Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 145-152 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

# Incorporación de n-gramas discriminativos para mejorar un reconocedor de idioma fonotáctico basado en i-vectores

## *Incorporation of discriminative n-grams to improve a phonotactic language recognizer based on i-vectors*

**Christian Salamea Palacios[1,2], Luis Fernando D'Haro[1], Ricardo Córdoba[1], Miguel Ángel Caraballo[1]**

[1]Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica
E.T.S.I. Telecomunicación. Universidad Politécnica de Madrid.
Ciudad Universitaria S/N, 28040 - Madrid, España.
{csalamea, lfdharo, cordoba, macaraballo}@die.upm.es

[2]Universidad Politécnica Salesiana del Ecuador
Calle Vieja 12-30 y Elia Liut, Casilla 26, Cuenca, Ecuador
csalamea@ups.edu.ec

**Resumen:** Este artículo describe una nueva técnica que permite combinar la información de dos sistemas fonotácticos distintos con el objetivo de mejorar los resultados de un sistema de reconocimiento automático de idioma. El primer sistema se basa en la creación de cuentas de posteriorgramas utilizadas para la generación de i-vectores, y el segundo es una variante del primero que tiene en cuenta los n-gramas más discriminativos en función de su ocurrencia en un idioma frente a todos los demás. La técnica propuesta permite obtener una mejora relativa de 8.63% en $C_{avg}$ sobre los datos de evaluación utilizados para la competición ALBAYZIN 2012 LRE.
**Palabras clave:** Posteriorgrama, i-Vectores, rankings discriminativos, fonotáctico, n-gramas.

**Abstract:** This paper describes a novel technique that allows the combination of the information from two different phonotactic systems with the goal of improving the results of an automatic language recognition system. The first system is based on the creation of posteriorgram counts used for the generation of i-vectors, and the second system is a variation of the first one that takes into account the most discriminative n-grams as a function of their occurrence in one language compared to all other languages. The proposed technique allows a relative improvement of 8.63% on $C_{avg}$ over the official set used for the ALBAYZIN 2012 LRE evaluation.
**Keywords:** Posteriorgram, i-Vectors, discriminate rankings, phonotactic, n-grams

## *1 Introducción*

El presente artículo describe una técnica novedosa que permite mejorar las tasas de reconocimiento de idioma mediante la unificación de dos técnicas que emplean información a nivel fonotáctico a partir de la salida de un reconocedor de fonemas que permite determinar las secuencias más probables de éstos para un determinado conjunto de ficheros de audio. En este caso, nos hemos decantando por el uso de técnicas fonotácticas, ya que son ampliamente usadas en el reconocimiento de idioma y/o locutor por las ventajas que presentan, su versatilidad, la posibilidad de incorporar información a alto nivel y el hecho que de forma congruente siempre permite mejorar las tasas de reconocimiento cuando se combina con otras técnicas basadas únicamente en información acústica (Kinnunen and Li, 2010).

La técnica que proponemos, aparte del uso de la información fonotáctica, aprovecha un elemento en común entre las dos técnicas que

ISSN 1135-5948 

hacen posible su unificación: son las llamadas "cuentas" que se calculan a partir de la ocurrencia de la secuencia de fonemas (i.e. n-gramas) reconocidos mediante el reconocedor automático de fonemas. En el caso de la primera técnica, esta utiliza las cuentas con el objetivo de entrenar un modelo basado en i-vectores que es la técnica actual que mejores resultados da en reconocimiento tanto usando información acústica como fonotáctica (Dehak et al, 2011)(Martinez et al, 2011)(D'Haro et al, 2012). Este sistema fonotáctico basado en i-vectores en combinación con otros sistemas basados en información acústica fue uno de los factores determinantes en la obtención de los buenos resultados conseguidos durante la evaluación de reconocimiento de idioma Albayzin 2012 LRE, tal y como se describe en (D'Haro et al, 2013).

Por otra parte, en el caso de la segunda técnica, se utilizan las cuentas de los fonemas reconocidos con el objetivo de crear un ranking de los n-gramas más discriminativos para reconocer un idioma frente a los otros. El proceso de creación de los rankings implica la estimación de un valor de discriminación que se utiliza como factor de ordenación de los rankings. Esta técnica también se ha probado previamente con muy buenos resultados (Cordoba et al, 2007) en el contexto de un sistema de reconocimiento que utiliza múltiples reconocedores de fonemas (Zissman, 1996).

En este artículo proponemos una nueva técnica en la que se modifican los valores de las cuentas de posteriorgramas usados por el primer sistema en la generación de los i-vectores mediante la utilización de la información de discriminación de los n-gramas generados para crear los rankings de la segunda técnica. Conviene mencionar que como figura de mérito sobre la eficacia de la técnica propuesta hemos utilizado los mismos datos de la evaluación Albayzin utilizando tanto la métrica oficial de la evaluación (con el objetivo de facilitar la comparación de resultados) como la métrica $C_{avg}$ que es una de la más empleada en las evaluaciones de reconocimiento de idioma. La métrica $C_{avg}$ permite ponderar los errores de falsa aceptación (i.e. reconocer un determinado fichero con un idioma distinto al que realmente es) y el falso rechazo (i.e. no reconocer el idioma real de un determinado fichero).

Este artículo se organiza de la siguiente manera. En la sección 2 se hace una breve descripción del sistema base empleado. Posteriormente se describe cada uno de los dos sistemas empleados por separado, para terminar describiendo la técnica propuesta. Luego, en la sección 3 describimos la base de datos y metodología seguida para la realización de los experimentos. En la sección 4 presentamos y discutimos los resultados obtenidos. Y finalmente, en la sección 5, presentamos las conclusiones y líneas futuras.

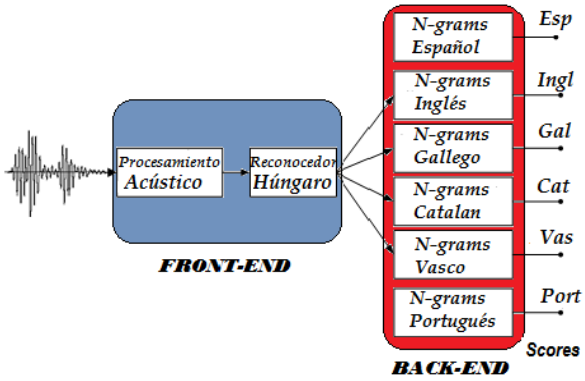


Figura 1: Sistema de reconocimiento de lenguaje basado en PRLM.

## 2 Descripción de los sistemas

Una gran mayoría de los sistemas de reconocimiento automático de idioma que emplean información fonotáctica lo hacen aplicando una técnica denominada PRLM (Phone Recognition Followed by a Language Model), en el que, como indica la Figura 1, se tienen dos componentes claramente diferenciados: uno denominado Front-End y otro denominado Back-End.

En el primero se realiza la parametrización de los ficheros de audio de entrada y se ejecuta un reconocedor automático de fonemas que se encarga de determinar la secuencia de fonemas más probables. Dado que esta salida contendrá errores de reconocimiento los resultados finales serán menos buenos, pero tiene dos grandes ventajas: 1) que los modelos utilizados por el reconocedor de fonemas no tienen por qué corresponder con los idiomas a reconocer (aunque evidentemente se obtienen mejores resultados cuando hay correspondencia), y 2) porque esto permite reaprovechar reconocedores mucho mejor entrenados, a la vez que se minimiza la necesidad de disponer de una gran cantidad de datos etiquetados para el entrenamiento de dichos reconocedores.

Para los experimentos presentados en este artículo hemos utilizado como Front-End el reconocedor de fonemas de la Universidad de

Brno (Schwarz, 2009), el cual se distribuye libremente e incluye modelos de fonemas de 3 idiomas distintos: checo, húngaro y ruso. En nuestro caso, hemos utilizado únicamente el modelo de fonemas de Húngaro dado que con éste se han conseguido resultados satisfactorios en experimentos previos (D'Haro et al, 2012), además de que fue el que se utilizó durante la evaluación de Albayzin LRE 2012. Este reconocedor permite identificar un total de 61 clases de fonemas aunque para nuestros experimentos hemos reducido este número a un total de 33, iguales a los enumerados en (Diez et al, 2013), unificando tres fonemas que permiten detectar ruidos y pausas en el habla, así como otros con un gran parecido lingüístico.

Por otra parte, en el Back-End se toman las secuencias de fonemas reconocidos y se entrena un modelo de lenguaje por cada idioma a reconocer. Estos modelos se utilizan uno a uno durante la fase de evaluación para calcular la perplejidad de la frase a identificar, siendo finalmente el clasificador quien decide cuál es el idioma reconocido en función al modelo que presente la menor perplejidad.

La diferencia entre el sistema PRLM y el utilizado en este trabajo para el reconocimiento radica principalmente en el Back-End, dado que no se utilizan las "cuentas" de los n-gramas para generar modelos de lenguaje, sino que las utilizamos para generar dos modelos de idioma distintos que luego se combinan. Por una parte, el primer sistema se crea a partir de los i-vectores generados para los ficheros de entrenamiento, mientras que el segundo utiliza las cuentas para crear un ranking de n-gramas discriminativos junto con su valor de discriminabilidad al comparar las ocurrencias de los n-gramas en un idioma frente a los otros. Posteriormente, los dos sistemas se unifican modificando las cuentas usadas por el primer sistema mediante el valor de discriminabilidad calculado en el segundo sistema. Finalmente, como clasificador hemos utilizado un sistema basado en regresión logística que utiliza como entrada los i-vectores reconocidos y les asigna una puntuación (score) según la similitud de cada i-vector con los diferentes modelos de idioma entrenados.

## 2.1 El sistema fonotáctico basado en cuentas de posteriorgramas

La creación de las cuentas de posteriorgramas se describe en los siguientes pasos (Figura 2):

a. El primer paso consiste en extraer los valores de las probabilidades a posteriori de cada uno de los posibles fonemas a reconocer para cada trama. Estos valores se obtienen directamente del reconocedor de fonemas. En la figura podemos ver que para cada trama del fichero de audio se obtienen 3 valores correspondientes a los 3 posibles fonemas a reconocer (i.e. hemos usado 3 fonemas para simplificar el ejemplo, aunque en el sistema real fueron 33 fonemas distintos).
b. En un segundo paso se suman y se promedian las probabilidades a posteriori de todas las tramas que se consideran que pertenecen a la misma unidad fonética. Esta agrupación de las tramas en fonemas es realizada también por el reconocedor de fonemas empleando el algoritmo de Viterbi sobre las probabilidades a posteriori obtenidas en el paso anterior.
c. El tercer paso es calcular las probabilidades condicionales de que ocurra un determinado fonema considerando los n-1 fonemas previos (i.e. n-gramas). Para ello, en el caso de usar bigramas, como se muestra en la figura, se realiza el producto exterior (outer-product) entre las probabilidades promediadas del fonema previo con las del fonema actual. Para solventar el problema de la primera trama se crea un fonema tipo "dummy" en el que la todos los fonemas son equiprobables.
d. El cuarto paso consiste en sumar todas las matrices producto generadas antes a lo largo de todo el fichero cuidando de sumar adecuadamente los mismos contextos (i.e. la probabilidad condicional $p_{ij}(t-1)$ con la probabilidad $p_{ij}(t)$). El resultado es lo que denominados cuentas de posteriorgrama condicionales.

Por último se convierte la matriz de cuentas de posteriorgramas en un supervector de dimensión [ $F^n$ x 1 ], donde F es el número de fonemas a emplear y n es el orden de los n-gramas. En nuestro caso, al emplear 33 fonemas y usar bigramas obtenemos un vector de dimensión 1089, y en el caso de trigramas tenemos un vector de dimensión 35937. Estos supervectores se crean para cada fichero y para cada uno de los idiomas a reconocer. Luego se utilizan en el entrenamiento de los i-vectores.

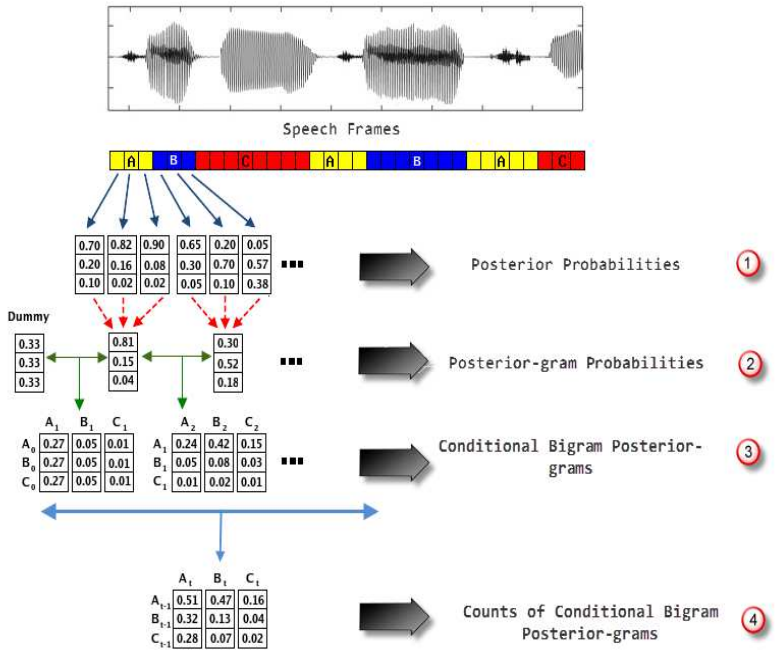


Figura 2. Procedimiento de creación de las cuentas de posteriogramas.

Una vez obtenidas las cuentas de posteriorgramas para todos los ficheros, se procede a modelar las probabilidades globales de ocurrencia de cada una de los n-gramas para posteriormente calcular los i-vectores usando un modelo de subespacios mutinomiales (SMM, Subspace Multinomial Model) propuesto por (Povey, 2010), que permite entrenar vectores de baja dimensión (i-vectores) en el subespacio de la variabilidad total para luego usarlos como vectores de características en el entrenamiento de un clasificador discriminativo de reconocimiento de lenguaje.

El entrenamiento de los i-vectores se realiza mediante el método de estimación y maximización (EM) y la optimización se lleva a cabo aplicando el método de Newton-Raphson. Para mayores detalles acerca de las formulaciones matemáticas de los SMM y su aplicación en sistemas de reconocimiento de idioma se recomienda la lectura de (Kockmann et al, 2010) y (Soufifar et al, 2011). Por otra parte, la formulación matemática de los i-vectores (Dehak et al, 2011) se realiza empleando la siguiente relación matemática:

$$m_x = M + T\omega_x \qquad (1)$$

Donde $m_x$ es un vector de dimensión [ $F^n$ x 1] que contiene las medias de las características que se modelan para un fichero determinado *x*. En nuestro caso, a partir de las cuentas de los supervectores de los posteriorgramas hallamos las probabilidades medias de ocurrencia de cada fonema para cada fichero. Por otra parte, la $M$ es un vector de dimensión [ $F^n$ x 1 ], que contiene las medias globales independientes de idioma mejor conocido como UBM (i.e. Universal Background Model). La $T$ se conoce como extractor de i-vectores y tiene una dimensión de [ $F^n$ x r ] donde r es un valor de baja dimensionalidad que se selecciona de tal forma que r << $F^n$. Por último, la $\omega_x$ son los i-vectores en sí y tienen una dimensión de [r x 1].

El procedimiento para entrenar la matriz T y crear los i-vectores es un proceso iterativo en el que se parte de unos valores de i-vectores inicializados aleatoriamente para con ellos obtener la matriz T; luego, con esta T se regeneran los i-vectores y a partir de los nuevos se vuelve a crear una nueva matriz T, y así sucesivamente. El proceso se detiene cuando entre iteración e iteración no se obtiene una reducción en la verosimilitud global del modelo. Tal como se ha comentado previamente, la gran ventaja de los i-vectores es la posibilidad de trabajar con vectores de baja dimensión ya que esto reduce los problemas de dispersión de datos y facilita el entrenamiento del clasificador. En nuestro caso hemos trabajado con i-vectores de dimensión 400 ya que con ellos se obtuvieron los mejores resultados durante la competición oficial.

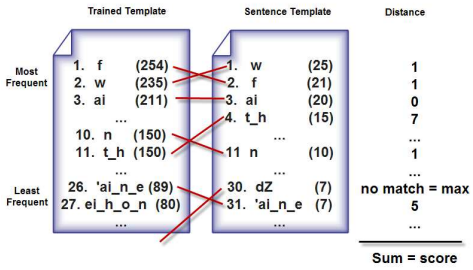


Figura 3. Categorización de texto en base de la ocurrencia de n-gramas

## 2.2 El sistema fonotáctico basado en rankings discriminativos de n-gramas

Este sistema se basa en el uso de una técnica de categorización de textos propuesta por (Cavnar y Trenkle, 1994) que permite combinar información local (i.e. n-gramas) e información de largo alcance (i.e. las cuentas de n-gramas recogidas a lo largo de una frase). En la Figura 3 se muestra a modo de ejemplo la técnica original en la que se propone durante la fase de entrenamiento la creación de una plantilla con los n-gramas más frecuentes (típicamente los primeros 400) empleando hasta un orden de 5-gramas, obtenidos a partir de las secuencias de caracteres (i.e. fonemas reconocidos en nuestro caso) de los ficheros de entrenamiento para cada idioma y ordenados por su ocurrencia de mayor a menor. Durante la fase de evaluación, se crea una plantilla dinámica a partir de la frase reconocida y se ordena siguiendo el mismo procedimiento que en la fase de entrenamiento. Para realizar la detección del idioma se suma la diferencia absoluta entre las posiciones de los n-gramas de las dos plantillas utilizando la ecuación 2.

$$d^T = \frac{1}{L}\sum_{i=1}^{L} abs(pos\ w_i - pos\ w_i^T) \qquad (2)$$

Donde L es el número de n-gramas generados para la frase a reconocer. Para aquellos n-gramas que no aparecen en los rankings entrenados se aplica una penalización en función del tamaño de la plantilla. Finalmente, el idioma reconocido es aquel que presente la mínima distancia entre las dos plantillas.

En (Cordoba et al, 2007) presentamos diferentes mejoras a la técnica original. Las más importantes fueron:

- Aplicamos lo que denominamos "posición golf" en la que para aquellos n-gramas cuyo número de ocurrencias sean iguales se ubican en el ranking dentro de la misma posición; tal como en el golf, donde aquellos jugadores que tengan el mismo número de golpes ocupan la misma posición en el ranking.
- La creación de diferentes rankings para los diferentes órdenes de n-gramas con lo que no se penalizaban los n-gramas de ordenes mayores en preferencia a aquellos más bajos (e.g. unigramas o bigramas) que suelen aparecer mucho más.
- Finalmente, inspirados en el trabajo presentado en (Lamel et al, 2002), donde se obtuvieron mejores resultados de identificación usando unidades más discriminativas, decidimos incorporar el mismo concepto aquí. En este caso, ubicando en las posiciones más altas del ranking aquellos n-gramas que aparezcan más en un idioma que en los demás y, por tanto, son más discriminativos.

Con el objetivo de calcular el valor de discriminación de cada n-grama probamos diferentes fórmulas basadas en la conocida métrica tf-idf. Para describir la fórmula que usamos, partimos de la ecuación 3 en la que $n_1(w)$ es el número de veces que aparece un n-grama en un idioma concreto, y $n_2(w)$ las veces que ocurre ese mismo n-grama en los otros idiomas, y T son las plantillas creadas para cada idioma.

$$N_1 = \sum_{\forall w: w \in T_1} n_1(w) \qquad N_2 = \frac{1}{|T-1|}\sum_{\forall w: w \in T; T \neq T_1} n_2(w) \qquad (3)$$

Como el número de cuentas será diferente para cada idioma y orden de los n-gramas, antes de hacer las comparativas aplicamos un proceso de normalización utilizando la ecuación 4. En esta fórmula: $N_1$ es la suma de todas las cuentas de todos los n-gramas para el idioma actual y $N_2$ es el promedio para el resto de idiomas.

$$n_1'(w) = \frac{n_1(w) \times N_2}{N_1 + N_2} \qquad n_2'(w) = \frac{n_2(w) \times N_1}{N_1 + N_2} \qquad (4)$$

Finalmente, también se puede aplicar un umbral sobre estos valores normalizados con el fin de eliminar n-gramas no representativos que aparecen muy poco. En nuestro caso no usamos estos umbrales. La ecuación 5 muestra las

fórmulas empleadas tanto en el caso de que el n-grama aparezca más en un idioma que en los otros ($n_1' > n_2'$).

$$\mathrm{Disc}(n_1) = \begin{cases} \alpha * \left( \frac{n_1' * (n_1' - n_2')}{(n_1' + n_2')^2} + \delta \right), n_1' > n_2' \\ \alpha * \left( \frac{n_2' * (n_1' - n_2')}{(n_1' + n_2')^2} + \delta \right), else \end{cases} \quad (5)$$

Donde, $\propto$ y $\delta$ son valores que normalizan el ranking discriminativo entre 0 y 1. Donde el valor "0" significa que el n-grama no es nada discriminativo o nada relevante para el idioma, y el valor "1" que es muy discriminativo o relevante para el idioma a reconocer.

## 2.3 Descripción de la técnica propuesta

Tal como hemos comentado en la introducción, la técnica propuesta se basa en la modificación de las cuentas de los posteriorgramas utilizadas para la generación de los i-vectores y le incorpora la información procedente del sistema que genera el ranking discriminativo de las cuentas de n-gramas. Para ello, lo que proponemos es modificar las cuentas de posteriorgramas incrementando su valor en función de cuán discriminativo sea el n-grama; es decir, que aparezca más para un idioma que para los demás. Para ello, utilizamos la ecuación 6:

$$C_{d,n}^{i} = \left(1 + \omega_n^i\right) \times Co_{d,n}^{i} \quad (6)$$

donde $C_{d,n}^{i}$ es el nuevo valor de la cuenta para el n-grama *n* obtenido para el fichero *d* e idioma *i*; $Co_{d,n}^{i}$ es el valor de la cuenta original en el supervector y $\omega_n^i$ es el valor de discriminabilidad del n-grama calculado al crear el ranking discriminativo para el idioma $i$.

Como puede observarse, el resultado es que la cuenta permanece inalterada si el n-grama no es discriminativo y puede llegar a duplicarse en caso de máxima discriminatividad.

El resultado es almacenado como un nuevo supervector denominado (SVs-RkDis) que ahora contendrá el efecto discriminativo de los rankings y que permitirá la generación de nuevos i-vectores en los que las dimensiones relacionadas con los n-gramas discriminativos adquieran mayor relevancia.

## *3 Condiciones de la evaluación y los experimentos*

Para la realización de las pruebas de la técnica propuesta hemos partido del mismo conjunto de datos de entrenamiento, evaluación y desarrollo que se usaron durante la evaluación ALBAYZIN 2012 LRE. Para esta evaluación se distribuyeron ficheros de audio extraídos de un portal de vídeos en la web, con diferentes longitudes, condiciones de canal y número de locutores, así como la presencia de diversas señales sonoras como música y ruido (Rodriguez-Fuentes et al, 2012).

Por otra parte, para la evaluación se propusieron cuatro tipos de condiciones distintas: a) plenty-closed, b) plenty-open, c) empty-closed, y d) empty-open. Donde los términos plenty y empty hacen referencia a que se tiene, o no, un conjunto de datos de entrenamiento, así como una diferenciación en los idiomas a reconocer. En el caso de la condición plenty se debían reconocer los siguientes 6 idiomas: español, inglés, portugués, gallego, vasco y catalán. Para la condición empty los idiomas eran: francés, italiano, alemán y griego. En cuanto a los términos closed y open, se refieren a la posibilidad de reconocer únicamente los idiomas mencionados antes (i.e. closed) o a la posibilidad de que el sistema pudiera detectar que el idioma del fichero es diferente a los incluidos en la condición (i.e. open).

| Ficheros | Limpios | Ruidosos | Total |
|---|---|---|---|
| **Español** | 486 | 312 | 798 |
| **Inglés** | 322 | 365 | 587 |
| **Gallego** | 675 | 300 | 975 |
| **Catalán** | 440 | 209 | 649 |
| **Vasco** | 579 | 215 | 794 |
| **Portugués** | 558 | 295 | 853 |
| **Total** | 3060 | 1596 | 4656 |

| | Train | Dev | Test | Eval |
|---|---|---|---|---|
| **Ficheros totales** | 4656 | 458 | 457 | 941 |

Tabla 1. Estadísticas por idiomas de los ficheros de entrenamiento y de la distribución de todos los datos de la evaluación para la condición plenty-closed

Los experimentos presentados en este artículo se han realizado únicamente para la condición principal de la evaluación, i.e. plenty-

closed. En la Tabla 1, se muestra la distribución y el número de ficheros de entrenamiento disponibles para la condición plenty-closed, así como la distribución que hicimos de todos los datos para realizar ajustes y probar el sistema antes (test) y durante la evaluación oficial (eval).

Finalmente, conviene mencionar que durante la evaluación oficial se propuso la utilización de la métrica $F_{act}$ con el objetivo de medir y comparar la bondad de los sistemas propuestos. Esta métrica se puede entender como una medida del grado de "incertidumbre" que tiene el sistema para detectar los idiomas. De esta manera, un valor de 0.0 significa que el sistema no tiene ninguna duda para reconocer los idiomas, en tanto que un valor igual o superior a 1.0 que no es capaz de mejorar la tasa de un sistema que escogiera de forma equiprobable cualquier idioma. Para mayores detalles se recomienda consultar (Rodriguez-Fuentes et al, 2012) y (Rodriguez-Fuentes et al, 2013).

Finalmente, también hemos decidido incluir en los resultados el cálculo de la medida $C_{avg}$ en porcentaje, ya que esta métrica ha sido ampliamente utilizada en todas las competiciones internacionales de reconocimiento de idioma. Esta medida tiene como objetivo ponderar los errores de falsa aceptación y falso rechazo del sistema por lo que un valor cercano a 0 significa que el sistema no comete ninguno de estos dos tipos de errores.

## 4 *Resultados*

La Tabla 2 muestra los resultados obtenidos al usar el sistema fonotáctico basado en el uso de i-vectores de tamaño 400 sobre los supervectores de cuentas de trigramas originales (SV) tanto para el conjunto de datos de test, como los proporcionados durante la evaluación. En la segunda línea vemos los resultados tras modificar las cuentas de los supervectores empleando las plantillas discriminativas (SVs+RkDis).

Tal como se puede ver en la Tabla 2, la modificación de las cuentas originales mediante la información discriminativa permite mejorar los resultados de $C_{avg}$ en un 8.63% relativo y un 0.4% en la tasa de $F_{act}$ para los datos de evaluación. En cuanto a los ficheros de test vemos que la mejora en Fact es un poco mayor (0.6%) en tanto que para Cavg empeora sólo un poco (0.43%).

| FICHEROS DE TEST | | | |
|---|---|---|---|
| **No.** | | **Fact** | **Cavg(%)** |
| **1** | SVs | 0.133658 | 6.94 |
| **2** | SVs+RkDis | 0.132864 | 6.97 |

| FICHEROS DE EVALUACIÓN | | | |
|---|---|---|---|
| **No.** | | **Fact** | **Cavg(%)** |
| **1** | SVs | 0.181393 | 9.85 |
| **2** | SVs+RkDis | 0.180704 | 9.00 |

Tabla 2. Resultados de los errores de reconocimiento para los ficheros de test y evaluación con trigramas.

## 5 *Conclusiones y lineas futuras*

En este artículo hemos presentado una técnica novedosa que permite combinar dos tipos de sistemas fonotácticos distintos empleando para ello información de largo alcance e información discriminativa como son las que proveen las plantillas y un sistema basado en i-vectores que es la técnica más exitosa para reconocimiento de idioma actualmente. Los resultados sobre los datos de la evaluación muestran que la técnica propuesta permite mejorar las tasas de reconocimiento hasta un 8.63% relativo, validando así sus prestaciones.

En relación con los trabajos futuros proponemos la inclusión de umbrales de decisión aplicados a los valores discriminativos de las plantillas, de forma que únicamente los n-gramas con un número mínimo de repeticiones vean modificadas las cuentas de sus posteriorgramas. En esta misma línea, consideraremos la creación de nuevas plantillas en las que el valor discriminativo pueda ser calculado a partir de nuevas fórmulas pudiendo también utilizar umbrales. Finalmente, también trabajaremos en ampliar esta técnica utilizando un sistema tipo PPRLM en el que tengamos no sólo un reconocedor de fonemas si no que podamos usar varios reconocedores en paralelo (e.g. aprovechando también los modelos de Checo o Ruso que viene incluido con el reconocedor de la Universidad de Brno).

## 6 *Agradecimientos*

Este trabajo ha sido posible gracias a la financiación de los siguientes proyectos: MA2VICMR (CC.AA. de Madrid, S2009/TIC-1542), y TIMPANO (TIN2011-28169-C05-03).

## *7 Referencias bibliográficas*


Kinnunen, T., H. Li, 2010. “An overview of text-independent speaker recognition: From features to supervectors”. Speech Communication, Vol 52, Issue 1, pp. 12-40.

Dehak, N., P. Kenny, R. Dehak, P. Dumouchel y P. Ouellet, 2011. “Front-End Factor Analysis for Speaker Verification”. IEEE Transactions on Audio, Speech, and Language Processing, 19(4), pp.788-798.

Martínez, D., O. Plchot, L Burget, O. Glembek y P. Matejka. 2011. “Language Recognition in iVectors Space". Proceedings of Interspeech. pp. 861-864.

D'Haro, L.F., O. Glembek, O. Plchot, P. Matejka, M. Soufifar, R. Cordoba y J. Cernocky. 2012. “Phonotactic language recognition using i-Vectors and phoneme posteriorgram counts”. Proceedings of Interspeech pp. 9-13.

D'Haro, L.F., R. Cordoba, 2013. “Low-Resource language recognition using a fusion of phoneme posteriorgrams counts, acoustic and glottal-based i-Vectors”. ICASSP 2013.

Cordoba, R., L.F. D'Haro, F. Fernandez-Martinez, J. Macias-Guarasa, y J. Ferreiros. 2007. “Language Identification based on n-gram Frequency Ranking”. 8th Annual Conference of the International Speech Communication Association, Interspeech, Vol. 3, pp.1921-1924.

Zissman, M. 1996. “Comparison of four approaches to automatic language identification of telephone speech”. IEEE Transactions on Speech and Audio Processing, vol.4, no.1, pp. 31-44.

Schwarz, P., 2009. “Phoneme Recognition based on Long Temporal Context”, PhD Tesis. Brno University of Technology. Disponible: http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context.

Diez, M., A. Varona, M. Penagarikano, L. Rodriguez-Fuentes, G. Bordel. "Dimensionality Reduction of Phone Log-Likelihood Ratio Features for Spoken Language Recognition". Interspeech 2013; Lyon, France, 25-29 aug., 2013

Povey, D., L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, S. Thomas. 2010. “Subspace Gaussian Mixture Models for Speech Recognition”, Proceedings of ICASSP, Dallas. pp 4330-4333.

Kockmann, M., L. Burget, O. Glembek, L. Ferrer, J. Cernocky, 2010. “Prosodic speaker verification using subspace multinomial models with intersession compensation,” Proceedings of ICSPL, Makuhari, Chiba, Japan.

Soufifar, M., M. Kockmann, L. Burget, O. Plchot, O. Glembek, T. Svendsen. 2011. “IVector approach to phonotactic language recognition”. Proceedings of Interspeech 2011, pp 2913-2916.

Cavnar, W., J. Trenkle. 1994.” N-Gram-Based Text Categorization”. Environmental Research Institute of Michigan.

Lamel, L., J-L. Gauvain, G. Adda, G. 2002. "Lightly Supervised and Unsupervised Acoustic Model Training". Computer Speech and Language, Vol.16, no.1, pp. 115-129.

Rodríguez-Fuentes, L., N. Brummer, M. Penagarikano, A. Varona, M. Diez, G. Bordel. 2012. “The Albayzin 2012 Language Recognition Evaluation Plan (Albayzin 2012 LRE)”.

Rodriguez-Fuentes, L. J., Brümmer, N., Penagarikano, M., Varona, A., Bordel, G. , Diez, M. "The Albayzin 2012 Language Recognition Evaluation". Interspeech 2013; Lyon, France, 25-29 aug., 2013.