Descripción de un motor de reconocimiento para habla continua

Antonio Cardenal López *Carmen García Mateo*

Departamento de Tecnoloxías das Comunicación
Universidade de Vigo
cardenal@gts.tsc.uvigo.es carmen@gts.tsc.uvigo.es

RESUMEN

In this paper we describe the recognition engine that is been developed in the University of Vigo[1]. The system has been designed to be used in continuous speech, large vocabulary applications, like automatic transcription and audio indexing.

1. INTRODUCCIÓN

Los avances obtenidos en los últimos años en la investigación de nuevos algoritmos y el aumento en la capacidad de los ordenadores, han permitido solucionar de forma satisfactoria el problema del reconocimiento de voz en muchos de los escenarios propuestos tradicionalmente. Sin embargo el desarrollo de un sistema sin restricciones para habla continua y en tiempo real, todavía pertenece al campo de la investigación.
Para lograr este objetivo, los reconocedores que se están desarrollando hoy en dia utilizan toda una serie de técnicas más o menos sofisticadas que implican el manejo de grandes cantidades de datos, lo que hace necesaria una programación altamente estructurada y optimizada. A continuación describimos el motor de reconocimiento para habla continua que está actualmente en desarrollo en este departamento y presentamos algunos de los resultados obtenidos hasta ahora.

2. DESCRIPCIÓN DEL RECONOCEDOR

El reconocimiento se realiza en dos fases diferenciadas. La primera consiste en una búsqueda en haz aplicando el algoritmo de Viterbi en forma síncrona. El espacio de búsqueda para esta primera fase se construye mediante la interconexión de una serie de nodos, cada uno de los cuales representa un fonema de una palabra del vocabulario y lleva asociado uno o varios modelos de Markov. Para reducir la complejidad, el vocabulario se organiza en forma de árbol. Existen tantos nodos iniciales (raíces) como fonemas, y tantos nodos finales (hojas) como transcripciones consideradas en el vocabulario. De esta forma se reduce de manera importante el número de nodos en el espacio y por tanto el coste computacional, pero se dificulta la aplicación del modelo de lenguaje, como se comentará más adelante.
La propagación de caminos propiamente dicha se realiza por medio de testigos [1]. Cada testigo es una estructura que representa un camino activo desde el comienzo del reconocimiento. Incluye diversa información acerca del nodo, modelo y estado en el que se encuentra, así como otros datos que facilitan la aplicación del modelo de lenguaje. Los testigos se propagan dentro de los modelos y entre los nodos según el algoritmo de Viterbi. Para limitar la búsqueda se utilizan varios metodos de poda basados en las puntuaciones acústica y lingüística del camino completo y de la palabra.
Para permitir el reconocimiento de varias palabras consecutivas es necesario realimentar el árbol. Empleamos para esto unos nodos especiales que almacenan todos los testigos que han alcanzado las hojas en un mismo instante de tiempo. De todos estos caminos sólo el mejor será propagado. Los demás se guardan en memoria y serán utilizados en una fase posterior.
Debido a la organización en árbol, la aplicación del modelo de lenguaje resulta complicada (no se conoce la identidad de la palabra hasta que se alcanza un nodo final). Puesto que el modelo de lenguaje, además de aumentar de forma considerable la tasa de reconocimiento, limita de manera significativa el número de testigos activos en el árbol, resulta conveniente su aplicación lo más temprana posible. Para esto se utiliza una técnica de *Look-Ahead* [2], que consiste básicamente en aplicar a cada testigo la probabilidad de la palabra predicha por el modelo, dada una historia de palabras y un nodo intermedio. Este mecanismo puede ser extremadamente costoso y requiere búsquedas rápidas para localizar cada probabilidad. En este reconocedor se ha comprobado que el coste se puede reducir a niveles aceptables utilizando tablas *hash* para el acceso a bigramas y trigramas e instalando una pequeña memoria cache en cada nodo.
Una vez terminada la fase síncrona de reconocimiento, se realiza una segunda búsqueda teniendo en cuenta las palabras no propagadas pero almacenadas en los nodos especiales. No resulta necesario un recálculo de las puntuaciones acústicas, pero sí de las probabilidades del modelo de lenguaje. Aunque el número de hipótesis que se extraen en esta fase puede ser enorme, se ha comprobado que utilizando un algoritmo de tipo *A* Stack*, el coste de esta búsqueda resulta pequeño, proporcionando sin embargo una mejora de varios puntos en la tasa de reconocimiento.

3. RESULTADOS PRELIMINARES

Presentamos en esta sección algunos resultados obtenidos en dos escenarios de prueba diferentes. El primero de ellos consiste en una aplicación de habla conectada con un vocabulario pequeño y un modelo de lenguaje estocástico perfectamente adaptado a la aplicación. En el segundo se muestran algunos resultados preliminares en un entorno de grandes vocabularios.

3.1. Marco experimental

La base datos de test está formada por frases cortas en idioma gallego, con voz telefónica y locutores masculino. Consiste en 936 ficheros con un vocabulario de 923 palabras. Se utilizaron monofonemas entrenados con voces masculinas y femeninas, con

[1]Este trabajo ha sido parcialmente financiado por la CICYT bajo el proyecto 1FD97-0077-C02-01

	Experimento 1			Experimento 2		
Poda	100	130	170	100	130	170
%Ac.	88.79	93.69	95.29	65.63	67.42	67.80
%Prec.	88.04	93.13	94.82	58.22	59.76	60.01
FTR	0.18	0.29	0.45	0.77	1.48	3.21

Tabla 1: *Tasa de acierto, precisión y factor de tiempo real.*

una parametrización que incluye energía, coeficientes cepstrales y sus derivadas primera y segunda. Existen tres estados por modelo y ocho mezclas por estado. En total disponemos de 26 modelos que representan fonemas y 5 para distintos tipos de silencio y ruidos. Los experimentos se realizaron sobre un procesador Pentium III a 650 MHz, con 125 Mb de memoria RAM, sobre sistema operativo Linux.

Los resultados pueden verse en las tablas 1 y 2. En la primera de ellas se muestra el porcentaje de palabras acertadas, la precisión (donde se tiene en cuenta las inserciones) y el factor de tiempo real (fracción del tiempo total disponible empleada en el reconocimiento), para tres niveles de poda diferentes. En la segunda tabla mostramos el porcentaje de tiempo empleado por las rutinas más importantes en el primer test.

Ambos experimentos han sido realizados sobre el mismo conjunto de frases. Para el primero se utilizó un modelo de lenguaje ideal, en el sentido de que fue extraído de las propias transcripciones, consistente en 1687 bigramas y 2586 trigramas. Para el segundo experimento se creó un modelo de lenguaje de texto periodístico, al que se le incluyeron las transcripciones para reducir la perplejidad, formado así por 21400 palabras, 251000 bigramas y 194200 trigramas. En este último caso la perplejidad de la tarea resulta bastante baja (15.97), aunque como contrapartida hay un 9.41% de palabras fuera de vocabulario.

3.2. Resultados de los experimentos

En el primer experimento el reconocimiento se realiza en tiempo real para todos los niveles de poda considerados. Las rutinas con mayor coste computacional son las que se encargan del cálculo de las probabilidades de observación, aunque su coste se reduce a medida que aumenta el nivel de poda. La razón está en la utilización de monofonemas. En total sólo existen 75 estados diferentes, por lo que a partir de un número de testigos propagados por trama suficientemente grande, no es necesario calcular más probabilidades. Lo mismo sucede en el segundo test, aunque el efecto aparece ya en el nivel de poda más bajo. Hay que tener en cuenta que al utilizar modelos más complejos, este porcentaje aumentará sensiblemente.

Para el segundo experimento sólo se obtiene una ejecución en tiempo real para el nivel de poda más bajo. Aunque no se muestran aquí los porcentajes obtenidos, las rutinas de cálculo del modelo de lenguaje aumentan considerablemente su coste debido al gran número de n-gramas considerados.

4. LÍNEAS FUTURAS

El motor de reconocimiento que acabamos de describir se pretende utilizar en dos aplicaciones diferentes. La primera de ellas consiste en un sistema de indexado de audio para noticiarios televisivos [3], que está comenzando a desarrollarse en este grupo de investigación. La segunda, considerablemente más compleja, es una tarea de dictado automático a través de

Poda	100	130	170
Cálculo LM	4.94	6.06	8.05
Cálculo Probabilidades	58.19	51.88	39.66
Propagación Testigos	17.11	23.18	35.32
Segundo Pase	8.62	11.67	12.12
Resto	11.14	7.26	4.86

Tabla 2: *Ocupación de rutinas para el experimento 1.*

la línea telefónica, que se pretende incluir en un sistema de consulta de correo electrónico por línea telefónica desarrollado en este departamento [4].

Respecto al trabajo actual, consideramos dos líneas principales. Por una parte el sistema todavía no ha sido validado en un entorno de test para habla continua adecuado. En estos momentos se esta adquiriendo una nueva base de datos de habla continua, en un entorno de texto periodístico, que será utilizada para realizar los tests adecuados.

La segunda línea de trabajo es la utilización de modelos más complejos. Ya han sido realizados experimentos con trifonemas y semifonemas, optándose finalmente por el empleo de demifonemas [5], que ofrecen buenos resultados sin complicar de forma excesiva la construcción del espacio de búsqueda. Los demifonemas deben mejorar los porcentajes de reconocimiento, pero aumentando considerablemente la carga computacional, debido por una parte al aumento de nodos en el sistema, y por otra al crecimiento en el número de modelos a considerar. Este aumento puede ser paliado de manera efectiva utilizando métodos eficientes de cálculo de las probabilidades de observación mediante la cuantificación de gausianas, tal como se describe en [6]. Estos métodos estan siendo probados actualmente en el reconocedor que se describe.

5. REFERENCIAS

[1] S. J. Young, N. H. Russell, and J. H. S. Thornton. Token passing: a simple conceptual model for connected speech recognition systems. Technical Report 38, Cambridge University Engineering Dept., July 1989.

[2] C. Lee, F. Soong, and K. Paliwal, editors. *Automatic Speech and Speaker Recognition, Advanced Topics*. Kluwer Academic Publishers, 1996.

[3] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava. Speech and language technologies for audio indexing and retrieval". *Proceedings of the IEEE*, 88:1338–1353, August 200.

[4] Leandro Rodriguez, Antonio Cardenal, Carmen Garcia, David Perez, Eduardo Rodriguez, and Xabier Fernandez. Telcorreo: A bilingual e-mail client over the telephone". In *TSD 2000. Brno (Czech Republic)*, pages 1215–1218, September 2000.

[5] J. Mariño, A. Nogueiras, and A. Bonafonte. The demiphone: an efficient subword unit for continuous speech recognition. In *EuroSpeech*, pages 1215–1218, 1997.

[6] Enrico Bocchieri. Vector quantization for the efficient computation of continuous density likelihoods. In *ICASSP*, pages 692–695, 1993.