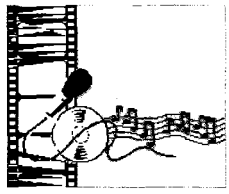

RECONOCIMIENTO DE VOZ MULTIDIALECTAL ESPAÑA - COLOMBIA

Mónica Caballero Galeote, Asunción Moreno Bilbao

Departamento de Teoria de la Señal y Comunicaciones
Centre de Tecnologies Aplicades al Llenguatge i la Parla (TALP)
Universidad Politécnica de Cataluña
almcg@gps.tsc.upc.es , asuncion@gps.tsc.upc.es

1. INTRODUCCIÓN

El idioma español se encuentra entre los idiomas más hablados en todo el mundo. Se habla en España y en toda Latinoamérica, exceptuando Brasil por más de 300 millones de personas. También es verdad que debido a la gran dispersión geográfica de estas zonas, el español de cada zona difiere y se establecen variedades dialectales.

Es por eso, que al plantearse un sistema de reconocimiento del habla para el castellano, parece interesante no limitarse a un solo dialecto, sino intentar crear un sistema multidialectal.

Más aún si se observa que para el castellano canónico o estándar, el hablado en España, se tienen bases de datos de habla de 1000 y 4000 locutores, que son bastantes recursos. Pero hay muchos dialectos del español de Latinoamérica de las que sólo se tienen pequeñas bases de datos, como es el caso del dialecto hablado en Colombia. Lo que se plantea en este proyecto es el poder aprovechar todos los conocimientos y datos que se puedan extraer de una gran base de datos para mejorar sistemas de reconocimiento del castellano hablado en otras zonas, con otras características fonéticas, con diferencias léxicas, pero a fin de cuentas, el mismo idioma.

Este estudio es una aproximación al sistema multidialectal que se plantea y se trabaja con un solo dialecto latinoamericano, el hablado en Colombia y el castellano hablado en España.

En el capítulo 2 se describirá el sistema de reconocimiento que se ha utilizado, así como también se tratará del tema relacionado con la fonética, la problemática y la resolución de la transcripción para atacar al sistema de reconocimiento. En el capítulo 3 se dan más detalles de las bases de datos que se han utilizado y finalmente en el capítulo 4 se describen los experimentos que se han realizado y los resultados obtenidos. En el capítulo 5 se comentan las conclusiones y algunas líneas de trabajo futuras.

2. SISTEMA DE RECONOCIMIENTO

2.1. Descripción del problema fonético

La representación de las variaciones dialectales es muy importante para el buen funcionamiento de un sistema de reconocimiento. En este caso, nos encontramos delante de una variación del español hablado en Colombia. Se ha hecho un estudio de las reglas de transcripción grafema-fonema del español de Bogotá, la zona más poblada de Colombia, dónde se habla el dialecto propio de las montañas o andino, así como también del castellano estándar hablado en España [1].

Las transcripciones se han realizado en símbolos SAMPA (Speech Assessment Methods Phonetic Alphabet) y algunos símbolos de SAMPA extendido para representar sonidos típicos colombianos, como el caso de /h/.

Las reglas para la transcripción del colombiano se basan en la transcripción del castellano estándar. Así, respecto a éste caracterizan el dialecto colombiano los siguientes fenómenos:

a) Pronunciación de /T/ siempre como /s/, el "seseo".
b) La /L/ siempre se pronuncia como /jj/, el "yeísmo".
c) La velar fricativa /x/ se realiza como la glotal /h/.
d) Los sonidos /b/, /d/ y /g/ se pronuncian siempre como consonantes oclusivas excepto cuando se encuentran en posición post-nuclear o en el inicio de una sílaba siguiendo a una vocal. En estos casos se transcriben como los alófonos aproximantes /B/, /D/ y /G/.

Las vocales en castellano representan aproximadamente un 50 % del total de alófonos y éstas no presentan grandes diferencias de pronunciación en los diferentes dialectos.

Dadas estas características se toman estas decisiones a la hora de transcribir:

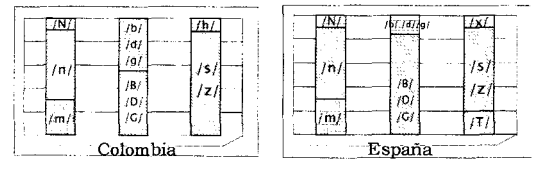

Figura 1. *Proporción de los fonemas nasales /m/, /n/ y /N/, de las consonantes /b/, /, /d/ y /g/ y las aproximantes /B/, / D/ y /G/ y de la presencia de los fonemas /s/ +/z/, /T/, /x/ y /h/ en los dialectos castellano y colombiano.*

Para el castellano: En castellano casi no se realiza la /N/. Por lo que se asimila la /N/ a la /n/. Igual pasa con los sonidos /b/, /d/ y /g/ que se asimilan a los alófonos aproximantes /B/, /D/ y /G/. También se considera que los sonidos /L/ y /jj/ están demasiado cercanos y se decide asimilarlos a uno común que se etiqueta como /y/.

Para el colombiano: Se transcribe siempre la /T/, la /L/ y la /x/ como /s/, /jj/ y /h/. En cambio se diferencian / b/ /, /d/ y /g/ de /B/, /D/ y /G/. Y al igual que en el castellano se asimila la /N/ a la /n/.

2.2. Descripción general del sistema

En este estudio se ha trabajado con un sistema de reconocimiento desarrollado por el grupo de procesado del habla del centro de Tecnologies aplicadas al llenguatge i la parla (TALP) de la UPC llamado *RAMSES* [2]. Este sistema se basa en el entrenamiento de modelos ocultos de Markov de unidades fonéticas. Este tipo de sistemas se presentan como unos de los más robustos debido a su capacidad de retener información estadística de los patrones de voz.

La estructura general de este tipo de sistemas es la que se presenta en la figura 2, a continuación:

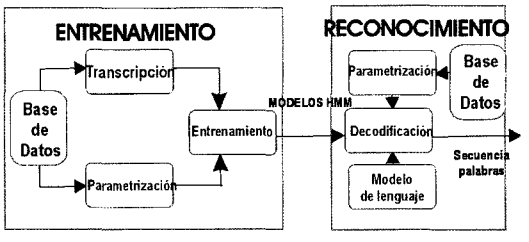

Figura 2. Sistema de reconocimiento del habla.

Ramses está compuesto por una serie de programas que permiten el entrenamiento de modelos de Markov continuos y semicontinuos, (en este estudio se crearan sólo modelos semicontinuos), y el reconocimiento de señal de voz a partir de estos. Ramses permite trabajar con un gran tipo de unidades fonéticas: palabras, fonemas, semifonemas incontextuales y semifonemas por umbral. Para obtener una primera aproximación al programa se trabaja con fonemas. El paso a semifonemas después resulta más fácil.

Debido a que estos programas ofrecen una gran variedad de opciones, existen unos scripts, que facilitan el uso de estos programas en los casos más habituales.

La forma en que se ha utilizado Ramses en este proyecto se explicará a continuación.

- **Transcripción :** Creación de ficheros de marcas que contienen la segmentación en unidades fonéticas, ya sean sólo en palabras (es el caso de las señales de test) o en palabras y fonemas (caso de las señales que participarán en el entrenamiento). Los ficheros de marcas se construyen a partir de los ficheros existentes en la base de datos junto a las señales que contienen la transcripción ortográfica de la señal e información del locutor.
  Es posible personalizar la transcripción inicial de forma que se adapte mejor a un propósito determinado. Sólo se ha de añadir un diccionario de sustituciones de las unidades fonéticas. Por ejemplo, no se van a crear modelos diferentes de las vocales tónicas respecto a las átonas. Al crear las marcas se sustituyeron todas las vocales acentuadas por las vocales sin acentuar. De esta forma se pueden asimilar fonemas de los cuales no se quieren calcular modelos.

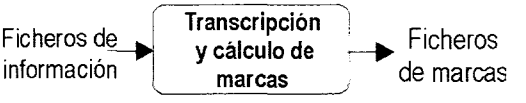

Figura 3. Transcripción.

- **Parametrización, cálculo del codebook y cuantificación:** La extracción de características consiste en una estimación espectral de la señal. Se usa una parametrización mel-cepstrum [3]. De esta parametrización se obtienen unos coeficientes que se han de cuantificar vectorialmente, por lo que se necesita calcular un diccionario de vectores significativos o codebook. Se calcula un codebook de 128 coeficientes o vectores diferentes y a continuación se cuantifica.

  También se puede realizar la cuantificación con estructura pipeline evitando así almacenar los datos intermedios si ya se tiene el codebook calculado.

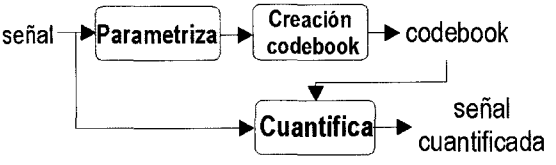

Figura 4. Parametrización, cálculo codebook y cuantificación.

- **Entrenamiento de modelos de Markov:** Para evitar que partes de señal donde hay silencios contribuyan al modelo, hay que determinar los tiempos donde se encuentran los silencios, tanto el inicial y el final, como los posibles silencios que se puedan encontrar en medio de la señal. Para ello se utiliza un programa que implementa el algoritmo de Viterbi para reconocimiento del habla, es decir busca el camino más probable. Se le llama con un fichero de arquitectura que le indica el grafo de búsqueda. Según la arquitectura con la que se le llame, varia su funcionalidad. En el entrenamiento de modelos se utilizará para alinear las señales. Se alinean las palabras que forman la señal de forma temporal, permitiendo la inserción de silencios en medio de la señal.
  Se presentan aquí dos posibles situaciones que varían el esquema en que se entrenan los modelos de Markov.

  1) Si se parte de cero, se crean unos modelos iniciales que contendrán partes de silencio y se alinean las frases. Con dos iteraciones de este proceso, se puede considerar que los silencios ya quedan bien definidos. Finalmente, se desprecian estos modelos y se entrenan los modelos finales.

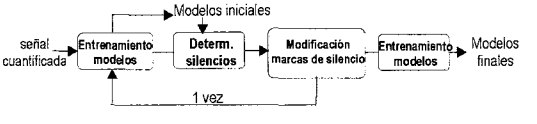

Figura 5. Det. de silencios y entrenamiento hmm's.

2) Si ya se tienen unos modelos iniciales y lo que se pretende es reentrenarlos, entonces basta con un alineado inicial de las señales del entrenamiento y entrenar los modelos finales que se utilizarán en el reconocimiento.

- **Modelado del lenguaje:** Es aquí donde se explota el conocimiento sobre el lenguaje. La forma en que utiliza aquí Ramses viene determinada porque la tarea a reconocer es un cuerpo finito de palabras. A nivel léxico, se construye un diccionario con la transcripción de las palabras a reconocer en las unidades fonéticas de las cuales se han creado los modelos de Markov. Así "casa" tiene la representación *k a s a*. En la transcripción no se diferencian las unidades tónicas de las átonas. Como la estructura del mensaje a reconocer es sencilla, con un autómata de estados finitos (FSA, gramáticas N-1 grama) basta para modelar la sintaxis del mensaje. La gramática se genera a partir de un fichero de texto donde han de estar listadas las palabras a reconocer. El script que calcula el diccionario lo puede hacer a partir de un corpus, de un fichero de texto o de una gramática, tal como se ha utilizado.

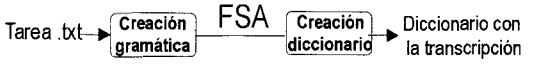

***Figura 6.** Modelado del lenguaje.*

- **Reconocimiento:** Una vez entrenados los modelos (hmm's) y creada ya una gramática y un diccionario, ya podemos pasar a esta etapa, también denominada decodificación.
  En primer lugar se han de cuantificar vectorialmente las señales a reconocer. Después se ha de determinar la secuencia de palabras que con mayor probabilidad ha generado una secuencia de observaciones. Para esta labor se utiliza el algoritmo de Viterbi.. El grafo de búsqueda, o lo que es lo mismo, la arquitectura, se construye de forma jerárquica: en el nivel más alto, la gramática, que indica las palabras a reconocer y las relaciones entre ellas, por debajo se tiene la representación de las palabras en unidades fonéticas y por último, se dispone de hmm's que relacionan las observaciones acústicas con los fonemas.
  Ramses, además dispone de scripts que calculan las tasas de aciertos y que muestran que señales no se han reconocido.

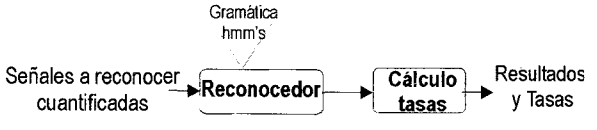

***Figura 7.** Reconocimiento y evaluación.*

3. BASES DE DATOS

Para este estudio se han utilizado 2 bases de datos, la base SpeechDat para castellano y la base de locutores colombianos que se incluye en el proyecto SALA [4].

El proyecto SALA consiste en la creación de bases de datos de los principales dialectos del castellano en Latinoamérica. Los criterios para la creación de estas bases, entre la que se encuentra la del colombiano son los mismos que los que se siguieron para la creación de SpeechDat en España.

Las especificaciones que debían seguir estos corpus son las siguientes:

- Cada locutor pronuncia más de 40 frases, leídas o espontáneas.
- La parte del corpus de habla leída contiene palabras de aplicación, frases con palabras clave, números, dígitos, fechas, expresiones relativas al tiempo, nombres, fechas, empresas, frases y palabras fonéticamente balanceadas.
- La parte del material espontáneo incluye preguntas si/no, ciudades, horas y fechas.

A parte de estas especificaciones más relacionadas con el contenido del corpus existen otras que también se han tenido en cuenta en las dos bases, tales como:

- Factores dialectales
- Características del locutor: edad, sexo
- Características específicas del entorno.

Ambas bases se han grabado sobre la red telefónica fija, mediante una línea RDSI.

Las señales quedan almacenadas en ficheros tras ser muestreadas a 8 khz, cuantificadas con 8 bits y comprimidas con la ley A. Cada señal se guarda en un fichero distinto. A cada fichero le acompaña otro con información del locutor y la transcripción. Esta transcripción se ha hecho con los símbolos SAMPA.

La base SpeechDat está compuesta por grabaciones de 4000 locutores.

La base colombiana englobada dentro del proyecto SALA está compuesta por grabaciones de 1000 locutores.

Ambas bases tienen especificados un conjunto de señales para el test. En el caso de SpeechDat se dedican 500 locutores para el set de test, en el caso de la base colombiana son 200 locutores los que componen este conjunto de test.

4. EXPERIMENTOS Y RESULTADOS

Se han llevado toda una serie de experimentos a fin de mejorar la tasa de reconocimiento tanto en el castellano, como en el colombiano.

4.1. Ficheros entrenamiento y test

Se pretende evaluar el sistema reconociendo palabras de aplicación y dígitos aislados. Para éste propósito se decide entrenar con frases fonéticamente balanceadas y ampliar el entrenamiento con palabras fonéticamente balanceadas.

Entrenamiento. En Colombia se tienen 2700 frases pertenecientes a 800 locutores y 1550 palabras, también pertenecientes a los 800 locutores. En castellano, se dispone de un total de 15156 frases y 9174 palabras pertenecientes a 3500 locutores.

Test. Se diseñan dos tests, uno para el español y otro para el colombiano con las palabras de aplicación y dígitos aislados. En colombiano se tienen un total de 751 ficheros de test de 200 locutores. En Castellano, aunque se dispongan de 500 locutores para realizar el test, sólo se utilizan 200 locutores, al igual que en el caso del colombiano, para así tener unos resultados comparables. En total se dispone de 626 palabras pertenecientes a los 200 locutores.

En total se entrenan un total de 26 fonemas en castellano y de 30 en colombiano y se reconocen un total de 30 palabras de aplicación más los 10 dígitos tanto en castellano como en colombiano.

4.2. Experimentos y resultados

En primer lugar, se evaluó cada sistema por separado. Esto es, entrenar y reconocer con castellano por una parte, y por otra, entrenar y reconocer colombiano. Se ha trabajado con fonemas.

Para el caso del castellano, se partió del mismo número de frases que se disponían en el caso colombiano, para tener un resultado comparable, y posteriormente se fueron añadiendo locutores poco a poco, de modo que se reflejara la mejora al aumentar el material de entrenamiento. Cada vez que se aumentó el número de frases, se usaban como modelos iniciales los que resultaron del anterior entrenamiento. Se pudo comprobar que una vez alcanzados un número suficiente de frases, el sistema no mejoraba notablemente, por lo que se decidió, después de entrenar con 4327 frases añadir a las frases las palabras fonéticamente balanceadas. Esto hizo mejorar la tasa de reconocimiento en más de un punto. Por este motivo se siguió reentrenando el sistema con más locutores, incluyendo además las palabras. Para el test escogido, este incremento de entrenamiento no se notó en el resultado del reconocimiento, quedándose en una tasa de 94,4 %.

En el caso del colombiano, se entrenó con las 2700 frases, dando un resultado de un 94 % de aciertos, resultado muy bueno, considerando el poco número de frases de las que se disponía. Al añadir las palabras, difiriendo del caso del castellano, el sistema no mejoraba. Se decidió alinear las señales con estos últimos modelos y volver a crear modelos nuevos sólo a partir de las frases. De esta manera, la tasa de reconocimiento aumentó, aunque en menor grado, llegando a un 94,14 %.

Por otra parte, se diseñaron una serie de experimentos de reconocimiento cruzado. A la vez que se iban obteniendo modelos nuevos del castellano, se les aplicaba el test colombiano, observándose una buena evolución. De esta forma, también se comprobó que el sistema, aunque se quedase estancado con el test castellano, mejoraba al aplicarle el colombiano.

Otra cosa a denotar es que los resultados tan parecidos al del sistema entrenado en castellano y reconociendo en el mismo dialecto, indican que los canales son parecidos.

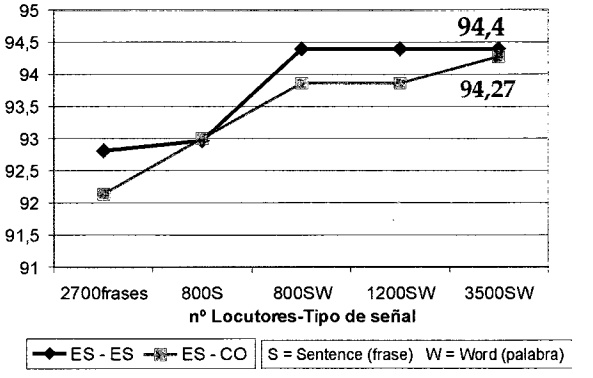

Figura 8. *Evolución tasas reconocimiento test castellano ES-ES y test colombiano ES-CO entrenando con base castellana.*

5. CONCLUSIONES

Este estudio demuestra que es posible conseguir unas buenas tasas de reconocimiento utilizando los modelos entrenados con la base de datos española. Quedan muchos experimentos por hacer. Los más inmediatos son, trabajar con las dos bases de datos a la vez, trabajar con semifonemas y trabajar más con la transcripción fonética. Sería interesante separar también en castellano los modelos de las oclusivas b, d y g para tener modelos separados para aplicar al sistema el test colombiano.

6. AGRADECIMIENTOS

Este trabajo ha sido subvencionado por el proyecto TIC-2000-1005-C03.

7. REFERENCIAS

[1] A. Moreno, J. B. Mariño. *Speech Dialects: Phonetic Transcription.*International Conference on Spoken Language Processing. Sidney, Australia. Dec,1998.

[2] A. Bonafonte, J. B. Mariño, A. Nogueiras, J.A. Rodríguez Fonollosa. *RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC*, VIII Jornadas de Telecom I+D (TELECOM I+D'98), Madrid.

[3] J.W. Picone *Signal Modeling Techniques in Speech Recognition.* Proceedings of the IEEE Vol 81 no9. September 1993

[4] A. Moreno, H. Höge, J. Koehler, J. B. Mariño. *Project SALA, SpeechDat Across Latin America.* First International Conference on Language Resources and Evaluation, LREC'98, Granada.