HACIA LA PREDICCIÓN DE RENDIMIENTO EN SISTEMAS DE RECONOCIMIENTO DE HABLA

J. Macías-Guarasa, J. Ferreiros y J.M. Pardo
Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica
Universidad Politécnica de Madrid
{macias,jfl,pardo}@die.upm.es – http://www-gth.die.upm.es

Palabras clave: complejidad de diccionarios, comparación de tareas, predicción de rendimiento

RESUMEN

En este artículo, y en la línea de trabajo del diseño de sistemas que dicen ser "independientes del vocabulario", presentamos algunas ideas relativas a mecanismos de evaluación a priori del comportamiento de sistemas de reconocimiento al enfrentarse a diccionarios de distinta composición. Nuestro estudio muestra cómo, para sistemas de reconocimiento de gran vocabulario y con enfoques arquitecturales y de modelado distintos (con guiado léxico explícito o sin él, sistemas integrados y no integrados, modelado dependiente e independiente del contexto), la longitud media de las palabras de un diccionario es determinante de la tasa de reconocimiento final alcanzable y puede servir de punto de partida para evaluar la complejidad de diccionarios y hacer predicciones sobre tasas esperables.

1 INTRODUCCIÓN

Los sistemas de reconocimiento automático de habla actuales dependen de forma crítica de su correcto comportamiento en tareas para las que no se dispone de datos de entrenamiento a priori. En este sentido, los sistemas que reclaman ser "independientes del vocabulario" [1] deben ser capaces de reconocer formas grafémicas que no han aparecido en el material de entrenamiento. La investigación en esta línea implica considerar aspectos de generación automática de pronunciaciones, modelos de variaciones fonológicas y dialectales, así como tratamiento específico de los errores producidos por el reconocedor fonético.

En este artículo pretendemos describir algunos planteamientos que nos permitan obtener conclusiones relacionadas con el concepto de independencia del vocabulario, intentando aislar aquellos factores que pudieran enmascarar variaciones en la tasa de reconocimiento, no debidas exclusivamente al cambio en el diccionario. El objetivo es validar las posibles comparaciones que pudieran hacerse para evaluar un sistema enfrentado a diccionarios de distintas características, manteniendo o no el número de entradas de los mismos. Esta línea de trabajo puede derivar finalmente en la posibilidad de predecir el rendimiento en sistemas de reconocimiento automático de habla.

La primera observación clara es que la tasa de reconocimiento depende, obviamente, del número de entradas del diccionario utilizado. Frente a este problema podemos usar medidas de rendimiento como las descritas en [2] en las que la tasa de reconocimiento de sistemas de preselección para un cierto número de candidatos se media de forma relativa con respecto a la longitud del diccionario utilizado (así, podríamos comparar el reconocimiento para una tarea con un vocabulario de 2000 palabras o de 10000, sin más que considerar en ambos casos una longitud de lista de preselección medida como porcentaje de dichos tamaños: Un 10% en ambos casos implicaría comparar la tasa de inclusión para 200 candidatos en el primer caso y 1000 en el segundo).

Mecanismos como el descrito son válidos como primera aproximación, pero hace falta abundar más en el estudio de factores relacionados.

2 BASES DE DATOS

La experimentación descrita se ha realizado sobre VESTEL [3], una base de datos telefónica capturada sobre la red telefónica pública y que está compuesta de dígitos, números, comandos, nombres propios, etc. y diseñada para soportar investigación y desarrollo en sistemas de

reconocimiento automático de habla, con independencia del locutor y basada en unidades inferiores a la palabra.

La parte de VESTEL que hemos utilizado define tres subconjuntos fundamentales de trabajo:

- PRNOK: Conjunto destinado al entrenamiento genérico de los sistemas que operan sobre VESTEL.
- PERFDV: Conjunto destinado al reconocimiento e inicialmente diseñado para realizar pruebas "dependientes del vocabulario", en el sentido de que comparte con PRNOK los grafemas.
- PEIV1000: Conjunto destinado al reconocimiento e inicialmente diseñado para realizar pruebas "independientes del vocabulario", en el sentido de que no comparte con PRNOK ningún grafema.

Además de estos subconjuntos definidos, se realizó una partición adicional utilizando la técnica de *leave-one-out* (VESTEL-L) para incrementar la fiabilidad de los resultados de reconocimiento obtenidos y aportar una visión distinta de la tarea. En total se generaron 10 conjuntos distintos de datos (90% del material para entrenamiento y 10% para evaluación).

En los experimentos descritos en este artículo haremos referencia a la tarea VESTEL-L como aquella que analiza el resultado medio para las 10 particiones de evaluación de VESTEL, usando obviamente los modelos entrenados con la partición de entrenamiento asociada..

3 DICCIONARIOS

Para conseguir una variabilidad suficiente en la composición de los diccionarios sobre los que evaluar nuestras pruebas, se diseñaron distintos grupos de diccionarios compuestos por 1175, 1996, 5000 y 10000 palabras, con distintas variaciones de los mismos en cuanto a su composición, de forma que tuviéramos un amplio rango de variaciones que estudiar.

El diccionario básico dependiente del vocabulario de VESTEL está compuesto por 1175 palabras distintas, y el independiente del contexto por 1996. Para llegar a completar los diccionarios hasta llegar a las 5000 o 10000 palabras, se añadieron nuevas entradas procedentes de los diccionarios obtenidos en el proyecto ONOMASTICA [5].

Los criterios de selección de nuevas entradas tenían en cuenta la longitud media de las palabras del diccionario resultante, dado que éste parámetro es uno de los que planteamos como objetivo a estudiar y era fácilmente seleccionable.

4 MOTIVACIÓN Y EXPERIMENTOS DE REFERENCIA

En la Figura 1 se muestran las curvas de tasa error de inclusión para la tarea de reconocimiento con un diccionario de 10000 palabras sobre PRNOK5TR, PERFDV y PEIV1000, usando los sistemas no integrados con estrategia ascendente (generación de cadena fonética seguida de acceso léxico) y sistemas integrados guiados por la estructura del léxico descritos en [4]. Si tenemos en cuenta que, como se ha comentado en el apartado 2, PEIV1000 es una base de datos diseñada para hacer experimentación en tareas independientes del vocabulario de entrenamiento, resulta sorprendente que se comporte mejor que PERFDV, cuyas palabras (grafemas) sí han sido vistas en la lista de entrenamiento y sería esperable que los modelos acústicos estuvieran mejor adaptados.

Evidentemente hay factores que influyen en dicho comportamiento y es nuestro objetivo en este artículo plantear algunas ideas al respecto.

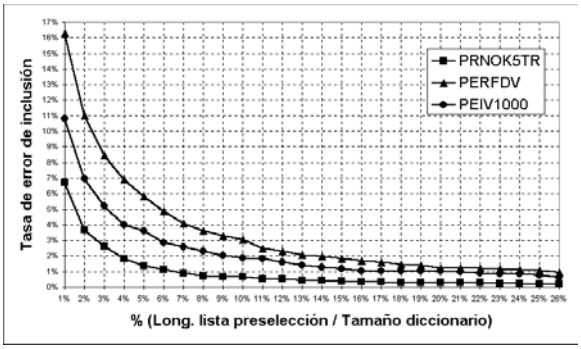

Figura 1. Tasa de error de inclusión para las tres bases de datos en función del tamaño relativo de la lista de preselección. Dicc. de 10000 palabras.

Nuestros esfuerzos en este sentido se orientaron en las siguientes líneas de actuación:

- Estudio del impacto en la tasa de reconocimiento del sistema de factores ligados exclusivamente a la composición del diccionario
- Estudio del impacto en la tasa de reconocimiento del sistema de factores

ligados exclusivamente a la composición de las bases de datos utilizadas

5 CRITERIOS DE DIFICULTAD

En [4] se discute sobre la posible correlación entre parámetros relacionados con la longitud de la palabra a reconocer con la tasa de reconocimiento final del sistema. Nuestro punto de partida aquí es precisamente ése, pero evaluado a partir del conjunto de formas grafémicas asociadas a diccionarios y a listas de bases de datos, calculando las posibles correlaciones entre el rendimiento de un sistema dado y las longitudes medias de aquellas.

Además de esa información, se plantearon parámetros adicionales completando el repertorio a estudiar, basándonos en la idea de que lo que perseguíamos era una medida de la confusabilidad (o dificultad, en definitiva) de cada diccionario o base de datos. Dichas medidas surgen de forma natural a partir de los sistemas estudiados en [4], usando los módulos de acceso léxico (que nos permiten evaluar el impacto de costes de alineamiento entrenados con el soporte acústico real del que disponemos, además de los estándar (Levenstein, por ejemplo)), sin entrar en más consideraciones acústicas ya que nuestro objetivo se centra en no llegar más allá.

Así, el repertorio final de parámetros sobre los que evaluaremos la correlación con las tasas de reconocimiento, es el siguiente:

- Longitud media de las palabras del diccionario o la lista dada
- Media de la diferencia entre el coste de acceso léxico para la palabra dada y la segunda en la lista de preselección, usando la distancia de Levenstein para los costes de alineamiento
- Media de la diferencia entre el coste de acceso léxico para la palabra dada y la segunda en la lista de preselección, usando los costes entrenados para la tarea considerada

Las diferencias de costes de acceso léxico se calculan sobre cada diccionario o lista particular alineando cada entrada del diccionario o la lista con todas las demás del diccionario o la lista. El uso de la distancia de Levenstein nos permite hacer una comparación en la que no intervienen medidas de confusabilidad fonética de nuestro modelo acústico, mientras que el uso de costes entrenados sí que aporta dicha información contenida.

Una alternativa al cálculo propuesto sería el enfrentar las listas con los diccionarios, lo que nos daría una medida más directa de complejidad de la tarea, pero nuestra intención es hacer el estudio de forma independiente.

En los siguientes apartados, se verán experimentos que analizarán la correlación entre las variables propuestas como relevantes de cara a medir la complejidad de un diccionario o lista, y los rendimientos obtenidos en tareas distintas.

6 EXPERIMENTOS SOBRE DIFICULTAD DE DICCIONARIOS

Los experimentos llevados a cabo en este apartado se realizaron tomando como puntos de evaluación los conseguidos con las listas de evaluación de VESTEL-L descritas en el apartado 2, así como el valor medio para todas ellas. En cada caso se estimaron los parámetros discutidos como relevantes en el apartado 5 y se estudiaron las correlaciones entre los mismos y, en este caso, los valores de tasa de inclusión obtenidos para longitudes de lista iguales a un 1% y un 10% del tamaño del diccionario usado. Los diccionarios estaban compuestos por 1996, 5000 o 10000 palabras, con distintas variantes para cada uno de ellos, lo que nos proporciona un amplio rango de efectos y valores.

1.1 Experimentos con parámetros dependientes de los diccionarios

En la Figura 2 se muestra un ejemplo de la tendencia de los valores de tasa de inclusión para una longitud de lista del 1% del tamaño del diccionario, (usando un sistema no integrado y modelado semicontinuo independiente del contexto). Como puede observarse, la longitud media del diccionario es un factor relevante para evaluar el rendimiento previsto de una tarea que lo use. La tendencia observada se mantiene también para distintas longitudes de lista (referidas como siempre al tamaño del diccionario).

Es interesante hacer notar que la dependencia vista se mantiene cambiando el diccionario y manteniendo el resto de condiciones de evaluación: misma lista de evaluación, misma arquitectura, mismo

modelado, mismo número de entradas en el diccionario, etc.

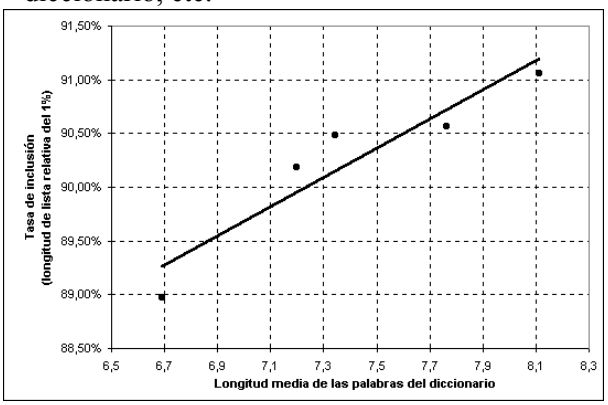

Figura 2. Tendencia del efecto de la longitud media de las palabras del diccionario en la tasa de inclusión (arquitectura no integrada)

En la Tabla 1 se muestra la tasa de inclusión para distintos candidatos (medidos respecto al tamaño del diccionario) evaluando el comportamiento medio para las listas de evaluación de VESTEL-L, usando diccionarios de 5000 y 10000 palabras con distinta longitud media de palabra (LMP).

El análisis de validez estadística muestra solape de bandas para puntos contiguos en la figura 2, pero no para los puntos extremos. Además, enfatizamos lo sistemático de los resultados, coherentemente mejores para todos los casos analizados (para modelos discretos y semicontinuos y todos los alfabetos disponibles) cuando se usaba un diccionario con el mismo número de entradas y mayor longitud media de palabra.

	Dicc 5000 LMP 7,76	**Dicc 5000 LMP 7,72**	**Dicc 10000 LMP 7,76**	**Dicc 10000 LMP 7,2**
Cand 1%	86,10%	85,68%	86,76%	85,80%
Cand 10%	97,66%	97,45%	97,82%	97,53%

Tabla 1. Tasas de inclusión para la tarea VESTEL-L completa en función del diccionario utilizado (LMP=longitud media de palabra)

Este hecho, junto con la tendencia observada en la Figura 2 da una clara idea de la necesidad de establecer diccionarios homogéneos, al menos en cuanto a longitud media de palabra si nuestro objetivo es realizar comparaciones entre tareas con dependencia o independencia del vocabulario (o cualquier otra en la que haya un cambio de diccionario implicado).

En la Figura 3 se muestra el efecto de la longitud media de las palabras del diccionario en la tasa media de inclusión para un tamaño de lista del 0'5% del tamaño del diccionario, usando el sistema integrado con modelos semicontinuos dependientes del contexto. Como puede verificarse, se repite la tendencia vista para el caso del sistema no integrado visto en la Figura 2, si bien en este caso las bandas de fiabilidad de todos los resultados se solapan, dadas las altas tasas alcanzadas y la falta de bases de datos mayores.

De los otros parámetros analizados, la diferencia de coste medio con la segunda palabra mejor reconocida usando la distancia de Levenstein no mostró correlación ninguna y lo mismo cabe decir de su versión normalizada por el número de símbolos.

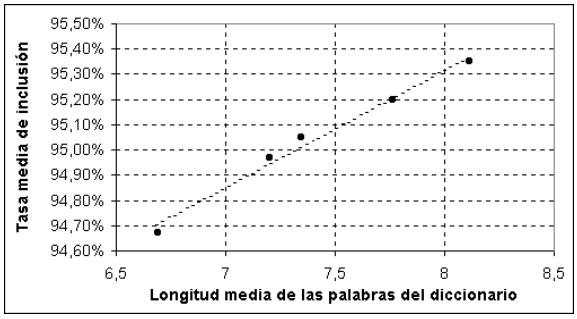

Figura 3. Tendencia del efecto de la longitud media de las palabras del diccionario en la tasa media de inclusión (arquitectura integrada)

Tampoco se encontró correlación apreciable al utilizar las diferencias medias de costes usando los costes entrenados para la tarea (ni la correspondiente versión normalizada).

Este resultado es sorprendente, ya que esperábamos que esa medida de complejidad, dependiente además de costes que habían visto información de la parte acústica sería mucho más relevante. En el apartado siguiente detallamos algún experimento adicional orientado a ofrecer alguna explicación a este efecto.

1.2 Experimentos con parámetros dependientes de las listas (bases de datos usadas)

El mismo análisis que el hecho en el apartado anterior se abordó para parámetros relativos al conjunto de palabras de evaluación. La dependencia con la longitud media de las palabras de evaluación se muestra en la Figura 4 donde, al contrario que en el caso anterior, no hay ninguna correlación apreciable, lo que da

idea de la independencia del rendimiento del sistema frente a la longitud media de las palabras de la lista a reconocer.

Al usar los parámetros dependientes de la distancia de coste medio tampoco encontramos correlaciones, lo que coincide con la observación hecha en el apartado anterior.

En este punto es importante hacer notar que las bases de datos de las que disponemos sólo cubren un vocabulario de unas 2000 palabras, con lo que la artificialidad de la generación de diccionarios usada introduce una dificultad adicional si queremos extrapolar los resultados vistos aquí. Nuestra intuición en el caso de una tarea en la que dispusiéramos de todos los ejemplos acústicos de las palabras del diccionario es que el impacto de medidas dependientes del conjunto de palabras a reconocer no tendrán una mayor correlación con el resultado esperado que la vista aquí (en el apartado anterior).

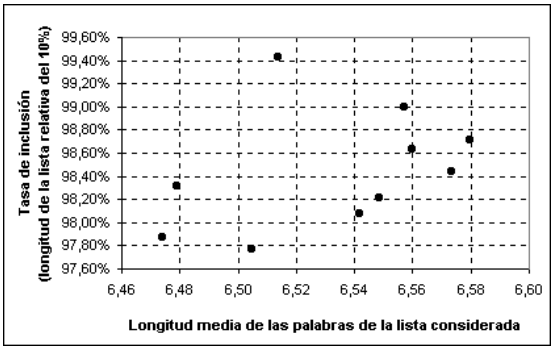

Figura 4. Efecto de la longitud media de las palabras a reconocer en la tasa de inclusión.

7 CONCLUSIONES Y LÍNEAS FUTURAS

En este artículo se ha hecho una incursión en el estudio del concepto de dependencia e independencia del vocabulario, lo que nos ha dado pie a abordar la problemática de la dificultad de diccionarios, proponiendo medidas concretas para evaluar la misma y verificando la alta correlación existente entre las tasas obtenidas en una tarea con la longitud media de las palabras de los diccionarios dados.

La principal conclusión a este respecto es que los diccionarios deben ser iguales en cuanto al tamaño medio de las palabras que lo componen si queremos hacer comparaciones homogéneas.

De cara al futuro, una propuesta adicional de las medidas de diferencia de coste de acceso léxico es el diagnóstico y asesoramiento sobre la calidad y/o dificultad de diccionarios dados. En vocabularios pequeños es planteable un diseño manual, pero cuando se usan grandes diccionarios, es imprescindible contar con alguna metodología de diagnóstico más objetiva y que permita realizar el proceso automáticamente. A partir de un diccionario, sería posible identificar las palabras más problemáticas en cuanto a confusabilidad, seleccionando aquellas con la menor distancia mutua. Dicha medida es más efectiva todavía si usamos costes de alineamiento entrenados para el tipo de modelado considerado. Esta estrategia de medida valdría igualmente para identificar elementos problemáticos de un diccionario, de cara a robustecer su modelado o, incluso, establecer modelos específicos para aquellos.

Por último, pretendemos aplicar las ideas sobre complejidad de diccionarios y medidas de dificultad a la predicción de tasas de reconocimiento sobre diccionarios no vistos, en la línea de los trabajos descritos en [6] y [7].

8 BIBLIOGRAFÍA

[1] Hon, H. y Lee, K. On Vocabulary-Independent Speech Modeling. in: IEEE ICASSP. Albuquerque, NM, 1990, pp. 725-728..

[2] Macías-Guarasa, J., Gallardo, A., Ferreiros, J., Pardo, J.M. and Villarrubia, L. "Initial Evaluation of a Preselection Module for a Flexible Large Vocabulary Speech Recognition System in Telephone Environment". ICSLP 96. pp. 1343-1346. 1996.

[3] Tapias, D., Acero, A., Esteve, J., and Torrecilla, J.C. "The VESTEL Telephone Speech Database". ICSLP 94: 1811-1814. 1994

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

[5] Trancoso, I.M. 1995. The ONOMASTICA Inter-Language Pronunciation Lexicon. Proceedings of the European Conference on Speech Technology. Madrid, 1995.

[6] P. Laface, L. Fissore, y F. Ravera. "Automatic Generation of Words Toward Flexible Vocabulary Isolated Word Recognition". Proc. of the International Conference on Spoken Language Processing (ICSLP), 1994, Yokohama, pp. 2215-2218. 1994.

[7] Roe, D.B y M.D. Riley. "Prediction of Word Confusabilites for Speech Recognition". Proc. of the International Conference on Spoken Language Processing (ICSLP), 1994 , pp 227-230. 1994.