Problemática de la recogida y anotación de una base de datos oral para el gallego

Begoña González Rei
Facultade de Filoloxía, Universidade de Santiago de Compostela
Tlf. 981 563100 - fgbego@usc.es

Antonio Cardenal López, Laura Docío Fernández y **Carmen García Mateo**
Departamento de Tecnologías de las Comunicaciones, Universidad de Vigo
Tlf. 986 812133 - Fax: 986 812116 - cardenal,ldocio,carmen@gts.tsc.uvigo.es

Resumen La creciente demanda de los denominados "teleservicios" requiere la recogida de bases de datos adecuadas para entrenar y evaluar los sistemas de reconocimiento automático de voz. Para lenguas habladas por grandes poblaciones se disponen en el mercado de bases de datos útiles que permiten la implementación de reconocedores. Sin embargo, las lenguas minoritarias sufren la falta de tales bases de datos por lo que casi cualquier investigación en el ámbito de las tecnologías del habla que se centre en una lengua minoritaria debe pasar por una fase en la que se capture una base de datos de voz con la que trabajar. En este artículo se presenta nuestra experiencia en la creación de una base de datos para el idioma Gallego. Se describen las cuestiones relativas a la captación de llamadas y al etiquetado de las mismas. También se muestran experimentos de entrenamiento y evaluación de reconocedores realizados sobre dicha base de datos que sirven como herramienta de validación de la base de datos en cuestión.

1 Introducción

La globalización y evolución de la tecnología de interfaces hombre-máquina que hacen uso del reconocimiento y síntesis de voz, ha provocado una creciente demanda de bases de datos en las lenguas minoritarias de los diferentes países. Este hecho se debe principalmente a que los actuales sistemas de reconocimiento automático de voz se basan en modelos estocásticos como son los modelos ocultos de Markov (HMM). Se necesitan bases de datos de voz adecuadas que permitan la obtención de unos modelos que hagan posible un reconocimiento robusto de voz.

En el mercado existen numerosas bases de datos de voz en lenguas mayoritarias. Sin embargo, para lenguas minoritarias la existencia de bases de datos de voz es poco habitual, por lo que cualquier deseo de desarrollo de un sistema que haga uso de la tecnología del habla en tales lenguas precisará previamente de la recogida de una base de datos en la lengua objeto.

Con el propósito de satisfacer la demanda de una de tales bases de datos de lenguas minoritarias, **el Gallego**, hemos creado una base de datos de voz telefónica para dicho idioma siguiendo las especificaciones del proyecto SpeechDat [1].

Este artículo presenta nuestro trabajo en la creación de esta base de datos. En primer lugar se tratará la cuestión de la captación de los locutores haciendo hincapié en la dificultad de encontrar locutores que permitan la creación de una base de datos que englobe todas las variantes dialectales del idioma Gallego.

En segundo lugar se considera la problemática que engloba la tarea de etiquetado de las llamadas. Nuestra experiencia en el ámbito del reconocimiento automático de voz nos demuestra que ésta es una tarea altamente importante que conviene realizar de forma precisa si se desea entrenar unos *buenos* modelos para reconocimiento de voz. Por ello se han realizado algunas desviaciones de las especificaciones de etiquetado impuestas por SpeechDat. Tales desviaciones consisten básicamente en:

- La inclusión de nuevas etiquetas para realizar una anotación bien diferenciada de determinados eventos acústicos que no son voz que en SpeechDat se engloban dentro de una etiqueta común. Estas etiquetas serán muy útiles para realizar una elección lo más adecuada posible del material a utilizar en el entrenamiento de los modelos, y para evaluar la robustez del reconocedor.
- La incorporación de ciertas etiquetas

y reglas de etiquetado que permitirán una transcripción ortográfica detallada y precisa de lo que ha pronunciado el locutor. Esto será muy útil para el entrenamiento de reconocedores y para la creación de pronunciaciones alternativas de las palabras.

Finalmente, se trata el tema de la validación del etiquetado que se realizará tanto de forma manual como automática a través de experimentos de reconocimiento de voz. Este último método de validación servirá para analizar la fiabilidad e importancia de las anotaciones realizadas de los eventos acústicos.

2 *Captación de llamadas*

El primer objetivo es obtener una base de datos de 1000 hablantes, que responden a un cuestionario de 46 items por teléfono. La recogida de una buena base de datos en el caso del gallego, así como de otras lenguas minoritarias, conlleva una dificultad extra con respecto a otras lenguas estatales o mayoritarias. En primer lugar por no tratarse de una lengua totalmente normalizada y por otra porque la especial situación sociolingüística del gallego (de dos lenguas en contacto) hace habituales las interferencias lingüísticas. El primer objetivo de nuestra base de datos es el de incluir 'buenos hablantes' en gallego. El segundo, el de obtener una representación equilibrada de la variedad de la población, con respecto a varios criterios como el sexo, la edad, el nivel sociocultural y la adscripción dialectal de los informantes.

En cuanto al primer objetivo, se valora la competencia lingüística de los informantes, su fluidez y su correcta pronunciación y prosodia en gallego, después de una preselección en función de datos recogidos en la ficha cubierta por el propio informante en la que declara cuál es su lengua nativa, la lengua hablada por sus padres (no siempre coinciden: es común el caso de padres gallego-hablantes entre si que a sus hijos los educan en castellano) y su lengua habitual. Aproximadamente el 70% de los informantes responde 'gallego' a las tres preguntas, un 15% declara que usa indistintamente las dos lenguas, y cerca del 15% que o bien tienen como lengua nativa el castellano, o bien que, aunque lo sea el gallego, su lengua habitual es el castellano.

En cuanto a la representación de la variedad de la población, en estos momentos en los que aproximadamente contamos ya con la mitad de la base de datos recogida y etiquetada, podemos hacer un balance de la campaña de recogida de llamadas, evaluar cómo se van cumpliendo los objetivos, y describir los principales problemas con los que nos encontramos.

-La representación semejante de los dos sexos es un objetivo fácilmente alcanzable, a pesar de que se observa una mayor colaboración de las mujeres: hoy por hoy los informantes masculinos constituyen aproximadamente un 40% de la base de datos y las mujeres un 60%. Por lo tanto es necesario potenciar la recogida de llamadas de hombres.

-Con respecto a la edad de los informantes, se van cumpliendo los requisitos mínimos del SpeechDat, aunque la mayor parte de los informantes se encuentran en las franjas medias (entre 20 y 45 años). Son dos los problemas fundamentales con los que nos encontramos: por una parte la castellanización de las nuevas generaciones, especialmente urbanas (esto limita la distribución de cuestionarios a aquellos informantes considerados buenos hablantes) y por otra parte el analfabetismo en gallego en las franjas de edad avanzada, que dificulta la lectura fluida del cuestionario por parte de estos informantes (muchos monolingües en gallego y otros gallego-hablantes pero alfabetizados en castellano).

-En cuanto al nivel sociocultural de los informantes, la mayor parte de los informantes pertenecen al medio urbano y tienen una cierta formación académica. Más del 50% tiene estudios superiores y el 75% por lo menos estudios medios. El nivel sociocultural es especialmente importante en el caso del gallego: en cierta medida está relacionado con la edad (analfabetismo en gallego entre los más mayores, como dijimos antes) pero, sobre todo, con la procedencia y con la extracción social. Por una parte en el mundo rural y entre los informantes de nivel sociocultural medio-bajo el gallego tiene una vitalidad mayor que en la ciudad y que entre las clases medio-altas, pero, por otra, en este grupo hay un mayor desconocimiento del estándar. En cuanto a la adscripción dialectal de los informantes, el objetivo que se persigue es representar los tres bloques dialectales principales del gallego (occidental, central y oriental).

3 Etiquetado

El etiquetado de la base de datos oral sigue los parámetros definidos por el proyecto SpeechDat general para todas las lenguas, pero con la incorporación de ciertas novedades que se consideraron imprescindibles para las necesidades del entrenamiento del reconocedor o por la especial idiosincrasia del gallego.

La representación elegida es la ortográfica, aunque también se indican los fenómenos paralingüísticos más importantes y con mayor incidencia en el entrenamiento y validación de reconocedores automáticos de voz.

Con respecto a la metodología seguida, la transcripción la efectúa un grupo de cuatro etiquetadores que trabajan en sesiones de medio día con una revisión posterior de todas las marcas por un supervisor.

La convención SpeechDat incluye transcripciones obligatorias y opcionales. A las marcas opcionales definidas se añadieron otras para facilitar la tarea de entrenamiento del reconocedor en función de necesidades exclusivas de la base de datos del gallego. Se trata de marcas fácilmente eliminables de manera que se pueda recuperar la forma básica de la transcripción.

3.1 La transcripción ortográfica

Con respecto a la **transcripción ortográfica**, los items lexicales se representan con su forma ortográfica normalizada (esto significa que, por ejemplo, en la segunda forma del artículo característica del gallego, los guiones se usan de manera normal). En cuanto a las variaciones de pronunciación (dialectalismos, vulgarismos, pronunciaciones relajadas, etc.), se transcribe la forma estándar, y el registro de las variantes se incluye en el diccionario, donde además de la transcripción fonética estándar se indican las diferentes transcripciones fonéticas de cada palabra. En cuanto a la puntuación, se conserva la puntuación incluida en el texto escrito que se les entregó a los informantes, en los cuales se evitó el punto final, para evitar confusión con el punto que indica pausa. En cuanto a anotación prosódica, opcional en la convención SpeechDat, se utilizan los dos signos propuestos: el punto que indica pausa con silencio y los dos puntos que representan la duración larga de los sonidos. Sólo utilizamos estos signos en casos claros y que nos parecen pertinentes.

3.2 La transcripción paralingüística

Además de la transcripción ortográfica se transcriben algunos eventos acústicos paralingüísticos. Para ello se emplean varios símbolos definidos entre corchetes. La convención SpeechDat señala cuatro categorías de eventos acústicos paralingüísticos, para ruidos originados por el hablante o procedentes de otras fuentes.

- [spk]: *Ruido del hablante.* Todas las clases de sonidos y ruidos hechos por el hablante que no forman parte del texto previsto (tos, gruñidos, clics de la lengua, fuerte respiración, risas, suspiros...)
- [fill]: *Pausas o dudas.* Estos sonidos pueden modelarse en un modelo de 'pausas de relleno' en los reconocedores de voz: *uh, um, er, ah, mm.*
- [sta]: *Ruido continuo.* Esta categoría contiene ruidos de fondo que no son intermitentes y tienen un espectro de amplitud más o menos estable. Ejemplos: ruido de tráfico, de la calle, de la tele, de un lugar público, voces de fondo.
- [int]: *Ruido intermitente.* Esta categoría contiene ruidos de naturaleza intermitente. Estos ruidos generalmente solo se dan una vez (como un portazo), o tienen pausas entre ellos (llamadas de teléfono), o que van cambiando de tipo. Ejemplo: música, conversación de fondo, lloros de un bebé, llamadas de teléfono, portazo, timbre...

El proyecto SpeechDat sólo admite la utilización de estas cuatro categorías. Sin embargo, en nuestro caso, procedimos a una subcategorización de dos de las categorías, utilizando marcas fácilmente traducibles automáticamente para conseguir la forma básica de la transcripción. Así, en el caso de los ruidos del hablante, distinguimos entre [spk1], [spk2] y [spk3].

- El [spk1] es el más frecuente: se corresponde con el ruido de la respiración, casi siempre al principio y al final de la realización, y a veces en pausas intermedias para tomar aire.
- El [spk2] también es muy frecuente, y se corresponde con los clic de la lengua. Su

espectro es muy distinto al del primer caso.

- El [spk3] marca el resto de los casos (risas, tos, gruñidos...), mucho más ocasionales. También se subcategorizan los ruidos de fondo estacionarios. En este caso empleamos dos marcas, que distinguen:
- El [sta1] el fuerte ruido de fondo de la propia linea.
- El [sta2] el ruido de fondo característico del tráfico, de la calle, de la tele, etc..

4 *Fenómenos lingüísticos que dificultan el etiquetado*

A la hora de etiquetar la base de datos del gallego, nos encontramos con problemas de distinta índole. Algunos de ellos son problemas que tienen su origen en la calidad de la grabación o en la presencia de fuertes ruidos de fondo, problemas comunes en el etiquetado de todas las lenguas. Pero otros son problemas lingüísticos, y por lo tanto exclusivos de cada idioma. En el caso del gallego, como de otras lenguas minorizadas, a las particularidades lingüísticas se tiene que sumar la realidad sociolingüística de dos lenguas en contacto. Son muy pocos los individuos que en una sociedad con dos lenguas en contacto mantienen los dos códigos sin interferir. Por otra parte, aunque se ha avanzado mucho a este respecto, el gallego no es aún una lengua totalmente normalizada (existen comportamientos diglósicos), y aún hay un gran número de gallego-hablantes alfabetizados exclusivamente en castellano. A estas dos realidades hay que añadir que el estándar del gallego es relativamente reciente. Todo esto sumado hace que en nuestra base de datos, como en la realidad, nos encontremos con casos de desviaciones de la norma e interferencias léxicas en los distintos planos del lenguaje (fónico, morfosintáctico y léxico).

A la hora de etiquetar nos encontramos fundamentalmente con tres tipos de problemas que comentaremos a continuación: variantes dialectales, vacilaciones fonéticas y castellanismos y realizaciones no estándar.

4.1 Dialectalismos

Los fenómenos fonéticos dialectales más relevantes en el caso del gallego son la *gheada* y el *seseo*. La *gheada* consiste en la pronunciación aspirada faríngea /h/ del fonema oclusivo velar sonoro /g/ en cualquier posición. Esta pronunciación es distinta a la fricativa velar sorda /x/ del castellano, por lo que a la hora de etiquetar se utiliza el dígrafo <gh> para marcar esta realización, y no la letra j. La utilización de este dígrafo es provisional, ya que la transcripción fonética posterior por medio de un transcriptor automático que no considera esta posibilidad falsearía la segmentación en trífonos y dificultaría la tarea de entrenamiento del reconocedor. En cualquier caso, no aparecerá en la transcripción final, sino que deberá indicarse como pronunciación alternativa en el lexicón.

El *seseo* en la pronunciación como una fricativa alveolar sorda de la consonante fricativa interdental sorda, exclusivamente en posición implosiva (seseo parcial) en algunas zonas de Galicia, o en cualquier posición silábica (seseo total) en otras. Es un fenómeno muy extendido, pero que no introduce ningún alófono nuevo en la base de datos. En la transcripción por lo tanto se reproduce la realización estándar y se lleva el registro de las variantes en la transcripción fonética del lexicón.

En cuanto a los dialectalismos morfológicos, los más relevantes atañen a la morfología nominal: se trata de las distintas pronunciaciones de los sufijos flexivos, tanto de género como de número. El caso más común es el del plural de las palabras acabadas en -n: -ns, -s, -is (*pantalóns, pantalós, pantalois; cans, cas, cais*), o en -l: a la solución estándar en -is (*animais*) se suma otra muy común en -s (*animás*) y casos de plurales en -les (solución minoritaria del gallego, pero favorecida por la semejanza al castellano). Otro caso se da en el resultado de las terminaciones latinas -ANUM, -ANAM: en el masculino hay dos soluciones, en -án o -ao (*irmán, irmao*), y otras dos en el femenino: -á o -án (*mañá, mañán*). Al ser la mayor parte de las realizaciones leídas, hay una preponderancia clara de las soluciones estándar sobre las otras, pero estas también tienen su presencia en la base de datos, y todas ellas se registran. También son diversas las variaciones dialectales en la morfología verbal del gallego. Ahora bien, en general podemos decir que la mayoría no aparecen en la base de datos, debido fundamentalmente al registro culto utilizado en el cuestionario. Las variantes diatópicas más recurrentes se dan en la tercera persona de los perfectos de la segunda y tercera

conjugación. Estas y otras variantes no se indican en la transcripción ortográfica sino en el lexicón, como pronunciaciones alternativas, siguiendo las indicaciones del proyecto SpeechDat, aunque se lleva un registro de todas las ocurrencias con el fin de no falsear la segmentación alofónica y facilitar el trabajo de entrenamiento del reconocedor de voz.

4.2 Otros fenómenos fonéticos

Otros fenómenos fonéticos que se reproducen en la base de datos consisten en **vacilaciones vocálicas**, más habituales en la lengua hablada en el caso del gallego que en el de otras lenguas, ya que se trata de una lengua de uso exclusivamente oral durante muchos siglos. Uno de estos fenómenos es la síncopa, que consiste en la pérdida de una vocal (generalmente átona) en interior de palabra: la síncopa de la vocal postónica es típica de un registro vulgar, consecuencia de una pronunciación poco cuidada, difícilmente reproducible en una grabación de este tipo, por lo que aparece solamente en casos aislados en la base de datos (*completísima* /completisma/). Nos encontramos con más frecuencia con la síncopa de la vocal pretónica precedida de oclusiva o fricativa y seguida de una líquida. El caso más habitual es el de la preposición para, cuya pronunciación común en el gallego es /pra/.

Otras vacilaciones vocálicas se producen en los resultados de los grupos latinos -ULT-, -UCT, -OCT-, y -ORI-, es decir, entre los **diptongos** -oi- y -ui- (la realización -ou- característica del gallego oriental es menos recurrente): *moito /muito/, oito /uito/...* También se producen alteraciones vocálicas cuando se produce el encuentro de dos vocales por fonética sintáctica: la casuística es variada y depende de la tonicidad de las vocales. Son comunes las **elisiones**, especialmente cuando las dos vocales son átonas: en este caso se reducen a una sola cuando se trata de la misma vocal (se *estableceu /s'estableceu/*), y a favor de una de ellas si son distintas (*desa estirpe /des'estirpe/, prometinme a min /prometinm'a min/, deica o venres /deic'o venres/*). Cuando una de ellas es tónica generalmente se conservan, pero pueden contraerse también a favor de la tónica. Uno de los casos más frecuentes es el de la asimilación de la preposición para y el artículo (*para un /pr'un/, para o /pr´o/*). Para marcar las elisiones vocálicas habilitamos un nuevo signo, el apóstrofe.

En cuanto al consonantismo, un caso digno de mencionar es la tendencia a la pronunciación relajada de los **grupos consonánticos cultos**. En la mayor parte de los casos consiste en la reducción del primer elemento del grupo consonántico (*victoria /vitoria/*), aunque también nos encontramos con casos de vocalización (*directorio /direitorio/*) y de asimilación regresiva (*significa /sinnifica/*).

Otro de los fenómenos fonéticos recurrentes en la base de datos está relacionado con la alteración de la **-s implosiva**. El caso más común es el **rotacismo**, que consiste en la articulación como /r/ de la /s/ en posición implosiva ante una consoante sonora, en interior de palabra (desde /derde/, mesmo /mermo/) o por fonética sintáctica (*as dúas /ar dúas/*). También registramos casos de **aspiración** de la -s, fenómeno típico del occidente de la provincia de La Coruña, en los mismos contextos (*despois do /despoih do/*). Este fenómeno fonético es muy común en el gallego y, a diferencia del seseo o la gheada, el hablante no es consciente de el. Por eso es difícil de controlar y aparece con mucha frecuencia en la base de datos, no sólo en las realizaciones espontáneas sino también en los ítems leídos.

4.3 Castellanismos y desvíos de la norma escrita

En la base de datos la mayor parte de las realizaciones son leídas, pero también hay un cierto número de realizaciones espontáneas y semiespontáneas (es el caso de los números, presentados en dígitos, o del deletreo, con letras mayúsculas). Es en estos casos, sobre todo, en los que nos encontramos en ocasiones con la presencia de castellanismos o realizaciones no estándar. Los castellanismos léxicos se marcan en la base de datos con el signo &. Se trata de un signo fácilmente eliminable y de esta manera se lleva el control del vocabulario final, distinguiendo estas palabras de las propiamente gallegas. La utilización de este signo es de gran importancia especialmente en aquellos casos en los que el corpus no se utilizará para entrenar unidades fonéticas (trífonos), sino modelos de palabras (números y letras).

En conclusión, toda esta variabilidad complica la tarea de etiquetado y de elaboración del lexicón. Para llevar un control de todas estas posibilidades, utilizamos signos extra fácilmente eliminables de manera que se pue-

da obtener una transcripción simple, ajustada a los parámetros del SpeechDat, y llevamos el control de las desviaciones de la norma, para no multiplicar el lexicón con la multitud de variantes diatópicas y diastráticas presentes por todas estas causas descritas en la base de datos.

5 *Validación del etiquetado*

La validez del etiquetado no sólo se realiza mediante la supervisión manual, sino también mediante la realización de distintas pruebas de reconocimiento automático de habla.

5.1 Pruebas de reconocimiento automático

Las bases de datos SpeechDat han sido diseñadas teniendo en mente el desarrollo de aplicaciones de comunicación hombre máquina. Por ello, en el diseño del material a recoger se ha prestado especial atención a dígitos, números, comandos, y fechas. De todas formas, se previó que con el avance de la tecnología los sistemas de reconocimiento basados en unidades inferiores a la palabra podrían llegar a ser transferibles a la industria. Sólo unos pocos años después, se considera ya como estado del arte estos sistemas y el material de la base de datos de frases equilibradas fonéticamente es al que se le debe prestar una mayor atención. Un sistema de reconocimiento basado en unidades inferiores a la palabra responde al diagrama de bloques de la figura 1. Las unidades acústicas pueden ser los denominados PLU´s ("Phone-Like Units") del idioma en cuestión, que para el caso del gallego hemos definido 26 unidades, o bien unidades dependientes del contexto: difonemas, trifonemas, semifonemas, que en nuestro caso hemos utilizado semifonemas tal como se describe en [2].

La experiencia en el trabajo con otras bases de datos nos dice que cuanto más se cuiden las tareas de transcripción, más fáciles y eficientes serán las tareas de entrenamiento de modelos acústicos y la evaluación de prestaciones. Aspectos a cuidar especialmente son todos los relacionados con el marcado de fenómenos paralingüísticos y ruidos de fondo e impulsivos.

El experimento que se ha diseñado para validar la utilidad/corrección del etiquetado realizado ha sido establecer una tarea de reconocimiento sobre un motor de reconocimiento de última generación. Este motor ha sido construido para funcionar con habla continua y grandes vocabularios.

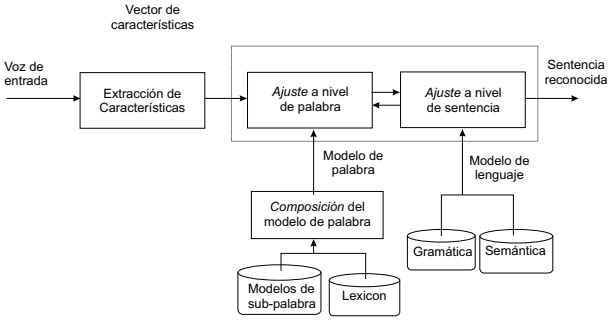

Figura 1: *Diagrama de bloques de un sistema de reconocimiento de habla basado en unidades inferiores a la palabra.*

5.1.1 Material para el entrenamiento de las unidades acústicas

La selección del material de voz utilizado para el entrenamiento de los modelos acústicos se debe realizar con bastante cuidado, ya que, la precisión y fiabilidad de los modelos determinará en gran medida las prestaciones del sistema de reconocimiento. En general, como material de entrenamiento se eligen grabaciones que se puedan considerar *"limpias"*, i.e., libres de ruidos de fondo, pronunciaciones erróneas, ruidos impulsivos, ..., por ello es muy importante que el etiquetado realizado sea preciso en cuanto a todos estos efectos.

En base a los criterios de transcripción impuestos por SpeechDat el material de entrenamiento no debería incluir ficheros que contengan alguna de las siguientes etiquetas: $**$, $[fill]$, $*[spk]$, $**[int]$, $[sta]$, $\sim$. Los ficheros que contienen las etiquetas "[int]" y "[spk]" se pueden utilizar siempre y cuando se consideren modelos que los representen. En la tabla 1 se muestra el número de ficheros que contienen alguno de los anteriores efectos en un conjunto de 1.349 frases etiquetadas. Se observa cuán frecuente es el fenómeno $[spk]$ producido por el locutor al hablar frente al microteléfono.

Si se seleccionan sólo aquellos ficheros libres de cualquier fenómeno paralingüístico nos encontramos con que el material de entrenamiento sería muy pequeño. Por ello, en nuestro caso, se seleccionarán ficheros *limpios* y ficheros con las etiquetas $[spk]$ e $[int]$ pues se consideran modelos específicos para estos dos fenómenos en el entrenamiento.

Teniendo en mente los semifonemas [2], a

Fenómeno	No. frases
[*sta*]	369
[*spk*]	1.094
[*int*]	275
[*fill*]	21
*[*spk*]	12
*[*int*]	14
**	14
*	222
~	12
&	10

Tabla 1: *Ocurrencias de algunas marcas de etiquetado*

priori el material diseñado en SpeechDat Galego contiene 740 semifonemas con lo que cubre el 86.5% de los semifonemas posibles. A esta alturas está etiquetado el 20% y se ha llegado al 80.2% de cobertura[1]. En la siguiente sección se muestran las prestaciones de estos modelos en comparación con los modelos obtenidos a partir de VOGATEL que también es una base de datos telefónica, recogida por Telefónica I+D para el diseño de reconocedores de habla en gallego [3].

5.1.2 Experimentos de reconocimiento

Se han realizado los experimentos de reconocimiento que detallamos a continuación.

El material que se ha validado está formado por un conjunto de 1.500 frases. Se han utilizado, en primer lugar, monofonemas y semifonemas entrenados a partir de la base de datos VOGATEL con voces masculinas y femeninas, y en segundo lugar monofonemas y semifonemas entrenados a partir de un conjunto de 2.700 frases de SpeechDat en gallego.

La parametrización utilizada incluye energía, 12 coeficientes cepstrales y sus derivadas primera y segunda. En total, para el caso de monofonemas se dispone de 26 modelos que representan fonemas y para el caso de semifonemas se dispone de 500 unidades. En ambos casos se dispone también de 5 modelos para distintos tipos de silencio y ruido.

Los experimentos se realizaron sobre un procesador Pentium III a 650 MHz, con 125 Mb de memoria RAM, sobre sistema operativo Linux. El motor de reconocimiento está diseñado para grandes vocabularios y habla continua con una estrategia de funcionamiento en dos pases. La primera consiste en una búsqueda en haz aplicando el algoritmo de Viterbi en forma síncrona. El espacio de búsqueda para esta primera fase se construye mediante la interconexión de una serie de nodos, cada uno de los cuales representa un fonema de una palabra del vocabulario y lleva asociado uno o varios modelos de Markov. Para reducir la complejidad, el vocabulario se organiza en forma de árbol. Una vez terminada la fase síncrona de reconocimiento, se realiza una segunda búsqueda teniendo en cuenta las palabras no propagadas pero almacenadas en los nodos especiales utilizando un algoritmo de tipo *A* Stack*.

Se han realizado dos experimentos de reconocimiento. En el primero se utiliza un modelo de lenguaje ideal, en el sentido de que éste se extrae del macrotexto constituido por las transcripciones que forman el corpus. El vocabulario consta así de 5.949 palabras, y el modelo de lenguaje está formado por 5.951 unigramas, 16.283 bigramas y 20.509 trigramas. La tabla 2 muestra los resultados obtenidos por cada uno de los conjuntos de modelos acústicos utilizados.

Vogatel		SpeechDat	
Monof	Semif	Monof	Semif
89.52%	89.26%	90.71%	96.5%

Tabla 2: Resultados de reconocimiento utilizando diferentes modelos acústicos y un modelo de lenguaje ajustado a la tarea.

Se observa como los mejores resultados son los proporcionados por los modelos entrenados con material de SpeechDat ya que, entre las bases de datos Vogatel y SpeechDat existe un *"desajuste"*, y las mejores prestaciones de un reconocedor se consiguen siempre en condiciones de *"ajuste"* entre los datos de entrenamiento y de evaluación. Además, se observa como los semifonemas funcionan apreciablemente mejor para el caso de "ajuste", mientras que no se aprecia un mejor comportamiento en el caso de "desajuste".

Para el segundo experimento se creó un modelo de lenguaje de texto periodístico, al que se le añadieron las transcripciones del corpus para reducir la perplejidad. El resultado fue un modelo de lenguaje formado por 22.395 unigramas, 198.953 bigramas y 154.470 trigramas. En este caso la per-

[1]No se ha considerado la coarticulación entre palabras

plejidad de la tarea resulta bastante baja (120.75), aunque como contrapartida hay un 29% de palabras fuera de vocabulario. La tabla 3 muestra los resultados obtenidos por cada uno de los conjuntos de modelos acústicos utilizados.

Vogatel	SpeechDat	
Monof	Monof	Semif
47.63%	53.19%	66.78%

Tabla 3: Resultados de reconocimiento utilizando diferentes modelos acústicos y un modelo de lenguaje general.

Se observa como al igual que en el anterior experimento los semifonemas superan en prestaciones a los monofonemas, y como los modelos de Vogatel se ven superados por los de SpeechDat debido al *"desajuste"* existente entre las bases de datos Vogatel y SpeechDat. También se comprueba, a partir de las tasas de reconocimiento logradas, como esta tarea es mucho más "difícil" a consecuencia de: la longitud del vocabulario y el modelo de lenguaje "general".

6 Conclusiones

En este artículo se han descrito las tareas necesarias para recoger y documentar una base de datos oral en gallego utilizable en el diseño y evaluación de sistemas de reconocimiento automático de habla.

En primer lugar, se ha tratado el tema de la captación de llamadas. En esta tarea el objetivo es obtener una representación equilibrada de la población que permita crear una base de datos que englobe todas las variantes dialectales del Gallego.

En segundo lugar, se ha analizado la problemática que engloba la tarea de etiquetado manual de las llamadas, lo que nos ha llevado a incluir desviaciones de las especificaciones de etiquetado impuestas por SpeechDat. Estas desviaciones consisten principalmente en la inclusión de nuevas etiquetas para tener en cuenta fenómenos paralingüísticos.

Por último se han realizado una serie de experimentos preliminares de reconocimiento de habla para la validación del etiquetado manual realizado. Los resultados obtenidos muestran la importancia de modelar las coarticulaciones y la necesidad de que exista un "ajuste" entre las bases de datos utilizadas para el entrenamiento de los modelos y para la evaluación.

Referencias

[1] http://www.speechdat.org

[2] J.B. Mariño, A. Nogueiras, A. Bonafonte. *"The demiphone: an efficient subword unit for continuous speech recognition"*. In Proceedings of EUROSPEECH'97. pp. 1215–1218.

[3] L. Villarrubia, P. León, L. Hernández, J.M. Elvira, C. Nadeu, I. Esquerra, J. Hernando, C. García-Mateo, L. Docío. *"VOCATEL AND VOGATEL: Two Telephone Speech Databases of Spanish Minority Languages (Catalan and Galician)"*. Workshop on Language Resources for European Minority Languages. Granada(Spain). May 1998.