EVALUACIÓN ALBAYZÍN-08 DE SISTEMAS DE VERIFICACIÓN DE LA LENGUA: SISTEMA DEL GRUPO SOFTLAB DE LA UC3M

M.J. Poza, B. Ruiz, L. Puente y D. Carrero

Grupo SoftLab, Universidad Carlos III de Madrid
mjpoza@entornotec.com, bruiz@inf.uc3m.es, lpuente@it.uc3m.es, dcarrero@di.uc3m.es

RESUMEN

El grupo SOFTLAB de la UC3M ha desarrollado un sistema de identificación de lengua basado en GMMs, y lo ha presentado a la Evaluación ALBAYZIN-08 de Sistemas de Verificación de la Lengua, organizada por el Grupo de Trabajo en Tecnologías Software de la UPV/EHU, en el marco de las V Jornadas en Tecnología del Habla organizadas por la Red Temática en Tecnologías del Habla y el Grupo AHOLAB de Procesado de Señal de la UPV/EHU. Este artículo describe el sistema de identificación presentado a dicho plan de evaluación.

1. INTRODUCCIÓN

Un sistema de identificación de lengua es, básicamente, un sistema de reconocimiento de patrones que hace uso de la señal de voz de un discurso o conversación inteligible de cualquier individuo para decidir si el idioma utilizado en la conversación o discurso es alguno de los que el sistema reconoce.

Todo sistema de clasificación basado en reconocimiento de patrones (ver Figura 1) tiene una fase previa de entrenamiento en que se capturan y modelan las características distintivas de cada uno de los 'usuarios' del sistema (en este caso, idiomas): en esta fase se generan los patrones o modelos que luego se usarán durante el funcionamiento normal, en la toma de decisión sobre si el discurso a clasificar está pronunciado en alguno de los idiomas para los que el sistema ha sido entrenado:

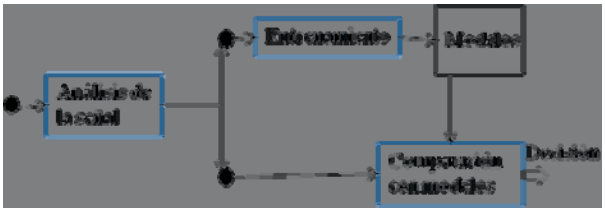

Figura 1. *Sistema de clasificación basado en reconocimiento de patrones*

Generalmente se distingue entre verificación e identificación de la lengua. Dado un segmento de habla Y, y un hipotético idioma I, la tarea de verificación consiste en determinar si el segmento Y fue pronunciado en el idioma I. La tarea de identificación considera un conjunto cerrado de idiomas (I_1,I_2... I_N), y trata de determinar en cuál de ellos fue pronunciado el segmento de habla Y.

Nuestro sistema de identificación de lengua actúa como un clasificador de patrones. Cada patrón está formado por un conjunto de características o parámetros, extraídos de una determinada locución, y es 'enfrentado' o comparado con distintos modelos generados para cada idioma. La salida del clasificador ofrece una verosimilitud o una medida de distancia, entre el patrón de entrada y el modelo; y en última instancia una decisión, basada en un umbral, que clasifica la locución como perteneciente o no a un determinado idioma.

Cada modelo de un idioma es generado mediante patrones extraídos de locuciones del mismo; siendo necesario que cada uno de los idiomas involucrados en el sistema, disponga de su propio conjunto de datos de entrenamiento. Este conjunto será distinto del conjunto de datos sobre los cuales se pruebe el sistema.

Es bien conocido, que una de las causas principales que degradan el rendimiento de los sistemas de reconocimiento basados en voz se debe a la variabilidad acústica entre los conjuntos de entrenamiento y test. Esta variabilidad no sólo es debida a la diferencia acústica en los distintos idiomas (en sistemas de reconocimiento de idioma), sino también a otro tipo de variaciones, como las distorsiones producidas por los distintos canales, las diferencias entre micrófonos, el ruido ambiental, etc. El uso de técnicas de compensación de canal, ya sea sobre el audio, los parámetros a modelar o el propio modelo, mejora las tasas de reconocimiento. Estas técnicas se basan en eliminar información no discriminativa que varía de forma no controlada entre las distintas locuciones.

Existen diversas y muy variadas técnicas aplicadas a la compensación o eliminación de la variabilidad de canal; Nuestro sistema se basa en la técnica CMS (cepstral mean substraction), también conocido como CMN (cepstral mean normalization). En una

parametrización basada en coeficientes cepstrales [1], una locución, es dividida en ventanas de tiempo, de la cual son extraídos un cierto número de coeficientes cepstrales. CMN se basa en sustraer para cada coeficiente cepstral extraído la media de dicho coeficiente a lo largo de toda la locución. De esta forma se reduce la distorsión introducida por elementos de variación lenta, como por ejemplo ruido estacionario.[2].

Los sistemas de reconocimiento de idioma sobre habla 'espontánea' tienen más limitaciones que los sistemas que usan voz 'limpia', como son el ruido de las conversaciones (ruidos de sillas, música, conversaciones de fondo, etc) y los silencios en las mismas. Además cuanto mayor sea el nivel de reconocimiento requerido mayor tendrá que ser la duración de la conversación.

2. DESCRIPCIÓN DEL SISTEMA

El diagrama de bloques del sistema verificador de idioma se muestra en la siguiente figura:

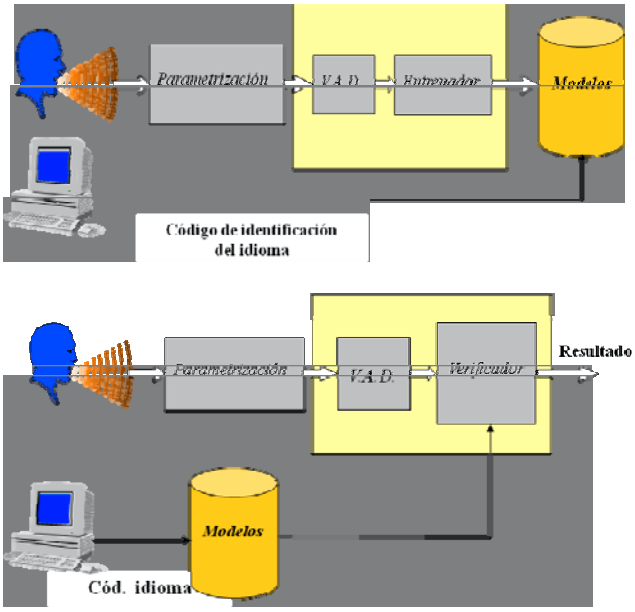

Figura 2. *Fases de entrenamiento y verificación del sistema de identificación de idioma.*

El reconocimiento automático del idioma comparte muchas técnicas con el reconocimiento de locutor, lo que hace que ambos problemas suelan ser abordados de un modo similar. Nuestro sistema se compone, básicamente, de tres módulos funcionales: parametrización (o extracción de características), detector de actividad vocal (para eliminar 'silencios' de la señal de entrada al entrenador y al verificador) y generación/comparación con mdelos.

2.1. Extracción de características

La extracción de características es el primer paso en cualquier sistema de reconocimiento automático. y comienza por la captación de la señal sobre la que se desea trabajar (voz) mediante un sensor, que en este caso será un sensor apto para la señal de voz (micrófono, teléfono, etc.).

La extracción de parámetros de nuestro sistema está basada en el análisis a corto plazo de la señal de voz, usando una de las técnicas más habituales en reconocimiento automático de voz: el análisis MFCC *(Mel-Frequency Cepstral Coefficients)*, junto con varias medidas de entropía [3] (la entropía de la señal, de su potencia y del logaritmo de su potencia, así como sus primeras derivadas respectivas). Así, los vectores de características estarán principalmente compuestos por algunos parámetros cepstrum y cepstrum diferenciales, pero también se añadirá otro tipo de información, como la energía, su derivada, y los valores de las entropías antes mencionados.

Sobre estos parámetros pueden llevarse a cabo mejoras con el fin de paliar las distorsiones sufridas por la señal de voz y mejorar el rendimiento de los sistemas, estas mejoras son conocidas como compensaciones de canal [4] y en nuestro sistema se utiliza la conocida como CMN *(Cepstral Mean Normalization)* [1], consistente en eliminar la media de los parámetros cepstrales con el fin de eliminar de ellos información de variación más lenta que la de la señal de voz, como puede ser la introducida por el ruido de fondo (ruido aditivo, no correlado con la señal de voz y casi estacionario en el tiempo). Esa técnica de eliminación de medias se ha empleado también para la energía y para la entropía de la señal.

La Figura 3 muestra un diagrama de bloques del módulo extractor de características.

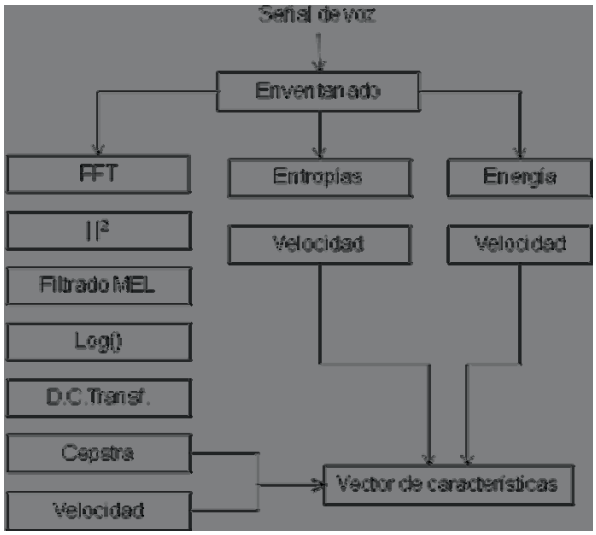

Figura 3. *Bloque extractor de características*

2.2. Detector de actividad vocal

Una señal vocal comprende hasta un 60% de silencio o de ruido de fondo, y ese 60% de la señal tiene características muy similares para todos los idiomas: por ello, eliminar de la señal de entrada esos tramos de

'solo-ruido' aumentará la eficiencia y la eficacia de los sistemas de reconocimiento, por aumentar el poder de discriminación entre los distintos idiomas (eliminando 'partes comunes') y disminuir la carga de trabajo del sistema. El módulo detector de actividad vocal se encarga de esa eliminación de silencios y/o ruidos.

Clásicamente hay dos formas de abordar la eliminación de silencios en una señal [5] [6]:

- Basándose en umbrales (umbrales para la energía, umbrales para los 'cruces por cero', etc: información 'local')
- Usando técnicas de clasificación (como pueden ser los modelos ocultos de Markov o las redes neuronales): esta familia de detectores de silencios se basa en características estadísticas de la señal en vez de en características locales.

El primer grupo de detectores de actividad vocal es el más utilizado, por su simplicidad de implementación, pero precisa ajustar varios umbrales para que funcione adecuadamente, y que esos umbrales se vayan adaptando a las características de la señal de entrada: a las variaciones ambientales. El segundo grupo de detectores elimina esa dependencia, con el coste de una fase previa de entrenamiento del clasificador.

El sistema propuesto por el grupo SOFTLAB incluye un módulo detector de actividad vocal basado en la energía de la señal, implementado como una máquina de 6 estados cuyas transiciones se realizan cuando la señal cruza por alguno de los 4 umbrales de energía que utiliza (dos para decidir transición silencio-voz y otros dos para decidir voz-silencio), dentro de ciertas limitaciones temporales.

2.3. SVMs: Módulo Decisor

En los últimos años se ha observado un incremento considerable de la utilización de las Support Vector Machines dentro del Aprendizaje Automático [7][8][9]. Su alto rendimiento hace de las SVM una de las metodologías más sólidas en este área.

Las SVM son básicamente clasificadores para 2 clases. Tienen como objetivo obtener un hiperplano óptimo capaz de separar lo mejor posible dos clases distintas de los puntos de entrada. Como un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un *kernel* Gaussiano u otro tipo de *kernel* a un espacio de características en un espacio de más dimensiones, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en distintas clases, cada una formando un agrupamiento.

El kernel utilizado en el sistema de identificación de idioma presentado por los autores del presente artículo es el RBF (Radio Basis Function) o gaussiana:

$$K(x, z) = \exp(- |x - z|^2 / \sigma^2)$$

3. AGRADECIMIENTOS

Este trabajo ha sido posible gracias a la financiación de los Ministerios de Industria (proyecto SEGUR@) y de Educación y Ciencia (proyecto TEC2006-12365-C02-01).

4. BIBLIOGRAFÍA

[1] Furui S., ”Cepstral analysis technique for automatic speaker verification”. IEEE Transactions on Speech and Audio Processing,Vol.ASSP-29,No.2 .April 1981

[2] Liu F., Stern R., Huang X. and Acero A. “Efficient Cepstral Normalization for Robust Speech Recognition”. Proceedings of ARPA Human Language Technology Workshop, March 1993.

[3] Cunha, S., Correira, S., Aguilar, B. "Pathological voice discrimination based on entropy measurements", BIODEVICES 2008: International Conference on Biomedical Electronics and Devices (Madeira, Portugal).

[4] Vaier C., Colibro D., Castaldo F., Dalmasso E., Laface P., "Channel factors compensation in model and feature domain for speaker recognition”. Odissey 2006

[5] Stadermann J., Stahl V., Rose G., “Voice activity detection in noisy environments”, Proc. of Eurospeech 2001.

[6] Carli G., Gretter R., “A start-end point detection algorithm for a real-time acoustic front-end based on DSP32C VME board”, ICSPAT’92.

[7] Campbell W. M., "Generalized linear discriminant sequence kernels for speaker recognition", Proceedings of the International Conference on Acoustics Speech and Signal Processing, 2002, pp. 161.164

[8] Vapnik V., "The nature of statistical learning theory". Springer, 1995.

[9] Burges, C. "A Tutorial on Support Vector Machines for Pattern Recognition", 1998 Journal on Data Mining and Knowledge Discovery, Vol 2, pp 121-167.