Estrategias de generación y reducción de variantes de pronunciación en sistemas de reconocimiento automático de habla: consideraciones arquitecturales

Javier Macías Guarasa, Javier Ferreiros, Ricardo de Córdoba, Juan Manuel Montero, José David Romeral y José M. Pardo
Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica. ETSI Telecomunicación. Universidad Politécnica de Madrid
Ciudad Universitaria s/n. 28040-Madrid
{macias,jfl,cordoba,montero,jdromeral,pardo}@die.upm.es

Resumen: En el contexto de sistemas de reconocimiento de habla de gran vocabulario es fundamental modelar de forma adecuada las variaciones alofónicas con las que se enfrentará el sistema en una tarea real. En esta comunicación describimos un estudio sobre la introducción de variantes de pronunciación dirigidas por datos, abordando tanto los procesos de generación y reducción de las mismas como los de evaluación de su impacto en la tasa del sistema. Las técnicas descritas se acompañan del correspondiente trabajo experimental, sobre dos sistemas radicalmente distintos en relación a su potencia de discriminación (basados en arquitecturas integrada y no integrada, pensadas para trabajar como módulos de hipótesis y verificación, respectivamente), de modo que es posible obtener conclusiones razonadas sobre el funcionamiento de cada uno de ellos en relación al incremento del tamaño de los diccionarios. Los resultados más relevantes muestran cómo, para el caso de la arquitectura no integrada es posible incrementar notablemente su tasa de inclusión, incluso para incrementos muy importantes del tamaño del diccionario (de hasta un 250%). Por el contrario, el incremento del número de variaciones tiene un efecto claramente negativo, cuando se utiliza el sistema integrado.
Palabras clave: Múltiples pronunciaciones, reconocimiento de habla, técnicas dirigidas por datos, arquitecturas para reconocimiento de habla.

Abstract: In the context of large vocabulary speech recognition systems, it is crucial to accurately model the allophonic variations that will be found in a real world task. In this paper we describe a study on the use of data driven pronunciation variations, considering the generation and reduction strategies, as well as their impact in the system performance. The described techniques are supported by the corresponding experimental evaluation on two radically different systems in what respect to their discrimination power (based on integrated and non-integrated architectures, designed to work as hypothesis and verification modules, respectively), so that it's possible to discuss on their relative performance as a function of the increase in dictionary size. The most relevant results show that in the case of the non integrated architecture, we can significantly improve the inclusion rate, even for huge increases in dictionary size (up to 250%). On the contrary, the increase in the number of pronunciation variants has a clearly negative effect when applied to the integrated system.
Keywords: Multiple pronunciations, speech recognition, data driven techniques, speech recognition architectures.

1 Introducción

En la bibliografía se pueden encontrar numerosas referencias al problema de la introducción de múltiples pronunciaciones (una excelente revisión al respecto puede encontrarse en (Strik y Cucchiarini, 1999)). En general, se trata de abordar el problema de la variabilidad en el modo de producción de la señal de voz debido a diferencias dialectales y modos de articulación específicos de ciertos locutores (variaciones estilísticas, sociales y culturales).

Estrictamente hablando, cualquier sistema de reconocimiento automático de habla se enfrenta de forma implícita con variaciones en

la pronunciación, dado que los modelos utilizados (típicamente modelos ocultos de Harkov, HMMs) se encargan de tener en cuenta todas esas variaciones segmentales y temporales. Sin embargo, lo que se pretende cuando se habla de variaciones de pronunciación, es introducir explícitamente conocimiento sobre las mismas, con lo que se consiguen sustanciales mejoras si los modelos acústicos se corresponden perfectamente con las transcripciones usadas (Saraclar et al., 2001). Las estrategias típicas usan aproximaciones dirigidas por conocimiento, en las que se aplican reglas de variación alofónica a los diccionarios canónicos, o bien dirigidas por datos (Strik y Cucchiarini, 1999).

La mayor parte de los casos descritos en la bibliografía se ocupan de estudiar métodos específicos y analizar su impacto en las tasas de reconocimiento. Cuando se introducen variantes de pronunciación en el léxico (diccionario) de una tarea de reconocimiento, el objetivo es mejorar la precisión de la decodificación acústica del reconocedor. Sin embargo, si las variantes introducidas no son adecuadas, la tasa final de error puede aumentar. Con esta restricción, los equipos de investigación son sumamente cuidadosos a la hora de introducir variantes y se han propuesto distintas estrategias para generar y limitar su número, como el uso de un criterio de máxima verosimilitud (Holter, 1998), el suavizado de transcripciones generadas automáticamente (Riley et al, 1999), la medida de la ocurrencia de las variantes añadidas (Kessens y Wester, 1999) (Adda-Dekker y Lamel, 1999) o el uso de medidas de confusabilidad (Wester, 2003), por citar algunas.

En esta comunicación, introducimos nuevas estrategias derivadas del enfoque de (Riley, 1999) pero, sobre todo, analizamos su impacto sobre dos sistemas arquitecturalmente distintos, basados en una búsqueda integrada o no integrada (Macías, 2001), pensando en un sistema final basado en el paradigma hipótesis-verificación en el que el módulo de hipótesis usaría el módulo con arquitectura no integrada, y el de verificación usaría el módulo de arquitectura integrada, con lo que conseguimos ofrecer una visión más amplia de los efectos que se pueden conseguir con el uso de diccionarios aumentados con múltiples variantes de pronunciación. La aplicación experimental se hará sobre sistemas de reconocimiento de habla aislada de gran vocabulario en entorno telefónico.

2 Nuestro enfoque

En nuestro trabajo, sólo nos preocuparemos del nivel segmental (como sucede en la mayor parte de los estudios de la bibliografía, salvo algunos intentos como el descrito en (Strik, 2002)), y dentro de él, del referido a variaciones dentro de cada palabra basadas en técnicas dirigidas por datos.

2.1 Consideraciones sobre la evaluación

La evaluación del impacto de la introducción de pronunciaciones alternativas en sistemas de reconocimiento automático de habla se hace tradicionalmente midiendo la reducción en tasa de error obtenida, y evaluando el incremento del tamaño del diccionario usado en cada caso. Además de esto, el hecho de que estemos pensando finalmente en una arquitectura basada en el paradigma hipótesis-verificación, nos obliga a introducir una consideración adicional: El módulo de hipótesis debe entregar al módulo de verificación un espacio de búsqueda más reducido que el original. En el caso del entorno experimental que nos ocupa (gran vocabulario y habla aislada), dicho espacio de búsqueda reducido estará compuesto por un subconjunto del vocabulario de la tarea, que llamaremos *lista de preselección*. El tamaño *n* de dicha lista de preselección debe diseñarse de forma que la *tasa de inclusión* (definida como la tasa de reconocimiento considerando que una palabra ha sido acertada si se encuentra en las primeras *n* posiciones de la *lista de preselección*) sea lo suficientemente alta para que no limite el rendimiento del sistema global. En nuestro caso, el objetivo definido es conseguir una tasa de inclusión superior al 98%.

Para el módulo de hipótesis, esta consideración nos obliga a discutir los resultados de la introducción de variaciones de pronunciación en función del tamaño de la lista de preselección: la introducción de variantes funcionará correctamente si conseguimos disminuir dicho tamaño para la misma tasa de inclusión al usar el diccionario aumentado.

Para el módulo de verificación, o incluso para el sistema conjunto hipótesis + verificación, la única consideración que se debe tener en cuenta se refiere a la tasa de reconocimiento final obtenida, por supuesto.

Finalmente es importante señalar que los resultados obtenidos deben ser siempre acompañado de consideraciones acerca de su fiabilidad estadística.

2.2 Estrategias de generación de variantes

La generación de variaciones de pronunciación se hace en nuestra propuesta a partir de la información proporcionada por un módulo de análisis fonético basado en el algoritmo de un paso que calcula la secuencia óptima de unidades acústicas para cada palabra de entrada. Obviamente dicha secuencia corresponderá en muy pocos casos con la pronunciación canónica de la palabra que se desea reconocer y es precisamente en los errores cometidos por ese decodificador acústico, donde buscamos extraer las múltiples pronunciaciones.

Para nuestros propósitos, veremos el proceso como una "corrección" del diccionario canónico (entendiendo por "corrección" la incorporación de transcripciones del diccionario canónico). Las estrategias planteadas usan las cadenas fonéticas generadas para la lista de entrenamiento como nuevas transcripciones de las palabras correspondientes, transcripciones que se incorporan o no al diccionario en función de los siguientes criterios:

- Corrección sin limitación: todas las producciones (cadenas fonéticas) contribuyen a generar variantes (*corrección total*).
- Corrección limitada a aquellas producciones para las que hay más de un determinado número de ejemplos, con el objetivo de no atender a variantes que no van a poder ser validadas con un mínimo de fiabilidad.
- Corrección limitada a aquellas producciones que introducen un número determinado máximo de errores de alineamiento, con el objetivo de no atender a variantes que introducen una variación excesiva en comparación con la pronunciación canónica.
- Corrección limitada a aquellas cadenas que producen errores de reconocimiento (*refuerzo negativo*) o aciertos (*refuerzo positivo*). Como parámetro de control se especifica el tamaño de la lista de preselección que se considerará como acierto (medido como porcentaje del tamaño del diccionario)

2.3 Estrategias de filtrado (reducción de variantes)

El objetivo de las estrategias de filtrado es, en todos los casos, limitar la complejidad introducida en el espacio de búsqueda acústico por el aumento en el número de entradas del diccionario, dejando aquellas que son realmente relevantes para nuestra tarea, por los beneficios que aporta al rendimiento.

Todas las estrategias de filtrado parten de la validación de las propuestas generadas por los mecanismos descritos en el apartado anterior, usando los nuevos diccionarios para reconocer la base de datos de entrenamiento. Nuestra propuesta consiste en estudiar el *grado de uso* de la estructura de árbol en la que se representan los diccionarios, entendiendo por *grado de uso* el número de veces (*ocurrencias*) en las que cada nodo del árbol formaba parte del camino óptimo recorrido tras el correspondiente alineamiento. Así, nuestro método permite tener una idea muy precisa de hasta qué punto hay alternativas que se utilizan de forma efectiva y cuáles no. El procedimiento práctico consiste en, para toda la base de datos de entrenamiento, alinear cada cadena con la palabra correspondiente del diccionario y anotar el número de veces que se recorre cada nodo.

Una vez disponibles las medidas de *uso* de árbol (realmente las de cada nodo del mismo), analizamos un amplio abanico de métodos de medida de importancia relativa de los mismos, de cara a su eliminación. En este punto introducimos el concepto de *grupo de nodos finales*, entendiéndolo como aquel conjunto de nodos finales que están asociados a una misma palabra (en la estructura de árbol, cada palabra puede tener varias pronunciaciones, lo que se traduce en distintos nodos finales, cada uno asociado a una de ellas). El número de ocurrencias permite estimar valores de probabilidad que se calculan para cada nodo final (normalizando por los valores que correspondan en cada caso). Las métricas usadas para validar las variantes de pronunciación introducidas en los diccionarios fueron las siguientes:

- Impacto en la probabilidad global de cada nodo final (calculado sobre el total de nodos finales): Se eliminan los menos probables.
- Impacto en la probabilidad parcial (calculado sobre el total de nodos del grupo de nodos finales al que pertenece el

considerado): Se eliminan los menos probables.

- Impacto en la entropía global (calculado como el aumento de entropía que supondría eliminar ese nodo en el conjunto de todos los nodos finales). Se eliminan los que menor aumento de entropía produzcan.
- Impacto en la entropía parcial (calculado como el aumento de entropía que supondría eliminar ese nodo en el conjunto de los nodos de su grupo). Se eliminan los que menor aumento de entropía produzcan.

Así, una vez etiquetados convenientemente los nodos y calculadas las métricas de validación asociadas a cada uno, se ordenan de acuerdo con el criterio que se desea aplicar en cada caso (de los cuatro vistos) y se elimina un porcentaje determinado de los mismos, con el objetivo de reducir el tamaño del espacio de búsqueda que tenemos tras la corrección, lo que especificamos como un porcentaje de incremento con relación al del diccionario canónico.

La consideración más importante en cuanto a la realización práctica de las medidas es el efecto del tamaño del grupo de nodos finales en las mismas. Para grupos muy pequeños nos encontramos con problemas de estimación fiable y, en general, con posibles valores nulos. Tras una experimentación previa, se llegó a la conclusión de que una aproximación razonable para evitar dichos problemas era aplicar un suavizado umbral, similar al típicamente usado en el entrenamiento de modelos acústicos.

Por último, es importante mencionar que el cálculo de aumento de entropía presenta problemas prácticos. Si pensamos en las consecuencias de una eliminación de un nodo final (lo que implica la eliminación de una variante de pronunciación), está claro que su pérdida debería implicar el reparto de la probabilidad asociada al mismo entre el resto de posibilidades. La aproximación inmediata al problema es repartir de forma proporcional al resto de probabilidades, pero en ningún caso tendremos la certeza de que dicho reparto se haría de esa forma si volviéramos a realizar el proceso de alineamiento. El cálculo exacto implica un coste computacional sumamente elevado y experimentos previos con listas reducidas mostraron que las diferencias en la calidad de la ordenación no son significativas, si comparamos el método exhaustivo con el aproximado que hemos descrito y que es el finalmente utilizado.

3 Resultados experimentales

3.1 Bases de datos y diccionarios

En nuestros experimentos, hemos usado parte de VESTEL (Tapias et al., 1994), una base de datos de habla telefónica realista, independiente del locutor, en idioma castellano, sobre la que se han definido dos tareas:

- VESTEL-S, dividida en PRNOK5TR (destinada al entrenamiento de modelos y compuesta por 5820 producciones de habla) y PERFDV (destinada a evaluación y compuesta por 2536 producciones de habla).
- VESTEL-L: Subconjunto de experimentación generado usando la técnica del *leave-one-out* sobre los casi 10000 ficheros de los que disponemos de VESTEL, para mejorar la fiabilidad estadística de los resultados. Cada subconjunto usa unos 9000 ficheros para entrenar y 1000 para reconocer.

En nuestras tareas hemos usado un diccionario "canónico" compuesto por 1175 palabras para la experimentación sobre VESTEL-S, y de 1952 palabras para VESTEL-L, aunque el sistema original maneja tamaños de hasta 10000 palabras. El problema es que, de cara al uso de múltiples pronunciaciones, es imprescindible limitar los diccionarios a aquellas palabras de las que se tienen ejemplos acústicos.

Se usan HMMs semicontinuos de tres estados, con autosalto y transiciones simples y dobles, independientes del contexto en el sistema no integrado (hipótesis) y dependientes de contexto en el integrado (verificación). El alfabeto consta de 45 unidades alofónicas, usando 800 distribuciones en los modelos contextuales.

3.2 Evaluación de las estrategias de generación

El primer conjunto de experimentos llevados a cabo para evaluar los procesos de generación se orientó a medir el impacto de cada una de las estrategias descritas en el apartado 2.2 en la tasa de inclusión del sistema.

A modo de resumen, es importante mencionar que prácticamente todas las estrategias usadas producen, obviamente, mejoras muy importantes al realizar la evaluación sobre la lista de entrenamiento, para

todo el rango de variación de los parámetros de control. Sin embargo, el impacto en la lista de evaluación es notablemente diferente: las diferencias entre el uso del diccionario canónico y los corregidos (aumentados) son razonablemente pequeñas, sobre todo teniendo en cuenta el tremendo incremento del tamaño del diccionario en algunos casos (que, por ejemplo, pasa en uno de los experimentos de 5086 nodos a más de 30000 y de 1175 entradas a 5984). A modo de ejemplo, en la figura 1 se muestra la curva de tasa de inclusión sobre PERFDV con distintas estrategias de generación de variantes (el eje de abscisas es la longitud de la lista de preselección medida como porcentaje sobre el tamaño del diccionario).

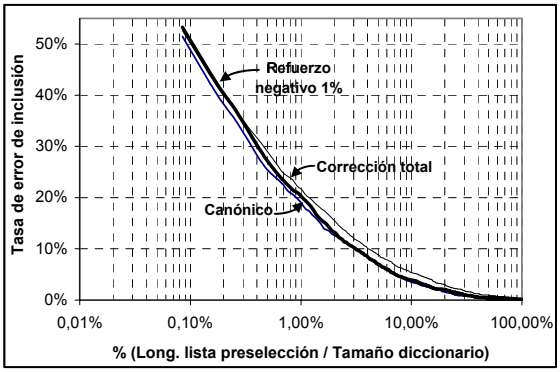

Figura 1: Tasa de inclusión sobre PERFDV con distintas estrategias de generación de variantes.

De nuestro estudio del efecto de distintos procesos de generación de variantes de pronunciación, y tras analizar el impacto de las estrategias de filtrado, se verificó que dicho filtrado es la etapa realmente importante, la que proporciona la verdadera potencia a nuestra estrategia, con lo que la metodología recomendada parte de realizar la corrección del diccionario canónico con toda la información acústica disponible (*corrección total*), centrando el esfuerzo en la etapa de filtrado de variantes.

3.3 Evaluación de las estrategias de filtrado sobre el módulo de hipótesis

En los experimentos realizados, nos centramos en estudiar el impacto en la tasa de inclusión de cada una de las estrategias de filtrado descritas en el apartado 2.3, para decidir cuál era la óptima, evaluándolo en función del incremento relativo en el tamaño del diccionario, ya que también estamos interesados en determinar para qué tamaño conseguimos los mejores resultados.

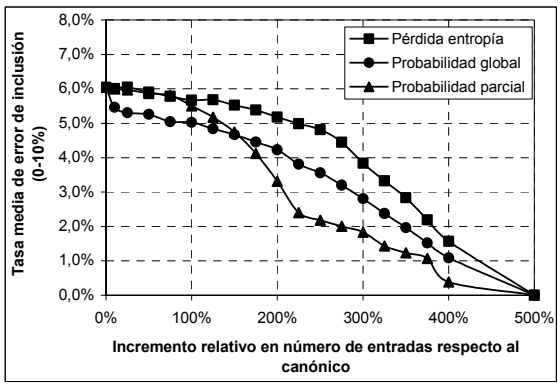

Figura 2: Efecto de distintos métodos de filtrado en la tasa media de inclusión para un 10% de la lista de preselección sobre PRNOK5TR.

Así, establecimos que las medidas basadas en medida de probabilidad del uso de variantes, son más potentes que las relacionadas con el aumento de entropía, como puede verse en la figura 2, en la que la métrica de calidad es la media de la tasa de inclusión para una lista de preselección de longitud igual al 10% del tamaño del diccionario.

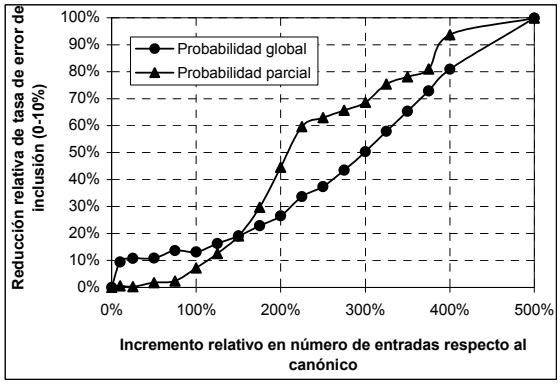

Figura 3: Efecto de métodos de filtrado basados en prob. en la tasa de inclusión (PRNOK5TR)

A partir de esos resultados, se realizaron experimentos más detallados sobre el impacto del uso de las estrategias basadas en probabilidad global o parcial, optando por evaluar en este caso, la reducción en tasa media de error de inclusión producida al compararlos con el resultado usando el diccionario canónico. De ese estudio, como se muestra en la figura 3, concluimos que el método de la probabilidad parcial es el mejor para un amplio rango de

condiciones, a partir de un incremento relativo de tamaño del diccionario del 150%, y que la máxima diferencia se obtiene en valores alrededor del 250%, que será el valor elegido en nuestro caso como incremento del número de entradas en el diccionario para la evaluación sobre la lista de reconocimiento. Obviamente la decisión sobre el método y el tamaño se hizo sobre la lista de entrenamiento, y al experimentar sobre la lista de evaluación se comprobó que los resultados seguían siendo mejores al comparar con el uso del diccionario canónico, tal y como se muestra en la figura 4, que representa la disminución relativa de la tasa media de error de inclusión para distintos tamaños de la lista de preselección, esta vez sobre PERFDV.

La explicación más razonable es que el método basado en probabilidad parcial captura toda la potencia de decisión necesaria al efectuar decisiones locales dentro de cada grupo de nodos finales. El basado en probabilidad global diluye las aportaciones individuales de cada grupo, suavizando el efecto y disminuyendo la relevancia de las probabilidades particulares.

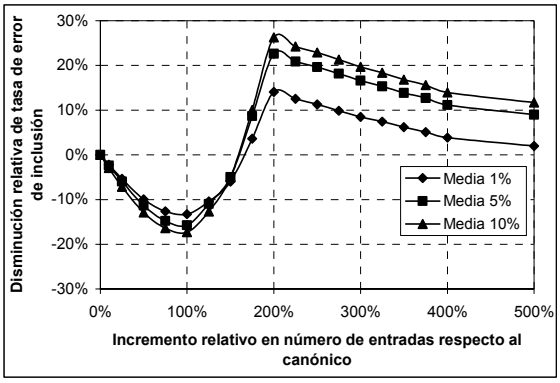

Figura 4: Disminución relativa de la tasa media de error de inclusión con filtrado basado en prob. sobre la lista de evaluación (PERFDV)

Merece la pena destacar cómo, a pesar del considerable crecimiento del número de entradas del diccionario, la tasa de inclusión llega a ser incluso mayor que la obtenida con el diccionario canónico, lo que ratifica la bondad de las estrategias de incremento del número de variantes de pronunciación basadas en criterios dirigidos por datos, al modelar deficiencias explícitas de los modelos acústicos de los sistemas sobre los que se aplica.

En la figura 5 se muestra un ejemplo ilustrativo en el que se dibuja la curva de tasa de inclusión para el diccionario canónico (etiquetada como *Antes*), para el corregido y filtrado incrementando el número de entradas del diccionario en un 250% (etiquetado como *Después*), y las zonas de la gráfica en la que las diferencias son estadísticamente significativas (marcadas con los puntos negros en la parte superior, etiquetado como *Validado*). En este ejemplo merece la pena destacar que la tasa de inclusión del 98% (nuestro objetivo) se consigue con el diccionario canónico para el candidato número 208 (17,7% del tamaño de diccionario), mientras que con el diccionario corregido y reducido a un tamaño un 250% superior al del canónico, se obtiene para el candidato número 123 (10,46% del tamaño del diccionario), lo que supone una mejora relativa del 40%. Esta significativa mejora implicará un considerable ahorro computacional en el sistema basado en hipótesis+verificación, ya que el módulo de verificación tendrá que enfrentarse a un espacio de búsqueda considerablemente más reducido para obtener los mismos resultados. En cuanto a la fiabilidad de los resultados y dado lo limitado de la base de datos (2502 ejemplos), no podemos garantizar la significancia estadística para todo el rango de valores de la longitud de la lista de inclusión, pero sí en la zona de interés, esto es, alrededor del 10% del tamaño del diccionario utilizado.

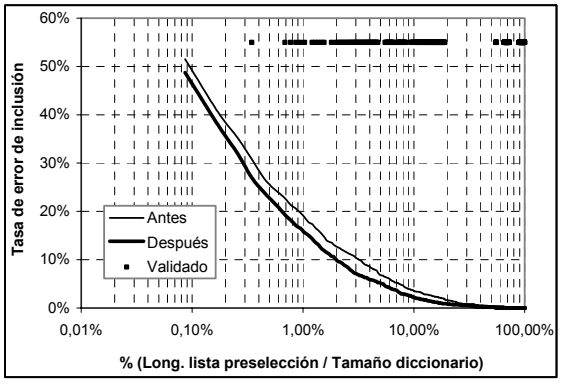

Figura 5: Tasa de inclusión sobre PERFDV con el diccionario canónico (*Antes*), el corregido y filtrado incrementando el diccionario un 250% (*Después*) y las zonas en las que las diferencias son estadísticamente significativas (*Validado*)

3.4 Aplicación a la misma tarea en distintas condiciones experimentales

Uno de los inconvenientes fundamentales de los métodos dirigidos por datos es la necesidad

de repetir la experimentación al cambiar las condiciones de la tarea, la base de datos, los diccionarios, etc.

Así, aplicamos las mismas ideas del anterior a la tarea VESTEL-L, usando el diccionario de 1952 palabras: no cambiamos sustancialmente la base de datos, pero sí las condiciones de entrenamiento y evaluación.

En la figura 6 se pueden ver los resultados obtenidos, que difieren cualitativamente de los mostrados en la figura 4 para la tarea VESTEL-S. Así, además de la elevada sensibilidad a variaciones en los datos que dirigen la técnica de generación y filtrado que hemos comentado, se muestra que también son muy sensibles a variaciones en las condiciones de experimentación, con lo que hay que prestar especial cuidado a las mismas para asegurar que el impacto de su introducción puede ser evaluado de forma adecuada, esto es, asegurando que las variantes introducidas van a poder ser utilizadas.

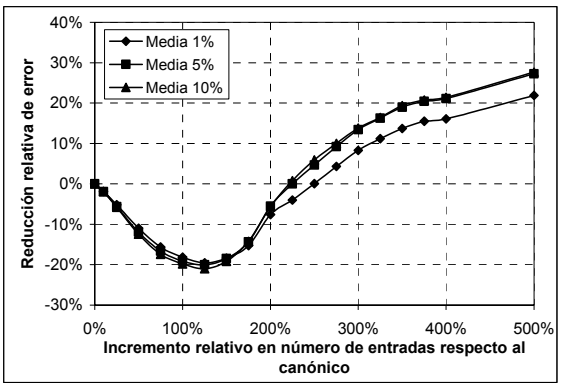

Figura 6: Disminución relativa de la tasa media de error de inclusión usando métodos de filtrado basados en prob. sobre VESTEL-L.

3.5 Aplicación al sistema integrado

Tras verificar las ventajas de la introducción de múltiples pronunciaciones en sistemas de preselección, aplicaremos por último la estrategia de corrección y reducción de variantes vista en los apartados anteriores a la tarea VESTEL-L usando un sistema integrado como base, mucho más potente que el no integrado, desde el punto de vista de su capacidad de discriminación.

Los resultados de la tasa de inclusión de la figura 7 muestran cómo el incremento del diccionario de un 400% sí tiene aquí un efecto negativo en la tasa de inclusión, siendo además dicha pérdida significativa estadísticamente. Por ejemplo, de una tasa de error del 14,3% para el diccionario canónico, se pasa a una del 16,1% cuando se incrementa el diccionario en un 100%.

El hecho de usar un modelado más fino y una búsqueda integrada no se ven beneficiadas en absoluto por el incremento de variaciones de pronunciación. Nuestra explicación a este efecto es la consideración de que dichas variantes han sido generadas con un modelado mucho más pobre, de forma que los errores producidos (y contemplados por tanto en las alternativas de pronunciación) no son coherentes con el nuevo modelado, que produciría otros.

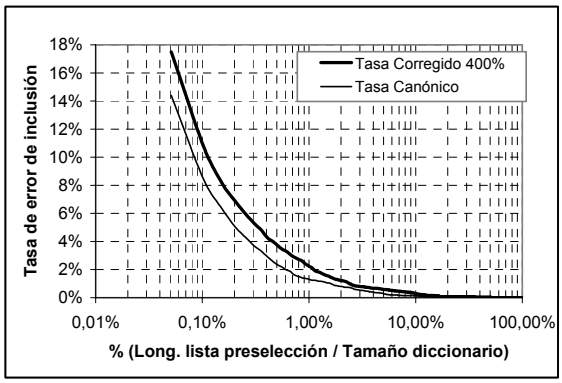

Figura 7: Curvas de tasa de inclusión usando el diccionario canónico y el corregido y filtrado usando un sistema integrado sobre la lista de evaluación (en VESTEL-L).

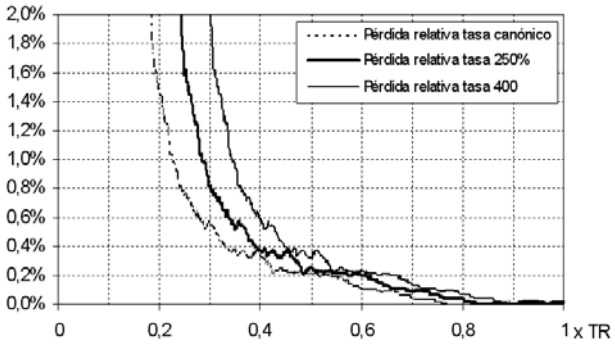

Figura 8: Pérdida relativa de tasa global de inclusión (hipótesis+verificación) en función de la fracción de tiempo real utilizado y para distintos incrementos en el tamaño del diccionario.

3.6 Consideraciones de coste computacional

Para concluir este apartado, queremos mencionar que el incremento de coste

computacional producido al incrementar el tamaño de los diccionarios del módulo de preselección todavía permite un amplio margen de maniobra. Incluso los incrementos tan importantes como los comentados más arriba del orden del 250% en número de entradas del diccionario, todavía permiten trabajar por debajo de tiempo real, como se muestra en la figura 8.

4 Conclusiones

En esta comunicación se ha hecho un estudio detallado de distintas estrategias para la introducción y el filtrado de variaciones de pronunciación dirigidas por datos y su aplicación en distintos esquemas arquitecturales (enfoques integrados, no integrados y los basados en el paradigma hipótesis-verificación).

Se ha introducido un nuevo enfoque basado en la corrección del diccionario y posterior reducción del espacio de búsqueda resultante, discutiendo sobre distintos criterios de corrección y reducción, basados en medidas probabilísticas y se ha comprobado cómo el proceso más importante es el de filtrado, siendo los resultados razonablemente insensibles al método de generación. Los experimentos han mostrado igualmente la potencia de cada criterio. Los resultados obtenidos han sido especialmente buenos, llegando a incrementar notablemente las tasas de inclusión a pesar del considerable incremento del número de variantes introducidas.

La aplicación de la misma estrategia y metodología a un sistema integrado ha mostrado un comportamiento contrario: el incremento de variantes ha producido incrementos en la tasa de error.

Nuestra línea fundamental de trabajo en estos momentos es la aplicación de los mecanismos de generación y filtrado aquí descritos a tareas de habla continua y espontánea (que ya comenzaron en (Ferreiros et al, 1998, 1999)), así como la investigación en métodos dirigidos por conocimiento y muy especialmente la definición de nuevas métricas de evaluación de la incorporación de múltiples pronunciaciones que ofrezcan más información sobre la bondad o no de cada una de ellas.

Bibliografía

Adda-Decker, M., y Lamel, L., 1999. Pronunciation variants across system configuration, language and speaking style

Ferreiros, J., Macías-Guarasa, J. y Pardo, J., 1998. "Introducing Multiple Pronunciations in Spanish Speech Recognition Systems". ESCA Tutorial and Research Workshop on "Modeling Pronunciation Variation for Automatic Speech Recognition"

Ferreiros,J. y Pardo, J., 1999. "Improving continuous speech recognition in Spanish by phone-class SCHMMs with pausing and multiple pronunciations. Speech Communication 29, pp. 65-76

Holter. T., 1998. "Maximum Likelihood Modelling of Pronunciation in Automatic Speech Recognition". PhD. Thesis. Norwegian University for S&T

Kessens, J. y Wester, M., 1999. "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation". Speech Communication, vol 29, p.193-207

Macías-Guarasa, J. "Arquitecturas y métodos en sistemas de reconocimiento automático de habla de gran vocabulario". Tesis Doctoral. Universidad Politécnica de Madrid. 2001.

Riley, M., Byrne, W et al., 1999. "Stochastic pronunciatio modeling from hand-labelled phonetic corpora". Speech Communication, vol 29, p. 209-224

Saraçlar, M., Nock, H., Khudanpur, S., 2001. "Pronunciation modeling by sharing Gaussian densities across phonetic models". Computer Speech and Language, vol 14, p. 137-160

Strik, H., 2001 "Pronunciation adaptation at the lexical level". Proc. of the ISCA Tutorial & Research Workshop (ITRW) 'Adaptation Methods For Speech Recognition', Sophia-Antipolis, France, pp. 123-131

Strik, H. y Cucchiarini, C., 1999. "Modeling pronunciation variation for ASR: a survey of the literature. Speech Communication, vol 29, p. 225-246

Tapias, D., Acero, A., Esteve, J., Torrecilla, J.C., 1994. "The VESTEL Telephone Speech Database". ICSLP 94, pp. 1811-1814

Wester, M., 2003. "Pronunciation modeling for ASR - knowledge-based and data-derived methods". Computer Speech and Language, vol 17, p. 69-85