EL SISTEMA DE IDENTIFICACIÓN DE LA LENGUA DE PRHLT

Míriam Luján-Mares, Vicent Tamarit, Roberto Paredes,
Vicent Alabau, Carlos-D. Martínez-Hinarejos

Instituto Tecnológico de Informática, Universidad Politécnica de Valencia,
Camino de Vera, s/n, 46022, Valencia, Spain

RESUMEN

El artículo explica los detalles de los sistemas de identificación de la lengua del grupo de investigación *Pattern Recognition and Human Language Technology* (PRHLT).

1. INTRODUCCIÓN

La tarea consiste en reconocer la lengua de un fragmento de habla, adquirido de un programa de televisión de una de las 4 lenguas objetivo (castellano, catalán, euskera y gallego) o una lengua desconocida.

La verificación de la lengua es un problema que puede ser estudiado desde dos puntos de vista: utilizando la información espectral de la señal o la información sintáctica y semántica de dicha señal.

Los sistemas que se presentan en este trabajo se basan solamente en la información espectral de la señal, ya que han sido entrenado únicamente con los datos proporcionados. Dichos datos contaban sólo con habla efectiva y su idioma correspondiente. Por tanto, no se han empleado corpora externo para la creación de modelos de lenguaje utilizables para un reconocimiento de habla, ni tampoco para mejorar modelos basados en señal.

Para ambos sistemas los datos han sido preprocesados con técnicas del estado del arte en verificación de la lengua. Los sistemas se basan en dos clasificadores diferentes para llevar a cabo la identificación: uno de ellos por k-vecinos y el otro por *Gaussian mixture models* (GMMs).

2. DATOS DE ENTRENAMIENTO

Los datos de entrenamiento provienen de programas de televisión (informativos, documentales, debates, entrevistas, reportajes, magazines, etc.). Dichos datos sólo cuentan con las cuatro lenguas objetivo (castellano, catalán, euskera y gallego).

Las señales se han adquirido a través de un mismo dispositivo (una grabadora digital Roland Edirol R-09) y se han depositado en ficheros WAV (monocanal, 16 Khz, 16 bits/muestra). Las grabaciones incluyen diversos tipos de habla: leída, planificada, conversacional formal, espontánea, etc. Asimismo, aunque la relación señal-ruido (SNR) es bastante buena en casi todos los casos, las condiciones ambientales y de canal son también muy diversas: entrevistas en estudio sin ruido de fondo, reportajes desde la calle, desde una fiesta, desde una manifestación, llamadas telefónicas en directo, reportajes con una ligera música de fondo, programas concurso o de humor con risas y aplausos, etc.

El conjunto de entrenamiento consta de aproximadamente 8 horas por lengua (unas 32 horas en total), en ficheros de duración variable. Estos ficheros contienen mayoritariamente voz (en condiciones ambientales y de canal diversas) y sólo pequeños fragmentos de silencio o ruido de fondo [1].

Para encontrar los parámetros idóneos para el sistema hemos utilizado un conjunto de desarrollo disjunto e independiente del grupo de entrenamiento. Esto significa, por ejemplo, que un programa utilizado para entrenamiento no aparecerá en desarrollo. Estos datos de desarrollo provienen también de programas de televisión y tienen las mismas características de grabación.

3. SISTEMA PRIMARIO

Este sistema se basa en el empleo de GMMs entrenados con SDCs.

3.1. Preproceso

Para la identificación del idioma utilizando GMMs se pueden utilizar los Shifted Delta Cepstrum (SDC) [2], una extensión de las derivadas de los cepstrales. Estos nuevos vectores de características se calculan sobre una ventana de cepstrales, buscando extraer información fonética de más largo alcance. Los cepstrales de partida fueron extraídos utilizando el parametrizador de audio de iatros [3], que fue ampliado con un nuevo módulo para generar los SDCs. Los SDCs se definen a partir de cuatro parámetros, N-d-P-k, donde N indica el número de cepstrales originales, d se refiere al número de cepstrales utilizados para el cálculo de las derivadas y P hace referencia a la distancia entre los sucesivos cálculos de derivadas. El último

Este trabajo ha sido subvencionado por VIDI-UPV bajo las becas FPI del programa PAID06 y por el EC (FEDER), por el proyecto subvencionado por el Ministerio de Educación y Ciencia Español TIN2006-15694-C02-01 y por el programa Español Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

valor, k, especifica el número de derivadas calculadas a partir de un vector de cepstrales.

3.2. Descripción del sistema

Como sistema primario presentamos un modelo de mixturas de gaussianas (GMM) para llevar a cabo el reconocimiento. Dicho sistema ha sido entrenado con el *toolkit* HTK [4] y los SDCs obtenidos en el preproceso. Para obtener los SDCs se llevó a cabo un estudio para determinar los parámetros N-d-P-k idóneos para dicha experimentación. Finalmente, los SDCs se obtuvieron con valores de 7-1-3-7, resultando vectores de características de 49 dimensiones calculadas cada 210ms. Se entrenaron desde 2 gaussianas hasta 8192, determinando que 4096 gaussianas es el número de gaussianas con las que se obtiene una mejor tasa de acierto para el conjunto de desarrollo.

Para llevar a cabo el reconocimiento se ha implementado un módulo dentro del reconocedor propio iatros [3]. Dicho módulo está concretamente diseñado para llevar a cabo reconocimiento con GMMs, basándose en sumar las probabilidades de emisión por el GMM de todos los vectores de características de una señal dada.

4. SISTEMA ALTERNATIVO

Este sistema se basa en clasificar las muestras de test con la técnica de k-vecinos sobre los vectores de características en nuestro caso, estos vectores son cepstrales.

4.1. Preproceso

Los vectores de cepstrales son la característica contínua utilizada para el reconocimiento del habla [5]. Se obtienen de la señal, aplicando una ventana que se desliza a lo largo de la misma a partir de la cual se calcula el vector de cepstrales. Los cepstrales eliminan de la señal los rasgos propios del locutor y resaltan la información fonética. Para su extracción se definen diversos parámetros, como el tamaño de la ventana (normalmente unos 0.025 segundos), la frecuencia de submuestreo (cada cuánto extraemos una ventana), tamaño de la transformada de Fourier, el factor de pre-énfasis y el número de elementos del vector (típicamente 11). Adicionalmente pueden calcularse también la primera y segunda derivada en el tiempo de estos vectores, obteniendo así los vectores de 33 elementos utilizados.

4.2. Descripción del sistema

Este sistema utiliza en primer lugar los datos de entrenamiento y el algoritmo c-medias [6] para aprender un *codebook* de C *codewords*. Con dicho *codebook* aprendido se realiza la cuantificación vectorial de los ficheros de audio.

Los ficheros de audio pasan a ser representados mediante un único vector. Dicho vector es el histograma de la frecuencia de aparición de cada uno de los *codewords* en dicho fichero de audio. Por lo tanto, el tamaño de dicho vector es C. Este proceso se aplica a todos los ficheros de audio de entrenamiento y test.

Con los vectores obtenidos se puede utilizar cualquier técnica clasica de clasificación, por ejemplo, la técnicas de k-vecinos [7]. El resultado se puede mejorar con un aprendizaje discriminativo.

Para aplicar este aprendizaje discriminativo se aprende una base de proyección de C a d dimensiones [8]. Dicha proyección se obtiene sólo con las muestras de entrenamiento, pero se aplica tanto a vectores de entrenamiento como de test.

Por las pruebas realizadas se puede afirmar que en el espacio reducido de d dimensiones la clasificación mejora. Por tanto, en este punto podemos llevar a cabo la clasificación de las muestras de test con la técnica de k-vecinos.

5. BIBLIOGRAFÍA

[1] "Kalaka," Speech database created for the 2008 Language Recognition Evaluation on Spanish Languages, organized by the Spanish Network on Speech Technology. Produced by the Software Technologies Working Group (GTTS, http://gtts.ehu.es), University of the Basque Country.

[2] B. Bielefeld, "Language identification using shifted delta cepstrum," in *Proc. Fourteenth Annual Speech Research Symposium*, 1994.

[3] Míriam Luján, Vicent Tamarit, Vicent Alabau, Carlos-D. Martínez-Hinarejos, Moisés Pastor, Alberto Sanchís, y Alejandro Toselli, "iatros: A speech and handwritting recognition system," *V Jornadas en Tecnología del Habla*, Noviembre Bilbao, 2008.

[4] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, y P. Woodland, *The HTK Book*, CUED, UK, v3.2 edition, July, 2004.

[5] Lawrence Rabiner y Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall Ptr, 1993.

[6] R. Duda y P Hart, *Pattern Classification and Scene Analysis*, Wiley, 1973.

[7] B. S. Kim y S. B. Park, "A fast k nearest neighbor finding algorithm based on the ordered partition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 761–766, 1986.

[8] Mauricio Villegas, Roberto Paredes, Alfons Juan, y Enrique Vidal, "Face verification on color images using local features," in *Proceedings of the IEEE Computer Society Workshop on Biometrics, in association with CVPR 2008*, Anchorage, AK, USA, June 2008, IEEE Computer Society.