionar y su funcionamiento se basa en el algoritmo de Viterbi. El modelo principal se estructura por niveles que van de lo más abstracto (frases) a lo más concreto (fonemas). Dichos modelos y el algoritmo de Viterbi se explican a continuación.

1.2.1. Modelos de lenguaje

El modelo de lenguaje define el conjunto de palabras, con un cierto orden, que debe de reconocer nuestro sistema de reconocimiento del habla. Es decir, define el conjunto de frases que se le van a introducir al sistema.

Por esta razón, el modelo de lenguaje viene determinado por la tarea a la que se vaya a destinar el sistema de reconocimiento del habla. Esto se debe a que la tarea fija las frases que se van a usar en el sistema, ya que no tendría ningún sentido incluir frases en las que se pregunta por el clima de una ciudad si lo que queremos que lleve a cabo nuestro sistema es atención al público por alguna empresa.

1.2.2. Modelos léxicos

Los modelos léxicos describen la pronunciación de las palabras que forman las frases del modelo de lenguaje. La pronunciación de dichas palabras se describe con una secuencia de símbolos que identifican los modelos acústicos que forman la palabra. Es decir, cada palabra está representada por un autómata que emite etiquetas de modelos acústicos en las aristas, lo que permite tener diferentes pronunciaciones de la misma palabra.

1.2.3. Modelos acústicos

Los modelos acústicos van asociados a cada sonido, de forma que se pueda hacer una comparación de las características de una secuencia acústica con los modelos descritos. De esta manera se trata de llegar a obtener una frase del modelo de lenguaje de la tarea, como una hipótesis de la frase que se pronuncia, comparando la señal con los sonidos modelados por estos modelos.

Los modelos acústicos que se utilizan habitualmente son los modelos ocultos de Markov (Hidden Markov Models (HMM)) [12]. Esto se debe a que es la tecnología para la implementación de modelos acústicos que goza de mejores resultados.

Un modelo oculto de Markov es un caso especial de un proceso de Markov, los cuales modelan cambios de estado de una variable a lo largo del tiempo. El habla humana presenta una característica importante: se desarrolla como un cambio de estados; es decir, podemos decir que que se cambia de un estado a otro al cambiar de fonema.

1.2.4. Marco estadístico

Los modelos definidos nos sirven para que la frase pronunciada por un locutor, que se convierte en una secuencia de observaciones O, pase a ser reconocida como una secuencia de palabras $\vec{W}'$.

Sea una pronunciación de una frase representada por una secuencia de vectores de características u *observaciones* $\vec{O}$, definida como

$$\vec{O} = \vec{o}_1, \vec{o}_2, \ldots, \vec{o}_T \tag{1.1}$$

donde $\vec{o}_t, t = 1, \ldots, T$, es el vector de características observado en el tiempo t. El problema del reconocimiento de una secuencia de N palabras $\vec{W}$ definida como

$$\vec{W} = w_1, w_2, \ldots, w_N \tag{1.2}$$

puede ser visto como el cálculo de

$$\vec{W} = \arg\max_{\vec{W}'} \Pr(\vec{W}'|\vec{O}) \tag{1.3}$$

donde $\vec{W}'$ es una frase perteneciente al lenguaje de la aplicación. Esta probabilidad es difícil de calcular directamente pero usando la regla de Bayes tenemos

$$\vec{W} = \arg\max_{\vec{W}'} \Pr(\vec{W}'|\vec{O}) = \arg\max_{\vec{W}'} \Pr(\vec{W}') \Pr(\vec{O}|\vec{W}') \tag{1.4}$$

donde $\Pr(\vec{O}|\vec{W}')$ es la probabilidad de que, dada la secuencia de palabras $\vec{W}$, ésta pueda generar la secuencia de observaciones $\vec{O}$, y $\Pr(\vec{W}')$ define las secuencias de palabras posibles y sus probabilidades asociadas. Así pues, $\Pr(\vec{O}|\vec{W}')$ viene dado por los modelos acústicos, y $\Pr(\vec{W}')$ por el modelo de lenguaje.

Las dos aproximaciones más usuales para construir modelos de lenguaje son:

- Autómatas de estados finitos [10]: se construyen de forma manual, semi-automática o automática (gracias a técnicas de inferencia gramatical aplicables sobre un corpus de entrenamiento); se suelen usar en dominios muy específicos y sencillos. El modelo correspondiente a este nivel es un autómata que emite una palabra del vocabulario en cada arco y, por lo tanto, va generando una frase a medida que va avanzando por el autómata.
- N-gramas [24]: es un conjunto de secuencias de N palabras que indica la frecuencia de aparición de dichas secuencias. Se obtienen a partir de un

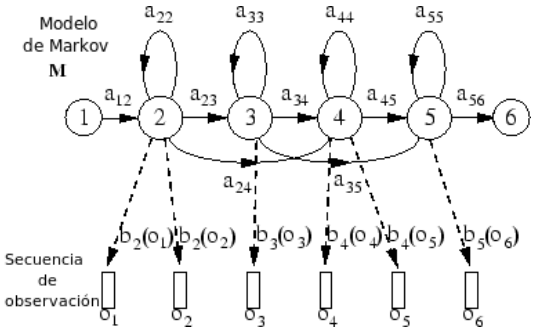

Figura 1.2: Ejemplo de HMM.

corpus de entrenamiento. Se emplean en tareas con un número elevado de palabras en el vocabulario porque permiten una mayor expresividad para tareas complejas.

Volviendo a la ecuación (1.4), dada la dimensionalidad de los vectores de observaciones $\vec{O}$, la estimación directa de la probabilidad condicional $\Pr(\vec{o}_1, \ldots, \vec{o}_T|\vec{W}')$ a partir de ejemplos de frases no es práctica. Sin embargo, si se asume un modelo paramétrico de producción de palabras como un modelo de Markov, la estimación a partir de los datos es posible, dado que el problema de estimar las probabilidades de densidad condicionales $\Pr(\vec{O}|\vec{W}')$ se convierte en el problema, mucho más simple, de estimar los parámetros de un modelo de Markov.

La definición formal de un HMM suele hacerse mediante la notación $\lambda = (A, B, \pi)$ en la cual A indica las probabilidades de transición (es decir, a_{ij} indica la probabilidad de pasar del estado i al estado j), B da las probabilidades de emisión (es decir, $b_j(O)$ indica la probabilidad de que en el estado j se emita el símbolo O) y π da la distribución de probabilidad de estados iniciales (es decir, π_i indica la probabilidad de que el estado i sea inicial). Para modelos continuos B se define habitualmente como una mixtura de gaussianas sobre las observaciones vistas en el entrenamiento. Esta es la aproximación que va a ser empleada a lo largo de este trabajo.

Para definir completamente un HMM se debe definir el número de estados, las características de las salidas (conjunto de símbolos en el caso discreto, número de componentes y dimensión de los vectores en el caso continuo) y las distribuciones de probabilidad A, B y π. De forma opcional, se puede definir un conjunto de estados finales. Se muestra un ejemplo de HMM en la Figura 1.2. También cabe comentar que en los HMM de primer orden suelen hacerse dos asunciones básicas:

- **Asunción de Markov**: en cada instante de la observación *t* se llega a un nuevo estado según la distribución de la probabilidad de transición A, la cual depende únicamente del estado previo.
- **Asunción de independencia de salidas**: las salidas producidas dependen

únicamente del estado actual, y no del instante actual ni de cómo se llegó al estado.

En los modelos de reconocimiento del habla basados en HMM, se asume que la secuencia de vectores de características observadas corresponde a un modelo de Markov como el representado en la Figura 1.2.

Un modelo de Markov es un autómata de estados finitos en el que en cada tiempo t se entra en un estado j y se genera un vector de características según la distribución de probabilidad b_j. Además, las transiciones entre dos estados i y j son también probabilísticas y se gobiernan por una distribución de probabilidad discreta a_{ij}. Por tanto, la probabilidad de que los vectores de observaciones $\vec{O}$ sean generados por el modelo oculto de Markov $\lambda = (A, B, \pi)$ moviéndose a través de una secuencia de estados $\vec{X} = x_0 x_1 \dots x_{T+1}$ se calcula simplemente por el producto de las probabilidades de transición por las probabilidades de emisión de los estados que atraviesa. Por tanto:

$$\Pr(\vec{O}, \vec{X}|\lambda) = a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(\vec{o}_t) a_{x_t x_{(t+1)}} \tag{1.5}$$

donde x_0 es el estado inicial y $x_{(T+1)}$ el estado final, ambos estados no emisores.

Sin embargo, en la práctica sólo se conoce $\vec{O}$ y la secuencia de estados subyacente $\vec{X}$ es desconocida. Este es el motivo por el cual se conoce como modelos ocultos de Markov.

Dado que $\vec{X}$ es desconocida, la probabilidad de $\vec{O}$ puede calcularse sumando la probabilidad condicional para todas las posibles secuencias de estados, esto es:

$$\Pr(\vec{O}|\lambda) = \sum_{X} \Pr(\vec{O}, \vec{X}|\lambda) = \sum_{X} a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(\vec{o}_t) a_{x_t x_{(t+1)}} \tag{1.6}$$

La probabilidad $\Pr(\vec{O}|\lambda)$ puede ser calculada aproximadamente según la expresión

$$\Pr(\vec{O}|\lambda) \simeq \hat{\Pr}(\vec{O}|\lambda) = \max_{X} a_{x_0 x_1} \prod_{t=1}^{T} b_{x_t}(\vec{o}_t) a_{x_t x_{(t+1)}} \tag{1.7}$$

comúnmente conocida como 'aproximación a la Viterbi' [36, 7]. Decir que $\hat{}$ suele indicar el óptimo.

Las ecuaciones (1.6) y (1.7) pueden ser calculadas de forma eficiente mediante técnicas de programación dinámica, pudiendo llevar a cabo poda en el proceso de

exploración. Es muy corriente expresarlo de forma recursiva por claridad, pero una implementación recursiva directa es muy ineficiente.

Notando $\delta_t(j)$ como la verosimilitud máxima de observar la secuencia de vectores de características $\vec{o}_1, \vec{o}_2, \ldots, \vec{o}_t$ y estando en el estado j en el instante de tiempo t, S_F como el conjunto de estados finales y $\psi_t(j)$ como el mejor camino hasta el instante t para el estado j, el algoritmo de Viterbi se puede expresar como sigue:

1. **Inicialización**: Para todo i estado Hacer

$$\delta_1(i) = \pi_i b_i(o_1); \psi_1(i) = 0$$

2. **Recursión**: Para t=2 Hasta T Para todo j estado Hacer

$$\delta_t(j) = \max_i[\delta_{t-1}(i)a_{ij}]b_j(o_t)$$

$$\psi_t(j) = \arg\max_i[\delta_{t-1}(i)a_{ij}]$$

3. **Terminación**:

$$\hat{\Pr} = \max_{s \in S_F}[\delta_T(s)]$$

$$\hat{s_T} = \arg\max_{s \in S_F}[\delta_T(s)]$$

4. **Recuperación del camino**: Para t=T-1 DecreciendoHasta 1 Hacer

$$\hat{s}_t = \psi_{t+1}(\hat{s}_{t+1})$$

Para comprender mejor el algoritmo, resaltar que $\hat{\Pr}$ da la probabilidad del camino más probable, $\hat{s}_T$ da el estado final más probable y los $\hat{s}_t$ constituyen la ruta de estados más probables.

1.3. Reconocimiento multilingüe

En los últimos años, los sistemas de reconocimiento de habla han sido mejorados, lo que ha permitido emplearlos en escenarios más reales y esto ha conllevado que se hayan planteado nuevos problemas con un grado de complejidad mayor, por ejemplo, los sistemas multilingües. Un sistema multilingüe es un sistema de Reconocimiento Automático del Habla (RAH) que reconoce locuciones en distintas lenguas.

Para construir un sistema multilingüe es necesario un corpus multilingüe con un número elevado de horas de grabación para cada lengua a reconocer, lo que implica que solamente un número reducido de lenguas pueda ser estudiado, ya que

sólo aquellas lenguas con suficiente apoyo económico y político serán empleadas en la grabación de corpora. Esto implica que las lenguas minoritarias cuentan con recursos limitados. Por esta razón, en la actualidad organizaciones como ”European Language Resources Association” (ELRA) o ”The Linguistic Data Consortium” (LDC) cooperan para reducir los costes e incrementar la disponibilidad de corpora. En [25] se puede encontrar una descripción de los prerrequisitos técnicos, legales y comerciales que deben ser considerados para producir recursos lingüísticos en un marco cooperativo. Gracias al interés suscitado por los sistemas multilingües se pueden encontrar corpus multilingües, como C-ORAL-ROM [18] o TC-STAR [1], entre otros.

Además, cada lengua tiene características únicas y diferentes del resto de lenguas, siendo necesario desarrollar modelos diferentes para cada una, aunque se puede compartir conocimiento común a todas ellas para obtener los modelos (como es el caso de sonidos o fonemas comunes a varias lenguas). Los modelos se pueden compartir en tres niveles: modelos acústicos, modelos léxicos y modelos de lenguaje. Para construir un reconocedor que sea capaz de reconocer simultáneamente varias lenguas se pueden emplear modelos dependientes para cada lengua o una combinación de modelos independientes del idioma entrenados con todas las lenguas [2, 35].

Otro aspecto importante en un sistema multilingüe es la identificación del idioma. Suele ser necesario hacer un estudio de identificación del idioma, ya sea antes del reconocimiento para determinar los modelos a utilizar o después del reconocimiento para determinar la lengua en la que el sistema debe responder al usuario [42].

Todas estas problemáticas se han estudiado con detalle en [32], donde se puede encontrar un detallado resumen del estado del arte en reconocimiento multilingüe.

1.4. Objetivos del trabajo

El objetivo de este trabajo es, básicamente, construir un sistema multilingüe castellano-valenciano para un sistema de información. La idea principal es trabajar con un corpus multilingüe de una tarea sintética para determinar las condiciones idóneas en un sistema preliminar y tratar de adaptarlas después a tareas reales.

Los sistemas de reconocimiento automático del habla multilingüe son de gran interés en entornos multilingües, pero conllevan muchos problemas, pues deben soportan muchos lenguajes y algunos de ellos son lenguas minoritarias (como el caso del valenciano). Por lo tanto, para construir un sistema multilingüe debemos tener en cuenta aspectos importantes tanto en el modelado de lenguaje como en los modelos acústicos.

Los sistemas multilingües deben incluir tantos modelos de lenguaje como lenguas puedan soportar. Esto implica que el sistema debe ser capaz de determinar la lengua que está siendo hablada. Además, la gran influencia que producen unas lenguas en otras implica una inapropiada pronunciación de algunos fonemas (caso claramente reflejado en la Comunidad Valenciana), teniendo en cuenta que cada lengua define sus propios fonemas (que pueden ser diferentes a los de otras lenguas). Por otro lado, si hablamos de lenguas minoritarias estamos hablando de pocos recursos para construir el sistema, aunque en la actualidad se está mostrando mucho interés por las lenguas minoritarias, ya que son lenguas con muy pocos recursos económicos pero en las que se pueden obtener buenos modelos con tasas de acierto competitivas. Para ello se requiere un buen estudio de las lenguas y la aplicación de técnicas de modelado multilingüe. Por esta razón, vamos a estudiar el caso de la Comunidad Valenciana con una lengua minoritaria (valenciano).

Por lo tanto, en este trabajo vamos a intentar solucionar todos los problemas que conlleva construir un sistema multilingüe, experimentando con todas las combinaciones de modelos y técnicas disponibles, con el fin de encontrar la combinación de mejores resultados.

CAPÍTULO 2

RECONOCEDORES DE HABLA DISPONIBLES

Para construir un sistema de reconocimiento automático del habla multilingüe necesitamos un sistema de reconocimiento. Actualmente, disponemos de dos reconocedores propios: ATROS (Automatically Trainable Recognition Of Speech) e iATROS (Improved ATROS). Además, existen reconocedores disponibles públicamente como HTK [40] o SPHINX [13], pero no se realizarán experimentos con ellos.

La mayor parte de la experimentación expuesta se llevó a cabo con el reconocedor ATROS. Debido a sus limitaciones, se implementó el reconocedor iATROS para poder mejorar los resultados obtenidos.

A continuación detallamos los dos reconocedores, ya que la implementación de iATROS ha sido llevada a cabo en buena parte por la autora de esta tesis.

2.1. ATROS

ATROS es un sistema de reconocimiento automático del habla basado en modelos acústicos estocásticos, modelos léxicos y modelos sintácticos de estados finitos. Se puede encontrar una explicación más detallada del sistema ATROS en [22].

Todos estos modelos pueden ser obtenidos de forma automática. Esto hace que el sistema sea fácilmente adaptable para diferentes tareas de reconocimiento. Por esta razón, lo utilizamos para nuestros experimentos.

ATROS puede ser considerado como un sistema compuesto por dos niveles diferentes:

- Nivel de preprocesamiento acústico: este nivel constituye el preproceso y la extracción de características (parametrización) en el que la información importante es extraída de la señal de habla. Este proceso de parametrización transforma la señal de habla en una secuencia de vectores de coeficientes cepstrales [6].

- Nivel de decodificación: este proceso se lleva a cabo con la información que aportan los modelos acústicos, modelos léxicos y modelos de lenguaje. De este proceso, se obtiene la secuencia de palabras pronunciadas en la señal acústica.

El sistema ATROS soporta modelos con las siguientes características:

- Modelos acústicos: cada uno de los fonemas (incluídos los silencios) son descritos con HMM continuos, donde la densidad de probabilidad de emisión de los vectores de características se asocia con los estados del HMM. ATROS sólo trabaja con HMM compuestos de mixturas de gausianas.

- Modelos léxicos: cada palabra se describe con un autómata de estados finitos, con fonemas asociados en los arcos.

- Modelos de lenguaje: los modelos de lenguaje sólo pueden ser autómatas de estados finitos, con palabras del vocabulario en los arcos, que modelan las frases que pueden darse en el sistema. Para dar más o menos peso a la información que nos aporta el modelo de lenguaje se emplea un parámetro llamado ”Grammar Scale Factor” (GSF). Además, se puede controlar que no se decodifiquen muchas palabras penalizando la inserción de palabras con un parámetro llamado ”Word Insertion Penalty” (WIP).

La decodificación en ATROS se basa en el algoritmo de Viterbi explicado en el apartado 1.2.4. El algoritmo obtiene la mejor secuencia de palabras por programación dinámica, pero el coste computacional puede ser alto para algunas tareas. Por lo tanto, para reducir el coste temporal y espacial se aplican técnicas de 'beam search' [38]. Estas técnicas podan los estados que en cada etapa no superan una cierta probabilidad dependiente de la mejor probabilidad en la etapa actual.

2.2. iATROS

iATROS es un sistema de reconocimiento muy parecido a ATROS, pero en iATROS se han implementado algunas funcionalidades que ATROS no soportaba. Es un sistema de reconocimiento automático de habla y de texto basado en modelos

morfológicos estocásticos (para voz y texto manuscrito off-line), modelos léxicos y modelos sintácticos.

iATROS puede ser considerado como un sistema compuesto por dos niveles diferentes:

- Nivel de preprocesamiento: en el caso acústico, este nivel constituye el preproceso y la extracción de características (parametrización) en la que la información importante es extraída de la señal de habla. Este proceso de parametrización transforma la señal de habla en una secuencia de vectores de coeficientes cepstrales.

  El aparato fonador humano puede representarse como la combinación de un generador de impulsos (a partir del tono fundamental) y un generador de ruido aleatorio, ambos asociados a un interruptor que elige uno u otro (voz sonora/sorda). La onda generada ha de pasar por un filtro lineal variable con el tiempo (el tracto vocal). La extracción de características busca modelar el habla eliminando cualquier rasgo propio del hablante, esto es, se centra en la extracción de los filtros lineales.

  Para extraer esa información iATROS se basa en el esquema típico de preproceso de audio, que consta de los siguientes pasos que son similares a los que realiza ATROS [33]:

  1. Adquisición de la señal.
  2. Aplicación de preénfasis.
  3. Aplicación de la ventana de Hamming.
  4. Aplicación de la Tranformada Rápida de Fourier (FFT).
  5. Transición a la escala de Mel.
  6. Aplicación del logaritmo.
  7. Aplicación de la Transformada Discreta del Coseno (DCT).

  En el caso de texto manuscrito, esta etapa pretende obtener aquellas características de la imagen que permitan discriminar lo mejor posible entre los patrones de cada clase, y al mismo tiempo elimina la información redundante. En esta fase se obtiene una representación compacta y discriminante de la señal de entrada en forma de vectores de características.

  En el caso de texto manuscrito se puede hablar de dos niveles: de página y línea de texto. Para llevar a cabo el preproceso de texto manuscrito iATROS se basa en los siguientes pasos [14]:

  1. Umbralización.
  2. Reducción del ruido.

3. Corrección del *skew* o desencuadre (falta de alineamiento del documento de papel con respecto a las coordenadas del escáner utilizado para su digitalización).
4. Segmentación en líneas.
5. Corrección del *slope* (pendiente o inclinación que presenta la línea base sobre la que está escrita una palabra).
6. Corrección del *slant* (ángulo que presentan las componentes verticales de los carácteres respecto al eje vertical).
7. Normalización del tamaño.

La extracción de características en texto manuscrito consiste en los siguientes pasos:

- Nivel de gris normalizado: explica la densidad del trazo en la zona analizada.
- Derivada vertical y horizontal: explican cómo varía esta densidad con respecto a los ejes de coordenadas.

- Nivel de decodificación: este proceso se lleva a cabo con la información que aportan los modelos acústicos o morfológicos, modelos léxicos y modelos de lenguaje. De este proceso se obtiene la secuencia de palabras subyacente en la señal.

  ATROS sólo trabajaba con habla, mientras que iATROS puede trabajar con habla y escritura. Esto se refleja en que ATROS calculaba automáticamente las derivadas y la aceleración de los vectores de características que se le suministraban, algo sólo útil para habla. En cambio, iATROS trabaja con los vectores de características que se le proporcionan, siendo el cálculo de las derivadas y la aceleración opcional. Por esta razón, iATROS puede trabajar con escritura, ya que en escritura no es necesario obtener las derivadas ni la aceleración de los vectores de características. Además, es importante recalcar que ATROS sólo trabaja con autómatas de estados finitos como modelo de lenguaje, mientras que iATROS puede trabajar también con n-gramas de cualquier grado como modelo de lenguaje.

El sistema iATROS soporta modelos con las siguientes características:

- Modelos acústicos: cada uno de los fonemas u otras unidades de sonido (incluídos los silencios) son descritos con HMM continuos, donde la densidad de probabilidad de emisión de los vectores de características se asocia con los estados del HMM. iATROS en principio sólo trabaja con HMM compuestos de mixturas de gaussianas, pero ha sido implementado para que se pudieran utilizar más tipos de distribuciones fácilmente. iATROS trabaja también con modelos morfológicos para llevar a cabo la decodificación de texto.

- Modelos léxicos: cada palabra se describe con un autómata de estados finitos, con fonemas asociados en los arcos. Además, se soportan múltiples pronunciaciones en las palabras del léxico.

- Modelos de lenguaje: los modelos de lenguaje pueden ser autómatas de estados finitos o n-gramas de cualquier orden, lo que da la posibilidad de que los sistemas sean más flexibles ante las frases de la tarea que pueden ser aceptadas. Los conceptos de GSF y WIP también son aplicables en la decodificación con estos modelos de lenguaje.

La decodificación en iATROS se basa en el algoritmo de Viterbi, en el que también se aplican técnicas de ”beam search”. iATROS basa la poda en un solo factor de poda, mientras que ATROS aplica dos factores, uno aplicado a nivel de modelo acústico y el otro a nivel de modelo de lenguaje. Además, se aplican podas para que la expansión del modelo de lenguaje no sea demasiado pesada, ya que para las n-gramas la expansión del modelo de lenguaje dispara el número de posibilidades[1]. Los parámetros de búsqueda son configurables mediante un fichero de configuración. Entre otras mejoras del nuevo iATROS, se pueden obtener los grafos de palabras o elegir entre una decodificación parcial o completa.

2.3. Diferencias principales entre los reconocedores usados

Entre los dos sistemas podemos encontrar diferencias importantes. iATROS incorpora:

- Reconocimiento de escritura, además de reconocimiento del habla.

- Reconocimiento directo sobre vectores de características (sin derivadas ni aceleración). Es decir, con cálculo de derivadas y aceleración opcional.

- N-gramas como modelos de lenguaje, además de los autómatas de estados finitos.

- Un solo factor de poda para llevar a cabo ’beam search’.

- Los parámetros de búsqueda son configurables mediante un fichero de configuración.

[1]iATROS se encuentra en desarrollo y esta poda ha sido sustituída, pero los experimentos de este trabajo han sido llevados a cabo con este tipo de poda.

- Se puede obligar a reconocer un símbolo inicial y final en la decodificación, algo muy interesante en habla para trabajar con los silencios y ruidos del principio y final de las grabaciones.
- Se generan grafos de palabras. Estos son configurables con dos parámetros: el número de aristas que llegan a los estados y el número máximo de mejores hipótesis desde las que se genera el grafo.[2]
- Se puede utilizar un factor de poda para las transiciones del modelo de lenguaje, porque los n-gramas pueden expandir un número muy elevado de hipótesis.
- Es posible añadir un símbolo en el fichero de configuración (con una cierta probabilidad de transición) que se añade siempre en las transiciones de modelo de lenguaje. Por ejemplo, en habla se puede utilizar el silencio y en escritura el espacio en blanco.
- Es posible obtener una decodificación completa o parcial.

[2]Una versión no estándar de Atros también genera grafos de palabras. Dicha versión ha sido empleada en los experimentos de la Sección 6.2.

CAPÍTULO 3

RECONOCIMIENTO MULTILINGÜE

Los sistemas de reconocimiento automático del habla multilingüe son de gran interés en entornos multilingües. Por un lado están los entornos donde se concentran personas de muy diversos ámbitos lingüísticos, como congresos internacionales, o lugares donde conviven dos lenguas oficiales (como por ejemplo la Comunidad Valenciana, donde conviven castellano y valenciano). Además, está la gran influencia que ofrecen los inmigrantes por su acento, ya que cada lenguaje define sus propios fonemas y la articulación de un mismo fonema en diferentes lenguas puede diferir en cada lengua. Por lo tanto, el multilingüismo implica que un sistema de RAH tiene que solucionar el problema de reconocer varias lenguas y diferentes acentos en un mismo entorno.

Los sistemas multilingües deben trabajar tanto en el modelado de lenguaje como en los modelos acústicos. Por lo tanto, para construir un sistema multilingüe es necesario trabajar principalmente en:

- Modelos de lenguaje: los sistemas multilingües deben incluir tantos modelos de lenguaje como lenguas puedan soportar; podemos construir modelos de lenguaje que intercalen varias lenguas en la decodificación o que la decodificación sea en una sola lengua.

- Modelos acústicos: teniendo en cuenta que cada lengua define sus propios fonemas (que pueden ser diferentes a los de otras lenguas) y que se produce una gran influencia de unas lenguas con otras, es necesario entrenar modelos acústicos para cada idioma o adaptarlos a partir de modelos acústicos de un idioma similar.

- Identificación del idioma: el sistema debe ser capaz de determinar la lengua que está siendo hablada; esto suele usarse para determinar qué modelo de lenguaje se tiene que utilizar o para emplear la lengua correspondiente en el caso de que se deba responder al usuario.

En la actualidad se está mostrando mucho interés por las lenguas minoritarias, ya que son lenguas con muy pocos recursos lingüísticos (corpora) debido a su menor interés económico. Sin embargo, haciendo un buen estudio de ellas y empleando las técnicas de modelado multilingüe se pueden obtener buenos modelos y sistemas de reconocimiento que obtengan unas tasas de acierto competitivas para dichas lenguas, aunque se tengan pocos recursos de las mismas.

En las siguientes secciones se explicarán las diferentes formas de construir un sistema multilingüe, una técnica para llevar a cabo identificación del lenguaje, el uso de grafos de palabras para mejorar el error a nivel de palabra en reconocimiento multilingüe y técnicas de adaptación para la obtención de modelos acústicos.

3.1. Sistemas multilingües

La construcción de sistemas de reconocimiento multilingüe se basa en la teoría sobre reconocimiento de habla detallada en el Capítulo 1. Además, la construcción de sistemas multilingües lleva asociado una serie de problemas, que se pueden dar desde los modelos acústicos hasta los modelos de lenguaje.

Cuando se diseña un sistema de reconocimiento de habla de un idioma en particular lo llamamos reconocedor monolingüe. Los modelos acústicos para este sistema se entrenan con datos del único idioma que se va a reconocer en el sistema.

Los reconocedores multilingües emplean modelos acústicos dependientes del idioma (un conjunto para cada idioma) o una combinación de modelos acústicos independientes del idioma (que fueron entrenados con datos compartidos de varios idiomas). El problema es que son necesarias grandes cantidades de datos para entrenar los modelos acústicos, y no siempre es posible encontrar corpus multilingües de todas las lenguas del mundo. Esto se debe a que el coste de obtener grandes corpora es elevado y el criterio para decidir qué idiomas son adquiridos cambia constantemente por razones económicas y políticas.

El desarrollo de un reconocedor multilingüe aplicando modelos acústicos multilingües conlleva muchos problemas. Una razón puede derivar de la naturaleza de la tarea, ya que según la misma será más fácil o más difícil unir el conocimiento de los múltiples idiomas. Según la similitud de los idiomas empleados acaba siendo necesario que un solo experto tenga que familiarizarse con todos los idiomas del sistema.

Además, dichos sistemas se enfrentan con usuarios que pueden ser nativos en alguna de las lenguas del sistema, pero puede darse el caso de que el usuario no sea nativo en ninguna de las lenguas del sistema, lo que provocará que el acento y la forma de pronunciar los fonemas que no ha aprendido en su lengua materna

dificulte el reconocimiento. Una de las posibilidades de mejora es permitir que ciertos fonemas puedan sustituirse por los más semejantes de otras lenguas. Esto significa que los modelos léxicos deberán contener múltiples pronunciaciones por cada palabra y para obtenerlas es necesario emplear grandes cantidades de datos de cada idioma.

Otro problema importante que se da en los sistemas multilingües es la utilización de los modelos de lenguaje, ya que se pueden utilizar los modelos de lenguaje independientes para cada lengua o utilizar un solo modelo de lenguaje para todas las lenguas. Esto se asocia a otro problema: determinar el idioma del usuario que está utilizando el sistema.

Por lo tanto, para construir un sistema multilingüe es necesario contar con un corpus de datos grande. Con ese corpus se puede llevar a cabo una gran experimentación que determine cuál es la mejor opción de todas las posibles. Es decir, se deben experimentar todas las combinaciones que se pueden dar en un sistema multilingüe respecto a los modelos acústicos, léxicos, de lenguaje y a la identificación del lenguaje.

Los sistemas multilingües se pueden construir de diferentes formas, dependiendo de los modelos de lenguaje y los modelos acústicos disponibles y su forma de uso:

- *Identificación del lenguaje primero.* Si se lleva a cabo primero una identificación del lenguaje (en el sentido de idioma) nos encontramos en un problema clásico de reconocimiento de habla, ya que con el idioma identificado podemos hacer uso del reconocedor monolingüe del idioma reconocido. La problemática de este caso es que si el idioma identificado no es el idioma correcto la frase será completamente mal reconocida, porque todas las palabras se reconocerán en un idioma incorrecto. Por supuesto, cuantos más idiomas puedan ser reconocidos más opciones de que el idioma se identifique incorrectamente.

  Para llevar a cabo la identificación del lenguaje se puede preguntar al usuario el idioma al inicio o se pueden emplear técnicas más avanzadas de reconocimiento de formas, como técnicas biométricas (que se estudiarán con más detalle en la Sección 3.2).

- *Se comparten sólo las unidades acústicas entre los lenguajes.* Para obtener modelos acústicos compartidos se usan todas las frases de entrenamiento disponibles de todas las lenguas. Es necesario que en el conjunto de modelos acústicos se encuentren todos los fonemas que hay en las lenguas, incluso los fonemas que se comparten. Los modelos de lenguaje están separados, es decir, utilizamos un modelo de lenguaje para cada lengua sin que sea posible pasar de un idioma a otro durante el reconocimiento. El problema es el coste temporal, ya que hay que llevar a cabo un reconocimiento en paralelo, es decir, hay que hacer un reconocimiento por cada idioma admitido.

- *Se comparten los modelos acústicos y modelos de lenguaje.* Por último, se comparten los modelos acústicos y el modelo de lenguaje, ya que el modelo de lenguaje permite transiciones entre los diferentes idiomas. Esto puede justificarse en el caso en que, por ejemplo, los nombres propios de una frase se digan en el idioma origen del nombre, a pesar de que la frase sea pronunciada en un idioma diferente.

  Para mejorar errores producidos por este tipo de modelo de lenguaje se pueden utilizar técnicas basadas en grafos de palabras (que se estudiarán con más detalle en la Sección 3.3).

- *Modelos acústicos adaptados para una lengua minoritaria desde una lengua mayoritaria.* Algunos estudios [20, 19] han demostrado que es posible mejorar un modelo acústico ya entrenado mediante un conjunto de datos reducidos, adaptando los parámetros de los modelos a dichos datos. Aunque esto se pensó inicialmente para adaptar el modelo a un locutor concreto, algunos autores [2, 31] han propuesto utilizarlo para adaptar los modelos acústicos de una cierta lengua a otros idiomas con características fonéticas similares.

  En [31] se muestra cómo para un conjunto de muestras reducido los modelos adaptados se comportan mucho mejor que los modelos de la lengua original. Sin embargo, si existen suficientes datos, los modelos de la lengua de interés entrenados con dichos datos sobrepasan claramente en rendimiento a los modelos adaptados. Por lo tanto, esta es una opción a tener en cuenta únicamente si se posee un conjunto reducido de datos de entrenamiento.

  Por lo tanto, en el caso de lenguas minoritarias semejantes a una mayoritaria, tenemos la opción de adaptar unos modelos acústicos entrenados con un corpus grande en la lengua mayoritaria (aunque la tarea del corpus no sea la misma que la tarea para la que se van a utilizar los modelos acústicos adaptados) con datos de la lengua minoritaria de la tarea.

  En la Sección 3.4 se estudiarán con más detalle las técnicas de adaptación implementadas para ser empleadas en la experimentación.

3.2. Identificación Del Lenguaje (IDL) basada en características de la señal

La identificación del lenguaje (IDL), entendido como lengua o idioma, tiene como objetivo conocer el idioma con el que el usuario interacciona con el sistema. Esto es necesario para que el sistema pueda comunicarse con el usuario en el idioma adecuado. Además, en el caso de que el sistema cargue un modelo de lenguaje con sólo un idioma, el sistema debe conocer (sin llevar a cabo un reconocimiento) el idioma en el que el usuario está interaccionando. Para este

trabajo, hemos estudiado técnicas de IDL biométricas basadas en características locales y k-vecinos.

3.2.1. Preproceso y extracción de características

Para llevar a cabo una identificación del idioma en una señal acústica es necesario preprocesar la señal. En primer lugar, la señal acústica se convierte en vectores de cepstrales como se puede ver en la Sección 2.2. El número de coeficientes para cada vector de cepstrales que se emplean puede variar, aunque normalmente en habla se utilizan 11 (incluyendo la energía). Además, en habla se suele calcular tanto la primera como la segunda derivada de los vectores de cepstrales.

En la Figura 3.1 se muestra un ejemplo de una señal acústica. En la parte superior (a) se puede ver la señal de habla de un sonido representado en el dominio del tiempo. En la parte central (b) se puede ver el correspondiente espectrograma, donde la frecuencia se representa en el eje vertical y el tiempo en el eje horizontal. Al final de la figura (c) se puede ver una representación de los coeficientes cepstrales con un pequeño esquema explicativo del proceso que hemos llevado a cabo y que se detalla a continuación.

Con la señal acústica convertida en vectores de cepstrales (parte inferior de la Figura 3.1) podemos aplicar un preproceso que consta de dos pasos:

- *Primer paso:* consiste en seleccionar aquellas partes de la señal que tienen la información más relevante. Para ello, calculamos la varianza local de pequeñas ventanas del espectro, y seleccionamos aquellas que tienen una mayor varianza local. Es recomendable probar diferentes tamaños de ventana a fin de obtener el que dé mejores prestaciones. El vector de características se obtiene concatenando los vectores de cepstrales de la ventana seleccionada.

- *Segundo paso*: consiste en reducir la dimensionalidad de los vectores obtenidos usando 'Principal component analysis' (PCA) [15]. De esta forma, obtenemos la información de una región local de la señal compactada.

El proceso completo se muestra en la Figura 3.1. Para concluir el preproceso de los mismos tenemos que etiquetar los datos con el idioma correspondiente. Por lo tanto, cada secuencia de entrenamiento queda representada por un conjunto de vectores locales y la etiqueta del idioma.

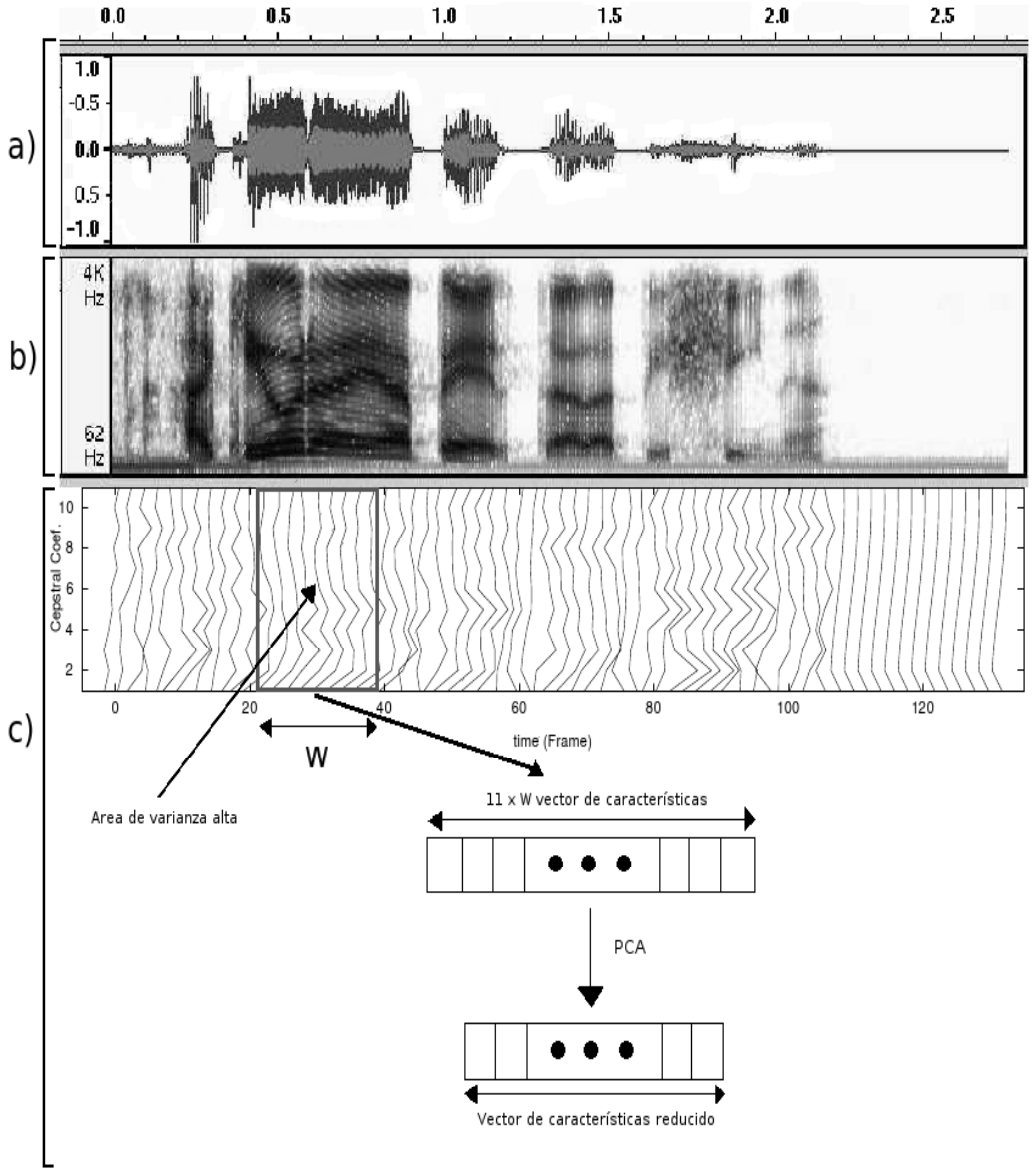

Figura 3.1: Parte superior (a): señal de habla de un sonido representado en el dominio del tiempo. Parte central (b): el correspondiente espectrograma; la frecuencia se representa en el eje vertical y el tiempo en el eje horizontal. Parte inferior (c): una representación de los coeficientes cepstrales y esquema del proceso de obtención de los vectores de características reducidos.

3.2.2. Esquema de clasificación: k-vecinos

El esquema de clasificación se basa en la técnica de k-vecinos más cercanos (K-NN) [16]. La idea es que, para cada característica local de test, se deben buscar los k-vecinos de entre todas las muestras de entrenamiento, y cada vecino debe votar a la clase a la que pertenece. Por lo tanto, la clase ganadora para la muestra de test es la clase con un mayor número de votos. Matemáticamente, el modelo de combinación de hipótesis locales se puede expresar como:

$$p(c|X) = \frac{1}{n}\sum_{i=1}^{n} p(c|x_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{k_{ic}}{K}$$

donde X es la señal acústica de test que puede ser clasificada en una clase c. Aquí n es el número de vectores de características que hemos obtenido en el preproceso, K el número de vecinos y k_{ic} el número de votos que tiene el vector de características i para pertenecer a la clase c. Una explicación más detallada de este clasificador puede consultarse en [27].

3.3. Grafos de palabras en reconocimiento multilingüe

Cuando se emplea un modelo de lenguaje mixto con los idiomas de la tarea, podemos obtener decodificaciones con una mezcla de palabras en ambos idiomas. Esto supone un aumento del error, ya que las frases se deben decodificar en un solo idioma. Por esta razón, pensamos utilizar grafos de palabras para obtener la frase en un solo idioma para intentar mejorar el error. Los grafos de palabras tienen información en una sola lengua de la frase decodificada. Por lo tanto, podemos obtener un grafo de palabras por cada lengua y por cada frase decodificada y decodificar la frase en un solo idioma.

Un grafo de palabras G es un grafo dirigido, acícilico y con pesos. Los nodos del grafo corresponden a puntos discretos en el tiempo. Las aristas del grafo son tripletas $[w, s, e]$ donde w es la palabra hipotética del nodo s al nodo e. Los pesos son puntuaciones asociadas a las aristas del grafo de palabras. El mejor camino que se forma desde el estado inicial hasta el estado final es la hipótesis más probable [30].

Los grafos de palabras con los que vamos a trabajar cuentan con diferentes probabilidades o puntuaciones (logaritmos de la probabilidad) en las aristas:

- Acústica: los pesos de las aristas son puntuaciones acústicas.
- Modelo de lenguaje: los pesos de las aristas son las probabilidades en el modelo de lenguaje.

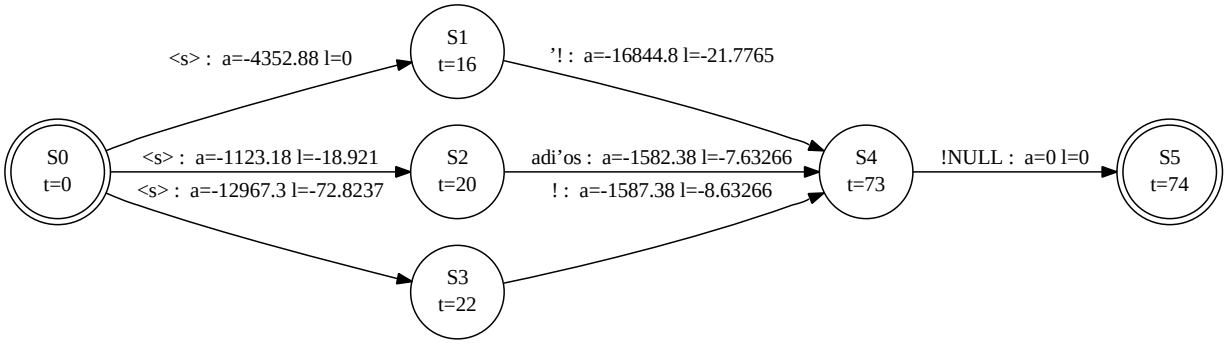

Figura 3.2: Ejemplo de grafo de palabras, donde a indica puntuación acústica y l indica puntuación de modelo de lenguaje.

- Total: los pesos de las aristas son una combinación de las puntuaciones acústicas y las del modelo de lenguaje.

Un ejemplo de un grafo de palabras puede verse en la Figura 3.2.

La problemática de utilizar grafos de palabras radica en aquellas frases que en el reconocimiento inicial fallan y no se decodifica nada, ya que dichas frases no tienen grafo de palabras. Por lo tanto, no se puede aplicar ninguna técnica que emplee los grafos de palabras para mejorar el reconocimiento en dichas frases.

Las probabilidades de los grafos de palabras que más nos interesan son:

- Probabilidad del modelo de lenguaje: es la probabilidad que tenía esa palabra en el modelo de lenguaje empleado para el reconocimiento.
- Probabilidad de reconocimiento: es la probabilidad acústica más la probabilidad del modelo de lenguaje por el GSF más el WIP (capítulo 2).

Empleando las distintas probabilidades de los grafos de palabras llevamos a cabo diferentes experimentos:

- *Segundo reconocimiento*: Utilizamos la probabilidad del modelo de lenguaje que aporta el grafo de palabras del idioma reconocido para construir un nuevo modelo de lenguaje empleado para hacer un segundo reconocimiento.
- *Viterbi monolingüe*: Utilizamos la probabilidad de reconocimiento que aporta el grafo de palabras del idioma reconocido para obtener la frase más probable del grafo de palabras.
- *Viterbi bilingüe*: Aplicamos el algoritmo de Viterbi al grafo de palabras después del reconocimiento. Obtenemos una frase para cada idioma y seleccionamos la frase más probable.

En los dos primeros casos, el reconocimiento o búsqueda se hace restringiendo a las palabras del idioma que se ha determinado como correcto.

3.4. Técnicas de adaptación al idioma: MLLR

El propósito de las técnicas de adaptación al locutor es obtener un sistema de reconocimiento dependiente del locutor. El sistema se obtiene a partir de unos modelos ocultos de Markov entrenados de manera independiente del locutor y con gran cantidad de datos. Dichos modelos se adaptan usando datos del locutor con la intención de conseguir modelos acústicos con un mejor rendimiento para ese locutor. Esto es equivalente a obtener un sistema dependiente del locutor, pero esta última opción suele requerir una cantidad de datos excesiva de dicho locutor. En cambio, la adaptación de habla permite obtener un sistema dependiente del locutor usando una cantidad limitada de datos.

MLLR (Maximum Likelihood Linear Regression) [21] es una técnica para adaptar un conjunto de modelos acústicos independiente del hablante usando una cantidad de datos de adaptación pequeña. Esta técnica puede ser usada también para hacer una adaptación al idioma usando material de adaptación en el idioma al que va a ser adaptado.

En adaptación al idioma, MLLR requiere un conjunto de HMM inicial del idioma original y datos de adaptación del idioma de la aplicación para adaptar los modelos acústicos. MLLR actualiza el parámetro media en el caso gaussiano (pero también puede adaptar varianzas) para maximizar la verosimilitud de los datos de adaptación. Las medias son actualizadas usando una matriz de transformación, que es estimada con los datos de adaptación. Nosotros hemos aplicado la formulación presentada en [21] para adaptación del idioma: utilizamos datos de adaptación de un idioma para adaptar los modelos acústicos usando MLLR de la misma forma que se hace en adaptación del locutor.

La teoría se basa en el concepto de clases de regresión. Una clase de regresión es un conjunto de mixturas (densidades de probabilidad de emisión) que comparten la misma matriz de transformación. Cuando las mixturas de todos los modelos están en la misma clase de regresión, tenemos una clase de regresión global. Sin embargo, cualquier conjunto de clases de regresión puede ser manual o automáticamente definido con las gaussianas del conjunto de HMM. No existe un método para determinar analíticamente el número óptimo de clases de regresión y las gaussianas que deben pertenecer a cada clase de regresión.

Para llevar a cabo la adaptación de las medias, se calcula una matriz de transformación $\vec{W}$ para cada clase de regresión. Esta matriz es aplicada sobre el vector media de todas las mixturas pertenecientes a la clase de regresión para obtener un vector media adaptado.

El vector media adaptado $\vec{\widehat{\mu}}_{qi}$ se obtiene como:

$$\vec{\widehat{\mu}}_{qi} = \vec{W} \cdot \vec{\xi}_{qi}$$

donde $\vec{\widehat{\mu}}_{qi}$ es el vector media adaptado y $\vec{\xi}_{qi}$ es el vector media extendido definido como:

$$\vec{\xi}_{qi} = [w, \mu_{qi}^0, \ldots, \mu_{qi}^{(D-1)}]' = [w : \vec{\mu}_{qi}]$$

donde D es el número de características, $\vec{\mu}_{qi}$ es el vector media original y w es un término de compensación.

Si tenemos un conjunto de datos de adaptación denotado por la secuencia de vectores de características acústicos $\vec{X} = \vec{x}_1\vec{x}_2...\vec{x}_T, \vec{x}_t \in \mathbb{R}^D, t = 1, ..., T$, podemos estimar la matriz de adaptación $\vec{\widehat{W}}$ usando el enfoque de máxima verosimilitud como:

$$\vec{\widehat{W}} = \max_{\vec{W}} p_{\vec{\theta}}(\vec{X})$$

donde $\vec{\theta}$ define los parámetros del modelo adaptado.

Para calcular la matriz de transformación podemos usar diferentes variantes: asumiendo matrices de covarianzas distintas para cada gaussiana (con una matriz de adaptación completa o diagonal) o asumiendo las mismas covarianzas en todas las distribuciones (mínimos cuadrados). Los detalles de la estimación de estas variantes se pueden consultar en [21]. La siguiente formulación asume sólo una muestra de adaptación ($\vec{X}$), pero se puede extender fácilmente a n muestras de adaptación.

3.4.1. Matriz completa

Dado un estado q de un HMM, y la i-ésima gaussiana de su distribución de salida, denotamos su vector media como $\vec{\mu}_{qi}$ y su matriz de covarianzas como $\vec{\Sigma}_{qi}$.

Para calcular una matriz de adaptación completa, es necesario calcular una matriz auxiliar tridimensional $\vec{G}$. En este caso, $\vec{\widehat{W}}$ debe ser calculada por filas porque $\vec{G}$ es una matriz tridimensional. Calculamos la fila k de $\vec{\widehat{W}}$ como:

$$\vec{\widehat{w}}_k' = \vec{G}^{(k)-1}\vec{z}_k'$$

donde:

$$\vec{z} = \sum_t \sum_q \sum_i (\gamma_{qi}(t))\vec{\Sigma}_{qi}^{-1}\vec{x}_t\vec{\xi}_{qi}'$$

$\gamma_{qi}(t)$ se define como la probabilidad a posteriori de ocupación en un estado q en tiempo t dado que la secuencia de observación $\vec{x}_t$ es generada en la i-ésima gaussiana.

La fila k de $\vec{G}$ se define como:

$$\vec{G}_{jq}^{(k)} = \sum_{q,i} \vec{v}_{ii}^{(qi)} \vec{d}_{jq}^{(qi)}$$

donde:

$$\vec{D}^{(qi)} = \vec{\xi}_{qi} \vec{\xi}'_{qi}$$

y

$$\vec{V}^{(qi)} = \sum_{t} \gamma_{qi}(t) \vec{\Sigma}_{qi}^{-1}$$

3.4.2. Matriz diagonal

Para calcular una matriz de transformación $\vec{\widehat{W}}$ restringida a que sea diagonal, definimos una matriz diagonal:

$$\vec{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & 0 & \cdots & 0 \\ w_{2,1} & 0 & w_{2,3} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ w_{n,1} & 0 & \cdots & 0 & w_{n,n+1} \end{bmatrix}$$

Obviando los elementos distintos de cero, podemos reescribir la matriz a un vector de transformación $\vec{w}$ como:

$$\vec{w} = \begin{bmatrix} w_{1,1} \\ \vdots \\ w_{n,1} \\ w_{1,2} \\ \vdots \\ w_{n,n+1} \end{bmatrix}$$

Definimos una matriz $\vec{D}_{qi}$ compuesta de elementos del vector media extendido $\vec{\xi}_{qi}$ como:

$$\vec{D}_{qi} = \begin{bmatrix} w & 0 & \cdots & 0 & \mu_{qi1} & 0 & \cdots & 0 \\ 0 & w & \ddots & \vdots & 0 & \mu_{qi2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w & 0 & \cdots & 0 & \mu_{qin} \end{bmatrix}$$

Entonces, $\vec{\widehat{w}}$ puede ser calculada como:

$$\vec{\widehat{w}} = \left[\sum_t \sum_q \sum_i (\gamma_{qi}(t)) \vec{D}'_{qi} \vec{\Sigma}_{qi}^{-1} \vec{x}_t \right]^{-1} \left[\sum_t \sum_q \sum_i (\gamma_{qi}(t)) \vec{D}'_{qi} \vec{\Sigma}_{qi}^{-1} \vec{D}_{qi} \right]$$

3.4.3. Mínimos cuadrados

En este caso consideramos que todas las covarianzas de la distribución son la misma. Así es posible seguir la aproximación de Viterbi, donde cada vector de características es asignado exactamente a una distribución.

En este caso, la matriz de adaptación puede ser calculada como:

$$\vec{\widehat{W}} = \left(\sum_t \vec{x}_t \vec{\xi}'_{qi} \right) \left(\sum_t \vec{\xi}_{qi} \vec{\xi}'_{qi} \right)^{-1}$$

La secuencia de $\vec{\xi}_{qi}$ se define a partir de un alineamiento de Viterbi de las muestras, es decir, es la media extendida de la gaussiana que ha dado la máxima probabilidad para el vector de características correspondiente, en un proceso de decodificación por Viterbi.

CAPÍTULO 4

UN SISTEMA ESPECÍFICO DE RECONOCIMIENTO MULTILINGÜE

4.1. Tarea

Después de definir las bases teóricas del reconocimiento del habla en un entorno multilingüe vamos a definir la tarea concreta que se ha llevado a cabo en este trabajo. En la Comunidad Valenciana hay dos lenguas oficiales, castellano y valenciano. Estas dos lenguas comparten muchos de sus fonemas y su sintaxis y vocabulario es similar porque se han influenciado mutuamente a lo largo de los años. Disponemos de un corpus con las dos lenguas de la misma tarea. En este trabajo pretendemos estudiar un caso concreto de bilingüismo, el de la Comunidad Valenciana, para construir un sistema de reconocimiento de habla bilingüe en castellano y valenciano.

4.2. Corpus

El corpus empleado para llevar a cabo toda la experimentación del trabajo es un corpus sobre un sistema de información que fue adquirido por la línea telefónica. Consta de 4 horas de grabación (2 horas para cada idioma) y consta de 120 frases distintas (60 para cada idioma) de cada uno de los 20 locutores del proceso de adquisición. El corpus está dividido en 5 grupos con hombres y mujeres castellano y valenciano parlantes hablando en castellano y valenciano. Por lo tanto, consta de 4 tipos para cada lengua (hombres y mujeres castellano y valenciano parlantes hablando en esa lengua). Las frases de entrenamiento corresponden a los grupos comprendidos entre el 1 y el 4, y las frases de test son las correspondientes al grupo 5. La mitad de los usuarios que grabaron frases eran castellano-parlantes y la otra mitad valenciano-parlantes, estando también equilibrada la proporción de sexos en

cada lengua. Todos los hablantes eran estudiantes universitarios y grabaron frases en ambas lenguas. La distribución de hombres y mujeres era igual.

La Tabla 4.1 resume las estadísticas más importantes del corpus. Para más detalle se puede encontrar una descripción completa del corpus en [3]. No hay palabras fuera de vocabulario en la parte de test de castellano y sólo 2 en la parte de test de valenciano.

		Castellano	Valenciano
Entrenamiento	Frases	240	240
	Palabras	2.887	2.692
	Duración	1 h 33 m	1 h 29 m
	Vocabulario	131	131
Test	Frases	60	60
	Palabras	705	681
	Duración	23m	21m

Tabla 4.1: Estadísticas del corpus.

4.3. Modelos de lenguaje

Los modelos de lenguaje definen qué tipo de frases son permitidas en el sistema. Por lo tanto, nuestro modelo de lenguaje debe aceptar todas las frases de la parte de entrenamiento del corpus.

Las frases del corpus se han generado de forma semi-automática mediante una plantilla. Por tanto, todas las frases del corpus tienen una estructura común en bloques de palabras: saludo, pregunta, información, título, persona y despedida. Los campos información y persona tienen el contenido semántico, siendo estos los más importantes en la tarea. Algunas frases de ejemplo se pueden ver en la Figura 4.1. No todos los campos son requeridos en las frases, y cada bloque modela sus subfrases con un autómata de estados finitos. De acuerdo con esta idea, para obtener el modelo de lenguaje final construímos un autómata combinando los autómatas correspondientes a cada bloque.

Para este trabajo desarrollamos tres modelos de lenguaje:

- Dos modelos de lenguaje separados: construímos un modelo de lenguaje monolingüe para cada lengua usando un autómata aceptor (en la correspondiente lengua) para cada idioma. El modelo de lenguaje fue construído juntando los autómatas aceptores de cada bloque en serie: por cada dos autómatas consecutivos, unimos el estado final del primer autómata aceptor con el estado

Castellano

- Por favor, quiero saber el e-mail de Francisco Casacuberta, adiós.
- Hola, ¿cuál es el horario de consultas de Enrique Vidal? Muchas gracias.
- Buenas noches, quería la extensión de la señorita Silvia Abrahao, muchas gracias.

Valenciano

- Per favor, vull saber l'e-mail de Francisco Casacuberta, adeu.
- Hola, quin és l'horari de consultes d'Enrique Vidal? Moltes gràcies.
- Bona nit, volia saber l'extensió de la senyoreta Silvia Abrahao, moltes gràcies.

Figura 4.1: Una selección de frases del corpus.

inicial del segundo autómata aceptor. Este proceso de serialización se puede ver en la Figura 4.2. Utilizar estos modelos de lenguaje equivale a hacer primero una identificación del lenguaje sin errores.

- Un modelo de lenguaje mixto: un solo autómata fue construído para poder aceptar las dos lenguas a la vez. Este autómata fue construído a partir de los dos modelos de lenguaje separados, el autómata aceptor de cada lengua en paralelo, juntando el estado inicial y el estado final para cada bloque. La Figura 4.3 muestra un ejemplo del proceso de paralelización. El autómata aceptor también fue unido en serie. La Figura 4.4 muestra un ejemplo del proceso de paralelización más serialización para unir autómatas.

Respecto a la perplejidad de los modelos de lenguaje, para el modelo de lenguaje de sólo español es de 5.98, para el modelo de lenguaje de sólo valenciano es de 6.46 y para el modelo de lenguaje mixto es de 8.42. La perplejidad de los modelos es muy baja; esto está de acuerdo con el tamaño del corpus, que también es pequeño.

4.4. Modelos acústicos

Cada modelo acústico se asocia a un sonido (por ejemplo, fonemas) para hacer una comparación de las características de una secuencia acústica con los

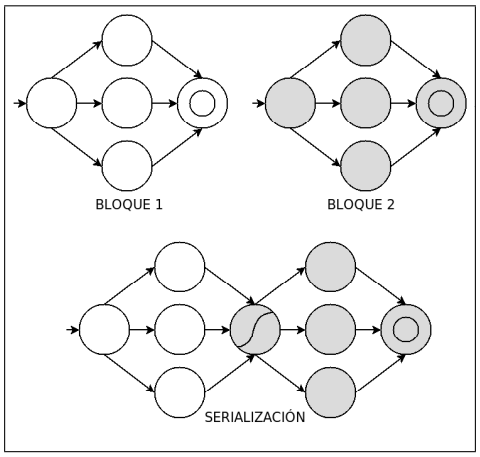

Figura 4.2: Ilustración del proceso de serialización.

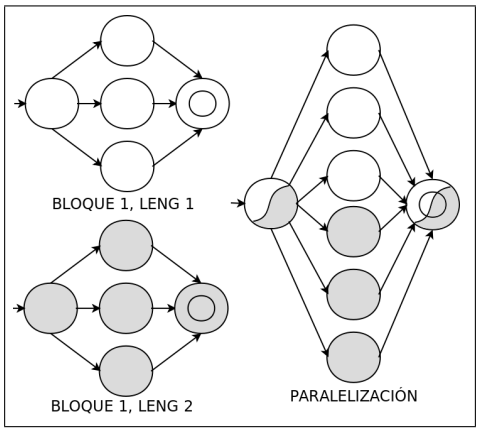

Figura 4.3: Ilustración del proceso de paralelización.

modelos acústicos. Los modelos acústicos son Modelos Ocultos de Markov (HMM) que entrenamos usando el *toolkit* HTK [40]. Los HMMs siguen una topología de izquierda a derecha sin saltos. Probamos desde una gaussiana por estado hasta 128 gaussianas por estado obteniéndose los mejores resultados con 32 gaussianas. Cada gaussiana se modela con un vector de características de 33 componentes (10 coeficientes cepstrales más energía con la primera y la segunda derivada).

En nuestro caso, cada HMM modela un solo fonema sin contexto (monofonemas). Se usaron diferentes conjuntos de modelos acústicos para llevar a cabo los experimentos:

- Modelos acústicos sólo en castellano: estos modelos fueron obtenidos con el conjunto fonético de castellano del corpus descrito en la Sección 4.2. Las transcripciones se hicieron de forma automática siguiendo las reglas descritas

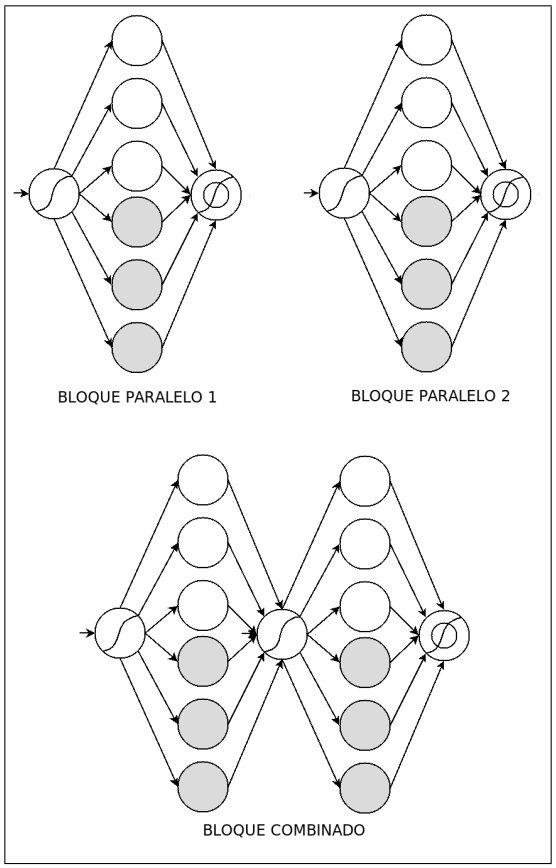

Figura 4.4: Ilustración de la combinación de los procesos de paralelización y serialización.

en [28] por el alfabeto fonético SAMPA [34]. Estos modelos acústicos fueron entrenados sólo con sonidos de palabras en castellano.

- Modelos acústicos sólo en valenciano: estos modelos fueron obtenidos con el conjunto fonético de valenciano del corpus descrito en la Sección 4.2. Los modelos acústicos de castellano y los modelos acústicos de valenciano comparten algunos fonemas. Sin embargo, en valenciano no disponemos de reglas que nos ayuden a transcribir las frases automáticamente. Por lo tanto, cada palabra del vocabulario fue transcrita con variaciones fonéticas conocidas. Este conjunto fue entrenado sólo con sonidos de palabras en valenciano.
- Modelos acústicos de castellano y valenciano mixtos: para obtener estos modelos acústicos, utilizamos todas las frases de entrenamiento que teníamos disponibles (tanto en castellano como en valenciano). Los modelos acústicos mixtos fueron obtenidos usando todos los fonemas de castellano y valenciano porque muchos de ellos se comparten en las dos lenguas. Para los fonemas

compartidos en castellano y valenciano usamos el mismo modelo acústico. Las transcripciones se llevaron a cabo con la información que disponíamos para castellano y valenciano.

- Modelos acústicos adaptados: para obtener modelos acústicos fiables, estos se deben entrenar con grandes cantidades de datos de entrenamiento. Para nuestro caso es fácil encontrar grandes corpora en castellano, pero no hemos encontrado un gran corpus en valenciano. Sin embargo, como el valenciano es muy similar al castellano fonéticamente, pensamos en usar la pequeña cantidad de datos de los que disponemos en valenciano para adaptar unos modelos acústicos que hubieran sido entrenados con un gran corpus en castellano. Para ello usamos como corpus de entrenamiento de los modelos iniciales el corpus *Senglar*, que ha sido usado con éxito en otras tareas [5]. Las condiciones de grabación de dicho corpus son diferentes de las condiciones de grabación del corpus que estamos empleando para llevar a cabo los experimentos. Por esta razón, obtuvimos modelos acústicos adaptando los modelos acústicos del corpus *Senglar* para castellano y valenciano, para que pudieran ser utilizados en nuestra tarea.

  Los modelos acústicos adaptados fueron obtenidos con la técnica MLLR, estimando una matriz global para cada lengua (sólo una clase de regresión). Usamos el corpus de entrenamiento como material de adaptación para obtener los modelos acústicos adaptados. Esta técnica ha sido usada antes para reconocimiento del habla multilingüe, y la cantidad de señal que utilizamos para adaptar los modelos es similar a la cantidad de señal que ha sido empleada en trabajos anteriores [32, 41]. La descripción de la técnica MLLR puede verse más detallada en la Sección 3.4.

CAPÍTULO 5

RESULTADOS EXPERIMENTALES

En esta sección se presentan los resultados obtenidos en el marco experimental definido. El objetivo de estos experimentos es evaluar las diferentes aproximaciones tanto monolingües como multilingües. La finalidad es encontrar la opción que obtiene los mejores resultados para el conjunto de test y estudiar el comportamiento del sistema multilingüe castellano-valenciano.

5.1. Medidas de evaluación

Se han tomado tres medidas de evaluación para medir aquellos aspectos de interés para el trabajo:

- *Tasa de error.* La tasa de error a nivel de palabra o Word Error Rate (WER) es una medida de la calidad del reconocimiento respecto a una referencia dada. Calcula la distancia de edición entre la frase reconocida por el sistema y la frase de referencia. Para ello las muestras de prueba deberán tener la transcripción correcta.

  La distancia de edición mide el número de sustituciones, inserciones y borrados que se deben hacer en la frase reconocida para convertirla en la frase de referencia. El WER del conjunto de prueba es el promedio de errores observados para las frases individuales.

- *Tasa de error semántico.* La tasa de error semántico (SemWER) es una medida de la calidad de reconocimiento de los campos semánticos. En este caso se considerará la etiqueta de información (por ejemplo, el tipo de mensaje y el nombre). Sobre esta frase semántica se realizará la distancia de edición, como en el caso de la tasa de error.

- *Tasa de identificación.* En entornos multilingües suele ser necesario saber la lengua del locutor para que, por ejemplo, el sistema pueda responder al locutor en la misma lengua. Por esta razón, cuando ambas lenguas pueden ser decodificadas, se computa el porcentaje de frases en que la lengua del locutor fue identificada correctamente.

Estas medidas se evalúan con todos los locutores de test. Sin embargo, necesitamos saber quién está cometiendo los errores en cada idioma para determinar la influencia de la lengua materna. Por esta razón, obtenemos los errores para la lengua materna como ”Locutor nativo”. De esta manera, podemos saber si todos los locutores del corpus pueden hablar las dos lenguas del sistema con la misma competencia.

5.2. Resultados con el reconocedor ATROS

En esta sección explicamos los experimentos llevados a cabo en el marco definido. Un primer intento se llevó a cabo en un trabajo previo [3], pero los resultados no fueron definitivos. En este caso, hemos llevado a cabo una experimentación más exhaustiva. El proposito de los experimentos es encontrar la opción que obtenga los mejores resultados para el conjunto de test y para estudiar el comportamiento de un sistema multilingüe castellano-valenciano.

Estos experimentos fueron llevados a cabo con modelos acústicos mixtos y separados. Los modelos acústicos mixtos fueron entrenados con todo el material de entrenamiento disponible (en ambas lenguas) y los modelos separados fueron entrenados con las frases de entrenamiento de la lengua correspondiente.

5.2.1. Modelos de lenguaje separados

Separamos los resultados respecto a si utilizamos modelos de lenguaje separados para cada idioma o un solo modelo de lenguaje para los dos idiomas. La primera opción equivale a hacer una identificación correcta del idioma antes de reconocer (para utilizar el modelo de lenguaje del idioma reconocido), mientras que la segunda implica hacer el reconocimiento con un modelo de lenguaje que soporta ambas lenguas, identificando después la lengua a partir del resultado del reconocimiento.

Los experimentos de este apartado se refieren a la primera opción.

Mod. Acústicos		No mixtos		Mixtos	
Mod. Lenguaje		Cast	Val	Cast	Val
WER		3.5 %	5.4 %	3.1 %	6.2 %
SemWER		1.7 %	3.8 %	1.7 %	3.6 %
Locutor Nativo	Cast	4.7 %	7.3 %	4.4 %	8.1 %
	Val	2.2 %	3.4 %	1.9 %	4.4 %

Tabla 5.1: Resultados para modelos de lenguaje separados con modelos acústicos no adaptados.

Modelos acústicos no adaptados

La Tabla 5.1 muestra los mejores resultados obtenidos para modelos de lenguaje separados con modelos acústicos no adaptados. Se puede observar que los resultados con modelos acústicos separados para castellano fueron mejores que para valenciano, tanto en WER como en SemWER.

Estos resultados pueden ser explicados por varios motivos; uno de ellos es la alta variabilidad de las pronunciaciones en valenciano; otro motivo podría ser que los nombres propios pueden pronunciarse tanto en valenciano como en castellano y además cada persona los pronuncia de diferente manera; otra razón puede ser la calidad de las transcripciones de las frases de valenciano que fueron usadas en el proceso de entrenamiento, que no fueron tan fieles como las transcripciones de castellano.

Con los modelos acústicos mixtos, el WER en castellano fue mejor que el WER para los modelos acústicos separados. Sin embargo, el WER para valenciano era el peor de todos.

Para explicar estos resultados, calculamos el WER para cada grupo de personas según la lengua materna del hablante. La Tabla 5.1 muestra que los hablantes cuya lengua materna era el valenciano hablaban ambos idiomas mejor que aquellos que su lengua materna era castellano. Además, los castellano parlantes hablaban valenciano con peor competencia de la que muestran los valenciano parlantes hablando castellano.

La posible razón es que actualmente en la Comunidad Valenciana es obligatorio aprender tanto castellano como valenciano. Sin embargo, hace unos años sólo era obligatorio aprender castellano. Como consecuencia, en general la gente de la Comunidad Valenciana actualmente habla mejor castellano que valenciano.

Modelos acústicos adaptados

Para obtener unos resultados iniciales con los que comparar los resultados obtenidos con los modelos acústicos adaptados, llevamos a cabo experimentos para castellano y valenciano con modelos acústicos de *Senglar* sin adaptar. Los modelos de *Senglar* en valenciano fueron construídos clonando los modelos más similares de castellano para cada fonema de valenciano.

El mejor resultado que obtuvimos fue un 11.0 % de WER y 7.9 % de SemWER para castellano, y 16.5 % de WER y 18.9 % de SemWER para valenciano. Estos resultados fueron los peores para este trabajo, porque estos modelos acústicos no fueron adaptados para nuestra tarea.

Mod. Acústicos		No mixtos		Mixtos	
Mod. lenguaje		Cast	Val	Cast	Val
WER		5.5 %	9.7 %	5.4 %	9.8 %
SemWER		2.2 %	7.8 %	2.0 %	8.2 %
Locutor	Cast	7.2 %	12.5 %	7.1 %	12.8 %
Nativo	Val	3.7 %	6.7 %	3.7 %	6.7 %

Tabla 5.2: Resultados para modelos de lenguaje separados con modelos acústicos adaptados con la técnica de mínimos cuadrados.

La Tabla 5.2 muestra los mejores resultados para modelos de lenguaje separados con modelos acústicos adaptados. Estos modelos fueron obtenidos de modelos acústicos telefónicos en castellano (modelos de *Senglar*), usando la parte de entrenamiento del corpus como datos de adaptación y calculando sólo una matriz global de adaptación por lengua mediante la técnica de mínimos cuadrados.

Los resultados mostrados en la Tabla 5.2 son peores que los presentados hasta ahora por la forma en la que los modelos acústicos se obtuvieron. Sin embargo, el comportamiento era similar al comportamiento observado en la Tabla 5.1. La explicación para estos resultados es la misma que la explicación de los resultados mostrados en la Tabla 5.1.

Implementamos las tres variantes de MLLR (matriz completa, matriz diagonal y mínimos cuadrados) presentadas en la Sección 3.4 a fin de comparar el rendimiento de los modelos adaptados con las distintas opciones. Las matrices completa y diagonal fueron calculadas con sólo una iteración del algoritmo EM ('Expectation-Maximization'), porque más iteraciones no mejoraban los resultados, algo que suele ser habitual [37].

La Tabla 5.3 muestra los mejores resultados que se obtienen con todas las variantes estudiadas de MLLR a partir de los modelos acústicos mixtos. En las mismas condiciones, otra herramienta estándar de adaptación (HTK) [40] obtiene

	Castellano	Valenciano
Iniciales	11.0 %	16.5 %
Matriz completa	6.0 %	10.4 %
Matriz diagonal	7.3 %	11.4 %
Mínimos cuadrados	**5.4 %**	**9.8 %**

Tabla 5.3: Resultados de las variantes de MLLR.

resultados semejantes: con una matriz completa obtiene 6.0 % para castellano y 11.0 % para valenciano.

La técnica de mínimos cuadrados obtiene los mejores resultados para esta experimentación porque la diferencia entre los modelos acústicos de Senglar y los datos de adaptación es suficientemente grande para hacer que la diferencia en las covarianzas sea una fuente de error cuando calculamos la ocupación del estado $\gamma_{qi}(t)$. La diferencia entre la matriz completa y la matriz diagonal es pequeña, pero con una matriz completa los resultados son mejores que con una matriz diagonal, algo que era previsible [21].

5.2.2. Modelo de lenguaje mixto

En este apartado, los experimentos se realizan sobre un modelo de lenguaje único que admite ambas lenguas. Por tanto, sólo se emplean modelos acústicos mixtos.

Mod. acústicos mixtos		No adaptados	Adaptados
WER		7.4 %	11.8 %
SemWER		3.0 %	5.2 %
IDL		98.1 %	97.1 %
Locutor	Cast	9.1 %/10.5 %	15.1 %/14.8 %
Nativo	Val	3.6 %/6.4 %	7.2 %/10.3 %

Tabla 5.4: Resultados para el modelo de lenguaje mixto. El WER de ”Locutor nativo” se presenta como castellano/valenciano WER.

La Tabla 5.4 muestra los resultados para el modelo de lenguaje mixto. Como en este caso es posible un reconocimiento multilingüe, también se calcula la tasa de identificación (IDL). Para determinar el lenguaje en que las frases han sido reconocidas, contamos el número de palabras que se han reconocido en cada lengua en las frases reconocidas. La lengua que tenía el número más alto de palabras en la cuenta era asignado como la lengua de la frase reconocida. Por ejemplo, ”bona nit quería el correo electrónico de carlos hinarejos gracias”, es una frase

que tiene dos palabras en valenciano (”bona” y ”nit”), cuatro palabras en castellano (”quería”, ”correo”, ”electrónico” y ”gracias”) y las otras palabras (”el”, ”de”, ”carlos” y ”hinarejos”) están presentes en las dos lenguas (por lo tanto, estas palabras se omiten en la cuenta de la tasa de identificación).

Los resultados (con modelos acústicos mixtos no adaptados) con el modelo de lenguaje mixto eran peores que con modelos acústicos mixtos en modelos de lenguaje separados (Tabla 5.1), porque en media el WER en ese experimento es 4.7 % y el SemWER es 2.7 %. El modelo de lenguaje mixto produce más errores porque la perplejidad es más alta (8.42) que la perplejidad para modelos de lenguaje separados (5.98-6.46). Por lo tanto, para los modelos acústicos no adaptados la mejor opción es modelos de lenguaje separados y modelos acústicos mixtos.

Los resultados (con modelos acústicos mixtos adaptados) con el modelo de lenguaje mixto eran peores que con modelos acústicos mixtos adaptados en modelos de lenguaje separados (Tabla 5.2), porque en media el WER en ese experimento es 7.6 % y el SemWER es 5.1 %. En estas condiciones el WER es el peor de todos, aunque el SemWER en promedio, es igual a la media de SemWER en la Tabla 5.2 (5.1). Esto puede ser explicado por la diferencia de las condiciones de grabación del corpus de *Senglar* y las condiciones de grabación de nuestro corpus.

Por lo tanto, a pesar de la adaptación, los modelos acústicos originales tenían más influencia que nuestros datos de adaptación. Además, los resultados para valenciano eran peores porque valenciano es diferente del idioma usado en los modelos acústicos de *Senglar* (castellano).

La tasa de identificación era muy buena, y esto es positivo porque en un sistema multilingüe es muy importante para el sistema adivinar qué lenguaje está siendo hablado (por ejemplo, para que el sistema pueda responder en el mismo lenguaje que el usuario está hablando).

”Locutor nativo” en la Tabla 5.4 se presenta como WER castellano/valenciano. Los resultados muestran la misma tendencia que en otras tablas, es decir, los locutores castellano parlantes hablan peor tanto en castellano como valenciano que los valencianos parlantes.

Toda la experimentación presentada en este apartado puede ser consultada en [23] y [26].

5.3. Resultados con el reconocedor iATROS

ATROS sólo permite llevar a cabo experimentación con modelos de lenguaje que sean autómatas de estados finitos. Por esta razón implementamos iATROS, para poder llevar a cabo experimentación con n-gramas de cualquier grado.

Repetimos el experimento con iATROS con modelos de lenguaje separados y modelos acústicos separados para cada lengua y sin adaptar. Los experimentos se repiten con los autómatas de estados finitos y se prueban n-gramas de diferentes grados, estimados a partir de todas las frases de entrenamiento, pero los mejores resultados se obtienen con dos bi-gramas, una para cada lengua.

Sistema	Modelo	Castellano	Valenciano
ATROS	AEF	3.5 %	5.4 %
	2-grama	–	–
iATROS	AEF	2.8 %	4.8 %
	2-grama	3.3 %	5.2 %

Tabla 5.5: Resultados ATROS/iATROS.

La Tabla 5.5 muestra resultados para el mismo experimento, pero con los dos sistemas de reconocimiento ATROS e iATROS. Con los autómatas de estados finitos, iATROS da una mejora relativa de un 20 % para castellano y un 11 % para valenciano. Además, iATROS es más rápido para esta tarea y más eficiente con la memoria que ATROS. Los n-gramas no mejoran demasiado los resultados porque la tarea es muy sencilla, pero si la tarea fuera más complicada nos interesaría poder probar n-gramas de diferentes grados. En iATROS, los n-gramas dan una mejora relativa de un 7 % para castellano y un 3 % para valenciano respecto a los resultados con estados finitos en ATROS.

CAPÍTULO 6

MEJORAS BASADAS EN LA IDENTIFICACIÓN DEL IDIOMA

6.1. Introducción

Los resultados obtenidos con el modelo de lenguaje que contiene las dos lenguas podrían ser mejores, debido a la posibilidad de incluir palabras del idioma incorrecto en la decodificación. Por lo tanto, vamos a trabajar para mejorar el WER que se obtiene por este tipo de errores.

Como se puede ver en la Tabla 6.1 los resultados a nivel semántico son muy buenos, y se puede decir que la identificación del idioma (IDL) es muy buena. En esta tabla la identificación del idioma la habíamos llevado a cabo por la técnica de conteo descrita en la Sección 5.1, pero con las técnicas explicadas en la Sección 3.2 identificamos sin errores el idioma correspondiente. Vamos a utilizar estas técnicas y grafos de palabras para intentar obtener una mejoría del WER.

Mod. acústicos mixtos no adaptados		
WER		7.4 %
SemWER		3.0 %
IDL		98.1 %
Locutor Nativo	Cast	9.1 %/10.5 %
	Val	3.6 %/6.4 %

Tabla 6.1: Resultados para el modelo de lenguaje mixto. El WER de ”Locutor nativo” se presenta como castellano/valenciano WER.

6.2. Identificación del idioma basada en grafos de palabras

Los errores que se producen a nivel de palabra en el reconocedor bilingüe se deben en buena parte al cambio de idioma que se produce en una frase que tiene el idioma bien reconocido, ya que la mayoría de palabras están bien reconocidas en el idioma que corresponde; incluso las palabras que están en el idioma incorrecto estarían bien reconocidas si estuvieran en el idioma correcto. Por lo tanto, con la identificación del lenguaje basada en la técnica de conteo y los grafos de palabras explicados en la Sección 3.3, se puede intentar mejorar el WER. El mejor WER que se puede alcanzar sería aproximadamente 4.7 % (como se puede ver en la Tabla 5.1, (3.2+6.2)/2). Esto sería equivalente a aplicar previamente al reconocimiento las técnicas presentadas en la Sección 3.2 para identificar el idioma, y luego emplear un modelo de lenguaje sólo del idioma identificado.

Con los grafos de palabras llevamos a cabo diferentes experimentos presentados en la Sección 3.3. Por lo tanto, vamos a explicar detalladamente cada uno de ellos y analizaremos los resultados obtenidos. Por último compararemos todos los resultados obtenidos con los grafos de palabras y el resto de experimentos, a fin de decidir cuál es la mejor opción para el sistema de reconocimiento.

La problemática de utilizar grafos de palabras radica en aquellas frases que en el reconocimiento inicial fallan y no se decodifica nada, ya que dichas frases no tienen grafo de palabras. Por lo tanto, no se puede aplicar ninguna técnica que emplee los grafos de palabras para mejorar el reconocimiento en dichas frases.

6.2.1. Segundo reconocimiento

La primera idea es utilizar la probabilidad del modelo de lenguaje que aporta el grafo de palabras para usarla en un nuevo modelo de lenguaje empleado para hacer un segundo reconocimiento, ya que lo que necesitamos es que la frase reconocida sea sólo del lenguaje correcto. Para ello, sabiendo el idioma de la frase a reconocer (porque la identificación del idioma funciona bastante bien), se construye un modelo de lenguaje a partir del grafo de palabras sólo con palabras del idioma correspondiente. Tras ello, se vuelve a reconocer con el nuevo modelo de lenguaje.

Para obtener los grafos de palabras correspondientes a las frases utilizamos los parámetros de reconocimiento con los que obteníamos el 7.4 % de WER. Para llevar a cabo el reconocimiento con los grafos de palabras filtrados y convertidos en modelos de lenguaje se hizo un nuevo ajuste de parámetros que mostró poca variabilidad en los resultados.

El mejor resultado obtenido es de 7.6 % de WER, que es peor que el 7.4 % que ya teníamos. Esto es debido a las frases que tienen el idioma mal identificado y

a las frases que en el segundo reconocimiento no reconocen nada. Las frases que tienen el idioma bien identificado mejoran el resultado porque si tenían alguna palabra en el idioma incorrecto, con el segundo reconocimiento han reconocido las palabras en el idioma correcto.

Sin embargo, en las frases que tenemos mal identificado el idioma estamos fallando más palabras que antes. Esto es debido a que las frases que tienen el idioma mal identificado mezclan palabras de los dos idiomas en la frase reconocida. Algunas de las palabras serán del idioma correcto, y puede que estén bien identificadas, aportando aciertos y decrementando el WER para esas frases. Como en este nuevo experimento obligamos a que la frase sea de un solo idioma, si el idioma escogido es incorrecto estamos fallando esas palabras que antes estábamos reconociendo bien.

Lo que mejoramos con las frases que tienen el idioma bien reconocido es menor que lo que perdemos al no reconocer unas cuantas frases en el segundo reconocimiento, más las palabras que fallamos en las frases que tenemos el idioma mal reconocido.

Un posible ejemplo de esta situación es:

- La frase real en castellano es ”hola quiero saber la extension de fernando hasta luego”.
- En el primer reconocimiento se reconocía ”hola quin es l adreca de fernando hasta luego”.
- El idioma de esta frase es valenciano, pero a pesar de fallar el idioma estamos acertando la despedida.
- Al obligar que la frase sea toda del idioma reconocido, la frase que se obtiene con el grafo de palabras es ”hola quin es l adreca de fernando per favor”.

En este caso fallamos la despedida que en el caso de la primera decodificación se estaba acertando. Por lo tanto, aquellas frases que tienen el idioma mal reconocido están introduciendo errores. Así se explica que el WER empeore de 7.4 % a 7.6 %, por culpa de estas frases y de los no reconocimientos.

Para corroborar que una parte de la culpa del empeoramiento del WER es por culpa de las frases que tenían el idioma mal identificado, repetimos la experimentación pero dotando al sistema de los idiomas correctos. Esta experimentación no puede ser usada y se lleva a cabo solamente para comprobar que el error obtenido se debe a la mala identificación del idioma. En este caso el WER desciende a 6.7 %, demostrando de esta forma que los casos en los que hemos fallado en la identificación del idioma son los que están produciendo gran parte del empeoramiento del WER.

6.2.2. Viterbi monolingüe

Como la experimentación con la probabilidad del modelo de lenguaje no obtiene una mejoría del WER, pensamos en utilizar la probabilidad de reconocimiento que tienen los grafos de palabras para obtener por Viterbi la frase más probable del grafo de palabras en un solo idioma. Para ello, usando la misma técnica de identificación del idioma, se obtiene un grafo de palabras en ese idioma y se decodifica la frase más probable del grafo de palabras. En este caso no es necesario hacer un ajuste de parámetros.

Con la identificación del idioma que tiene errores (es decir, la identificación del idioma que se obtiene por conteo después de hacer el reconocimiento), obtenemos un 7.3 % de WER, mejorando ligeramente el WER que teníamos sin utilizar los grafos de palabras. En este caso los grafos mejoran las frases obtenidas y además no hay frases sin reconocer.

Por ejemplo:

- La frase real en castellano es ”buenas noches por favor queria avisar al profesor alpuente frasnedo adios”.
- En el primer reconocimiento obteníamos la frase ”buenas noches por favor queria avisar professor alpuente frasnedo adios”.
- En esta frase el idioma se reconocía bien, pero falta el artículo ”al” y ”professor” es una palabra valenciana.
- En el grafo de palabras se obtiene ”buenas noches por favor queria avisar al profesor alpuente frasnedo adios”, acertando la frase por completo.

Otro ejemplo:

- La frase real en castellano es ”por favor me gustaria saber el mail de alvarez hasta luego”.
- En el primer reconocimiento se obtenía ”por favor me gustaria saber el mail de alvarez au”.
- En este caso sólo se falla la despedida, que se hace en valenciano en vez de en castellano.
- Con los grafos de palabras este problema se soluciona, ya que se obtiene ”por favor me gustaria saber el mail de alvarez hasta luego”, acertando toda la frase.

En el caso en que el idioma esté mal identificado ocurre lo mismo que en el experimento anterior, que las palabras de una frase que estaban en el idioma correcto y antes se acertaban ahora están fallando. El WER mejora de un 7.4 % a 7.3 % por los casos en los que mejora cuando el idioma está bien identificado y porque no hay frases sin reconocimiento (ya que en el experimento anterior se obtenían varias frases sin reconocimiento); los grafos de palabras también producen una mejoría en las frases. Si utilizamos los idiomas correctos, el WER desciende al 6.3 %.

6.2.3. Viterbi bilingüe

Por último, una nueva idea es aplicar el algoritmo de Viterbi para las dos lenguas (castellano y valenciano) en el grafo de palabras original (sin filtrar con el idioma reconocido). Así, se buscan las frases sólo en castellano y sólo en valenciano con mayor probabilidad, y escogemos aquella frase que obtiene una probabilidad más alta. Con esta técnica obtenemos un 7.0 % de WER.

La mejoría respecto a los anteriores experimentos es que el número de frases decodificadas en el idioma correcto es mayor que en los casos de Viterbi monolingüe y segundo reconocimiento, ya que con esta técnica decodificamos la frase en el idioma correcto un número mayor de veces. Esto ocurre porque la frase en el idioma correcto obtiene normalmente una probabilidad más alta que la frase en el idioma incorrecto. Esto supone una mejoría en la decodificación de las frases obtenidas, ya que al aplicar el algoritmo de Viterbi sobre los grafos de palabras para cada lengua obtenemos frases con todas las palabras en el mismo idioma. Por lo tanto, mejoramos en aquellas frases que se decodificaban palabras en el idioma incorrecto y ahora se decodifican en el idioma correcto. Además, en este caso los grafos mejoran las frases obtenidas en el idioma correcto y no hay frases sin reconocer.

Por ejemplo:

- La frase en castellano ”hola quiero saber la extension de fernando hasta luego”.
- En el primer reconocimiento se reconocía ”hola quin es l adreca de fernando hasta luego”.
- En los anteriores experimentos esta frase se fallaba porque el idioma reconocido de esta frase es valenciano.
- Con el algoritmo de Viterbi se obtiene ”hola quiero la extension de fernando hasta luego”, que es la frase decodificada perfectamente.

En esta frase se puede ver claramente lo que sucede en este experimento respecto a los anteriores. Como se puede apreciar, el idioma de la frase se reconoce perfectamente y la frase se decodifica sin errores.

Con esta técnica los fallos se deben mayoritariamente a errores de reconocimiento, ya que el 98.3 % de las frases se han decodificado en el idioma correcto.

6.2.4. Comparación de resultados

La Tabla 6.2 muestra los resultados obtenidos con los grafos de palabras.

Técnica	WER
Rec	7.4 %
Rec+lm	7.6 %/6.6 %
Rec+rec	7.3 %/6.3 %
Rec+vit	7.0 %
Monolingüe	4.7 %

Tabla 6.2: Resultados utilizando grafos de palabras y el caso monolingüe. El primer resultado es con la identificación de la lengua que se obtiene con la técnica de conteo y el segundo resultado es dotando al sistema del idioma correcto.

En esta tabla:

- ”Rec” muestra el resultado obtenido con un solo reconocimiento con el modelo de lenguaje conjunto. En el reconocimiento se pueden encontrar frases sin reconocer que no generan grafos de palabras.
- ”Rec+lm” muestra los resultados con un primer reconocimiento con el modelo de lenguaje conjunto y utilizando la información que da la probabilidad de modelo de lenguaje de los grafos de palabras para un nuevo reconocimiento (segundo reconocimiento). En el nuevo reconocimiento se pueden encontrar frases sin reconocer. El primer resultado es con la identificación de la lengua que se obtiene con la técnica de conteo y el segundo resultado es dotando al sistema del idioma correcto.
- ”Rec+rec” muestra los resultados con un primer reconocimiento con el modelo de lenguaje conjunto y utilizando la información que da la probabilidad de reconocimiento de los grafos de palabras del idioma reconocido para obtener la frase más probable (Viterbi monolingüe). Todas las frases son reconocidas. El primer resultado es con la identificación de la lengua que se obtiene con la técnica de conteo y el segundo resultado es dotando al sistema del idioma correcto.

- ”Rec+vit” muestra el resultado con un primer reconocimiento con el modelo de lenguaje conjunto y aplicando el algoritmo de Viterbi para las dos lenguas al grafo de palabras original (Viterbi bilingüe), escogiendo aquella frase que tiene la probabilidad más alta. Todas las frases son reconocidas.

- ”Monolingüe” muestra el resultado que se obtiene si se identifica el idioma correctamente antes de llevar a cabo el reconocimiento y se emplea el modelo de lenguaje de dicho idioma en el reconocimiento .

Como se puede observar, estas técnicas mejoran el resultado que se obtiene con sólo el reconocimiento, pero lo que realmente es beneficioso es utilizar la técnica de identificación del idioma primero y después utilizar un modelo de lenguaje del idioma que se ha identificado (siempre y cuando dicha identificación no presente errores). De esta forma se obtiene un 4.7 % de WER (como se puede ver en la Tabla 5.1, (3.1+6.2)/2).

6.3. Identificación del idioma basada en K-NN

Por último, vamos a llevar a cabo la identificación del idioma con las técnicas biométricas explicadas en la Sección 3.2

Para llevar a cabo esta experimentación elegimos un subconjunto del corpus. El corpus original estaba dividido en 5 grupos, con hombres y mujeres castellano y valenciano parlantes hablando en castellano y valenciano. Por lo tanto, consta de 4 tipos para cada lengua (hombres y mujeres castellano y valenciano parlantes hablando en esa lengua).

El corpus de entrenamiento para estas pruebas consta de 60 frases para cada lengua (120 frases en total), de las cuales hay 15 frases de cada tipo y todas ellas del grupo 5. El corpus de test consta de 80 frases para cada lengua (160 frases en total), de las cuales hay 5 frases de cada tipo de los 4 primeros grupos. Hay que hacer notar que se intercambian los grupos de entrenamiento y de test respecto al resto de experimentos presentados, a fin de que el número de frases del conjunto de test sea mayor que el número de frases del conjunto de entrenamiento. Esto se debe a que con los primeros experimentos realizados con la distribución habitual de entrenamiento y test los resultados ya eran muy buenos. Por lo tanto, se utilizaron las particiones de dicha forma para propiciar una situación más desfavorable. Con este conjunto de datos identificamos el idioma entre castellano y valenciano.

En primer lugar, preprocesamos las señales acústicas para convertirlas en vectores de cepstrales. Variamos el número de cepstrales con valores de 5, 10, 15 y 20 cepstrales; obtenemos la primera y la segunda derivada. Se experimentará sin derivadas, con la primera derivada y con la primera y la segunda derivada.

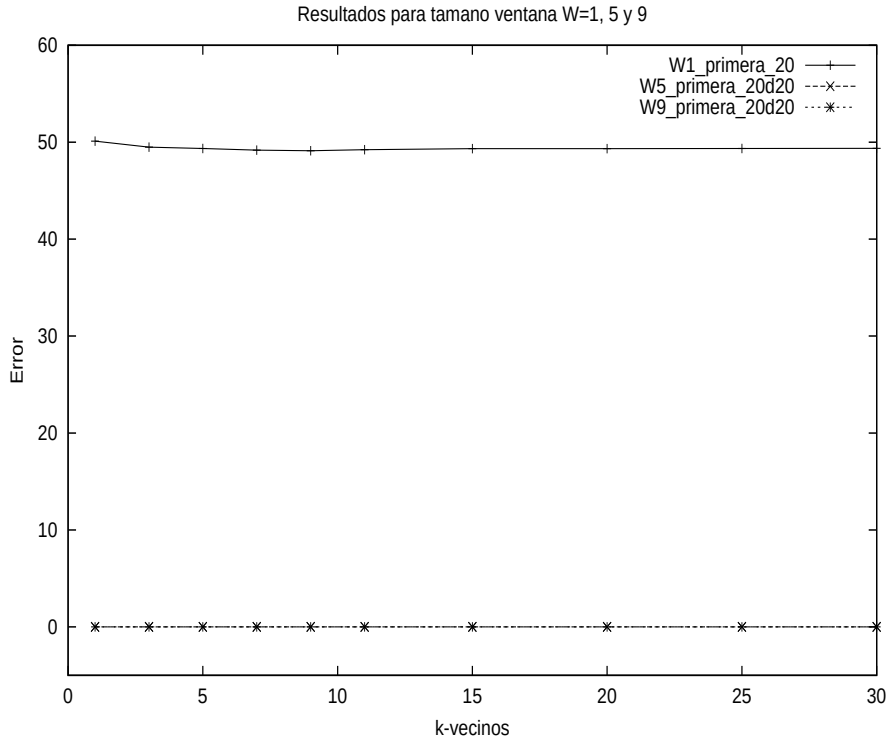

Figura 6.1: Resultados para tamaño de ventana W=1, 5 y 9

Con la señal acústica convertida en vectores de cepstrales llevamos a cabo los dos pasos del preproceso:

- Primer paso: seleccionamos las partes de la señal que tienen la información más relevante; variamos el tamaño de la ventana con valores de 1, 5 y 9.
- Segundo paso: reducimos con la técnica de PCA aquellos vectores que se habían obtenido con tamaños de ventana 5 y 9 a vectores de dimensión 10, 20 y 30.

Con este proceso terminado, etiquetamos los datos con el idioma correspondiente. Para clasificar utilizamos la técnica de k-vecinos con el cálculo de la distancia euclídea. Aplicamos la técnica de k-vecinos con valores de k de 1, 3, 5, 7, 9, 11, 15, 20, 25 y 30 vecinos.

En resumen, para llevar a cabo la experimentación variamos:

- Número de cepstrales: 5, 10, 15 y 20.
- Cepstrales: sin derivadas, con primera derivada, y con primera y segunda derivada.

- Tamaño de la ventana (W): 1, 5 y 9.
- Dimensión final en PCA: 10, 20 y 30.
- k-vecinos: 1, 3, 5, 7, 9, 11, 15, 20, 25 y 30.

Con todas las posibles combinaciones obtuvimos un gran número de resultados. Lo más importante es que con valores de ventana W igual a 5 y 9 se obtiene un 0.0 % de error de identificación, pero con valores de ventana W igual a 1 los resultados obtenidos rondaban el 50.0 % de error de identificación. Esto puede verse en la gráfica de la Figura 6.1, donde ha sido representado en el eje horizontal el número de vecinos y en el eje vertical el error de identificación. La función que se ajusta al valor 50 de error de identificación es para tamaño de ventana 1 y el resto son para tamaño de ventana 5 y 9. Los resultados que se muestran en la gráfica se obtuvieron con la primera derivada, 20 cepstrales y una dimensión final igual a 20. El resto de resultados no se muestran porque no aportan conclusiones diferentes. La conclusión que se obtiene de esta experimentación es que, para esta tarea, podemos acertar siempre en la identificación del idioma con esta técnica usando un tamaño de ventana lo bastante amplio.

CAPÍTULO 7

CONCLUSIONES Y TRABAJO FUTURO

El trabajo realizado ha demostrado que el reconocimiento bilingüe entre valenciano y castellano es factible, ya que las tasas de reconocimiento obtenidas son buenas. Por lo tanto, podemos afirmar que es posible construir un sistema bilingüe con un rendimiento adecuado de reconocimiento.

Como sistema de reconocimiento se puede decir que es una buena idea emplear iATROS frente a ATROS, ya que iATROS obtiene mejores resultados para esta tarea y permite llevar a cabo una experimentación más amplia (mayor variedad de modelos de lenguaje, más parámetros a ajustar,...).

Según los experimentos llevados a cabo, para obtener un buen sistema de reconocimiento automático bilingüe se deberían emplear modelos acústicos mixtos, ya que obtienen mejores resultados que los modelos acústicos no mixtos. Además, se deberían utilizar modelos de lenguaje separados para cada idioma, porque los resultados mejoran con dos modelos separados frente a un modelo conjunto.

La adaptación de modelos acústicos con unos modelos acústicos iniciales de mala calidad o muy diferentes a los datos de adaptación no resulta muy buena idea. Para este trabajo, el método de adaptación que mejores resultados ha proporciando ha sido la técnica de mínimos cuadrados de MLLR.

La identificación del idioma es muy importante en sistemas multilingües; por ello podemos afirmar que la técnica de conteo empleada después de un reconocimiento es buena, pero las técnicas biométricas empleadas para identificar el idioma antes de un reconocimiento son mejores. Además, resultan necesarias para emplear un sistema de reconocimiento con modelos de lenguaje separados, lo que permite mayor calidad de reconocimiento. Si es necesario emplear un modelo de lenguaje conjunto, los grafos de palabras ayudan a mejorar el WER final. Podemos afirmar que aplicar el algoritmo de Viterbi después del reconocimiento

(Viterbi bilingüe) es la mejor opción de las distintas alternativas presentadas en este trabajo.

Los resultados obtenidos en este trabajo son buenos, pero no se pueden considerar definitivos porque los experimentos se han llevado a cabo con un corpus pequeño. Por lo tanto, en el futuro es necesario repetir los experimentos con un corpus más grande y más usado por la comunidad de reconocimiento del habla. Además, se pueden estudiar e implementar más tareas de adaptación, como, por ejemplo, adaptación de los modelos de lenguaje. También se pueden estudiar más técnicas de identificación del idioma.

Otro punto que es necesario estudiar en el futuro es la transcripción ortofonética del valenciano, ya que en el presente trabajo dicha transcripción se ha realizado adaptando el transcriptor de español. Esto implica con bastante seguridad que estamos introduciendo errores, pues no podemos asegurar que la construcción de los modelos léxicos sea la adecuada. Esta razón puede explicar que en valenciano se alcancen peores tasas de reconocimiento que en castellano (para los mismos experimentos). Es necesario que en el futuro se estudie este problema con algún corpus de catalán que se encuentre disponible.

Bibliografía

[1] http://www.tc-star.org/.

[2] M. Adda-Decker. Towards multilingual interoperability in automatic speech recognition. *Speech Communication*, 35:5–20, 2001.

[3] V. Alabau and C.D. Martínez. Bilingual speech corpus in two phonetically similar languages. In *Proc. of LREC'06*, pages 1624–1627, 2006.

[4] James Allen. *Natural language understanding (2nd ed.).* Benjamin-Cummings Publishing Co., Inc., isbn 0-8053-0334-0, Redwood City, CA, USA, 1995.

[5] F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. García-Varea, D. Llorens, C. Martínez, S. Molau, F. Nevado, M. Pastor, D. Picó, and A. Sanchis. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*, 18:25–47, 2004.

[6] D. G. Childers, D. P. Skinner, and R. C. Kemerait. The cepstrum: A guide to processing. In *Proc. IEEE,*, volume 65 of *Institute of Electrical and Electronics Engineers, Inc. Conference*, pages 1428–1443, 1977.

[7] G. D. Forney. The viterbi algorithm. In *Proceedings of the IEEE*, volume 61(3):2, pages 268–278, mar 1973.

[8] S. Furui. Toward robust speech recognition under adverse conditions. In *Proceedings ESCA Workshop on Speech Processing in Adverse Conditions*, pages 31–41, nov 1992.

[9] G.Docherty and L. Shockey. *Speech Synthesis*. Aspects of Speech Technology. Edinburgh: Edinburgh University Press '88, 1988.

[10] John E. Hopcroft and Jeffrey Ullman. *Introducción a la Teoría de Autómatas, Lenguajes y Computación.* Ed. CECSA., 1995.

[11] H. G. Huang. Speech recognition in adverse environments. *Computer Speech and Language*, 5:275–294, 1991.

[12] X.D. Huang, Y. Ariki, and M.A. Jack. *Hidden Markov Models for speech recognition.* Edinburgh University Press., 1990.

[13] Xuedong Huang, Fileno Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, and Ronald Rosenfeld. The SPHINX-II speech recognition system: an overview. *Computer Speech and Language*, 7(2):137–148, 1993.

[14] Moisés Pastor i Gadea. *Aportaciones al reconocimiento automático de texto manuscrito*. PhD thesis, Dep. de Sistemes Informàtics i Computació, València, Spain, Oct 2007. Advisors: E. Vidal and A.H. Tosselli.

[15] I. T. Jolliffe, B. J. T. Morgan, and P. J. Young. A simulation study of the use of principal components in linear discriminant analysis. *J. Stat. Comput. Simul.*, 55:353–366, 1996.

[16] B. S. Kim and S. B. Park. A fast k nearest neighbor finding algorithm based on the ordered partition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):761–766, 1986.

[17] J. Köhler. Multilingual phone models for vocabulary-independent speech recognition tasks. *Speech Communication*, 35:21–30, 2001.

[18] Thomas Koller, Emanuela Cresti, and Massimo Moneglia (Eds.). C-oral-rom, integrated reference corpora for spoken romance languages. *Machine Translation*, 20(4):297–300, 2006.

[19] C.H. Lee, C.H. Lin, and B.H. Juang. A study on speaker adaptation of the parameters of continuous density hidden markov models. *IEEE Transactions on Signal Processing*, 39(4):806–814, apr 1991.

[20] C. Leggetter and P. Woodland. Flexible speaker adaptation using maximum likelihood linear regression. In *Proc. Eurospeech'95*, pages 1155–1158, Madrid., 1995.

[21] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, pages pp. 9:171–185., 1995.

[22] D. Llorens, V. M. Jiménez, J. A. Sánchez, E. Vidal, and H. Rulot. ATROS, an automatically trainable continuous-speech recognition system for limited-domain tasks. In A. Calvo and R. Medina, editors, *VI Spanish Symposium*

on Pattern Recognition and Image Analysis, pages 478–483, Córdoba, España, 1995.

[23] M. Luján, C. D. Martínez, and V. Alabau. A study on bilingual speech recognition involving a minority language. In *3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 138–142, Poznan, (Poland), October 2007.

[24] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT, 1999.

[25] V. Mapelli and K. Choukri. Report contributing to the design of an overall co-ordination and strategy in the field of lrs. Technical report, Paris, 2003. ENABLER Deliverable D5.2.

[26] Míriam Luján Mares, Carlos David Martínez Hinarejos, and Vicent Alabau Gonzalvo. Evaluation of several maximum likelihood linear regression variants for language adaptation. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008.

[27] R. Paredes, E. Vidal, and F. Casacuberta. Local features for speaker recognition. *SPR 2004.International Workshop on Statistical Pattern Recognition. LNCS 3138 of Lecture Notes in Computer Science*, pages 1087–1095, 2004.

[28] A. Quilis. *Tratado de fonología y fonética españolas*. Madrid (Gredos), 2nd edition, 1999.

[29] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall Ptr, 1993.

[30] A. Sanchis, A. Juan, and E. Vidal. New features based on multiple word graphs for utterance verification. In *8th International Conference on Spoken Language Processing*, pages 2545–2548, October 2004.

[31] T. Schultz and A. Waibel. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(Issues 1-2):31–51, August 2001.

[32] Tanja Schultz and Katrin Kirchhoff. *Multilingual Speech Processing*. Academic Press, Inc., Orlando, FL, USA, 2006.

[33] Zheng-Hua Tan and Brge Lindberg. *Automatic Speech Recognition on Mobile Devices and over Communication Networks (Advances in Pattern Recognition)*. Springer Publishing Company, Incorporated, 2008.

[34] UCL. *SAMPA computer readable phonetic alphabet*, 1993.

[35] U. Uebler. Multilingual speech recognition in seven languages. *Speech Communication*, 35:53–69, 2001.

[36] Andrew Viterbi. Error bounds for convolutional codes and a asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269, 1967.

[37] Phil C Woodland. Speaker adaptation for continuous density hmms: A review. *ITRW on Adaptation Methods for Speech Recognition*, pages pp. 11–19., August 29-30, 2001.

[38] Yuehua Xu and Alan Fern. On learning linear ranking functions for beam search. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 1047–1054, New York, NY, USA, 2007. ACM.

[39] S. Young. Large vocabulary continuous speech recognition: A review. *In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 3–28, dec 1995.

[40] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book*. CUED, UK, v3.2 edition, July, 2004.

[41] Xufang Zhao and Douglas O'Shaughnessy. An evaluation of cross-language adaptation and native speech training for rapid hmm construction based on very limited training data. In *Interspeech 2007*, August 27-31, 2007.

[42] M. A. Zissman and K. M. Berkling. Automatic language identification. In *Proceedings of the workshop on Multilingual Interoperability in Speech Technology. European Speech Communication Association - NATO.*, pages 93–101, 1999.