SEGMENTACIÓN DE FONEMAS NO SUPERVISADA BASADA EN MÉTODOS KERNEL DE MÁXIMO MARGEN

Yago Pereiro Estevan[1*], *Vincent Wan* [2]

Universidad Carlos III de Madrid[1]
Departamento de Teoría de la
Señal y Comunicaciones
Avda. de la Universidad, 30,
28911-Leganés (Madrid), ESPAÑA

Odette Scharenborg[2], *Ascensión Gallardo Antolín* [1]

University of Sheffield [2]
Department of Computer Science
211 Portobello Street,
Sheffield S1 4DP, UK.

RESUMEN

En este artículo se desarrolla un método automático de segmentación de fonemas no supervisado. Este método utiliza el algoritmo de agrupación de máximo margen [1] para realizar segmentación de fonemas sobre habla continua sin necesidad de información a priori para el entrenamiento del sistema.

1. INTRODUCCIÓN

En los últimos años ha quedado patente que la utilización de las máquinas de vectores soporte (*Support Vector Machines*, SVMs) [2, 3] en el campo del habla es una sólida línea de futuro. Existen campos, como la verificación de locutor [4], en los que las SVMs obtienen resultados muy competitivos, como se muestra en las evaluaciones del *NIST* [5]. Sin embargo, en otros campos como el reconocimiento automático del habla, su utilización todavía no es muy común.

Un algoritmo relativamente nuevo y muy interesante en este sentido es el algoritmo de agrupamiento de máximo margen (*Maximum Margin Clustering*, MMC) [1], que representa una aproximación no supervisada de las SVMs, consiguiendo localizar fronteras de máximo margen cuando el etiquetado no está disponible. Por lo tanto, el análisis del comportamiento de este algoritmo sobre la señal de voz es una tarea interesante.

En este artículo analizaremos la segmentación a nivel fonético de la voz. Sin embargo, el objetivo último no es la segmentación fonética en sí, si no el agrupamiento de la señal de voz en unidades que posteriormente puedan ser clasificadas por SVMs.

Desarrollaremos un análisis que nos permita realizar la segmentación fonética sin necesidad de un conocimiento a priori de la secuencia de fonemas contenida en la señal de voz. No se utiliza, como en otros métodos [6], la transcripción de la frase para determinar las fronteras entre fonemas. Tampoco es necesaria una secuencia de entrenamiento anterior a la aplicación del algoritmo.

El algoritmo se divide en tres etapas: en primer lugar, se realiza un preprocesado de la señal de voz, posteriormente, sobre el conjunto de vectores de parámetros que caracterizan la señal se aplica un algoritmo de agrupación por segmentos temporales basado en [1] y finalmente se realiza un postprocesado sobre dicho algoritmo con el fin de obtener la segmentación definitiva.

A continuación en la secciones 2 y 3 explicaremos brevemente la etapa de preprocesado de la señal y el algoritmo de agrupamiento utilizados respectivamente. En la sección 4 desarrollaremos el método de análisis de la señal de voz preprocesada y finalmente en la sección 5 explicaremos el postprocesado de dicho análisis con el fin de obtener la segmentación. En las secciones 6 y 7 explicaremos los experimentos realizados, las conclusiones obtenidas y las líneas futuras de trabajo.

2. PREPROCESADO DE LA SEÑAL DE VOZ

A partir de la señal de voz inicial, se extraen una serie de vectores de parámetros que contendrán las principales características de la voz. Entre los diferentes tipos de parametrización existentes [7], para nuestro algoritmo hemos escogido la parametrización MFCC (*Mel Filtering Cepstral Coefficients*) que extrae una serie de coeficientes derivados del espectro de la señal que son pasados por un banco de filtros en escala *mel*, que aproxima la percepción auditiva humana. En la figura 1 se puede ver la etapa de parametrización de la señal de voz del estándar ETSI ES 202 050 [8].

Es crítico encontrar un compromiso entre el desplazamiento entre vectores de parámetros y el tamaño de la

*Este trabajo ha sido realizado durante el periodo de investigación de Yago Pereiro Estevan en el departamento de ciencias de la computación de la universidad de Sheffield.

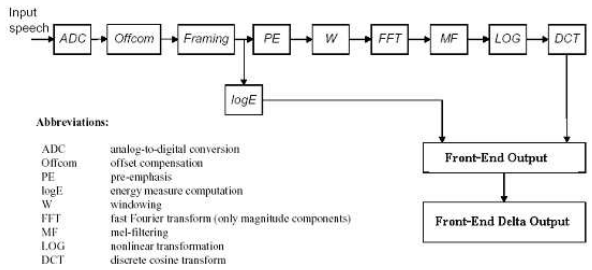

Figura 1. *Diagrama de bloques del sistema de parametrización para reconocimiento perteneciente al estándar ETSI ES* 202 050.

ventana de análisis de cada uno. Tomaremos segmentos de $15ms.$, en los cuales se puede considerar la señal casi estacionaria y, aplicando un solapamiento de $10ms.$ entre tramas, extraeremos vectores que caractericen cada segmento de señal de voz.

Inicialmente trabajaremos con vectores de 39 parámetros: 12 MFCCs más la log-energía y las correspondientes primera y segunda derivadas.

De este modo, la señal de voz quedará preprocesada obteniendo un conjunto de tramas de la forma

$$\mathbf{x}_i[n]_{i=1}^{k},$$

donde $n = 1 \dots N$ es el número de tramas y $i = 1 \dots k$ son los k coeficientes MFCCs de cada trama.

3. ALGORITMO DE AGRUPAMIENTO DE MÁXIMO MARGEN

El método de agrupamiento de máximo margen (MMC) [1] parte del principio de aprendizaje máquina de clasificación de máximo margen y lo modifica de modo que se pueda hacer una aproximación del mismo sin necesidad de que haya una etapa de aprendizaje supervisado o semi-supervisado. Esta modificación combinada con la utilización de métodos kernel consigue resultados a menudo mas eficientes que los métodos convencionales de agrupamiento espectral.

El objetivo de las máquinas de vector soporte (SVMs) es, dado un conjunto de entrenamiento etiquetado (x^1, y^1) ,. . ., (x^N, y^N) donde cada ejemplo es asignado a una clase, encontrar el discriminante lineal $f_{w,b}(x) = w^T\phi(x) + b$ que maximice el margen mínimo de no clasificación

$$\gamma^* = \max_{\omega,b,\gamma} \gamma \text{ dado } y^i(\omega^T\phi(x^i) + b) \geq \gamma, \forall_{i=1}^{N}, \|\omega\| = 1 \quad (1)$$

donde la constante de normalización euclídea sobre ω asegura que la distancia entre los datos y el hiperplano de separación (en el espacio de $\phi(\mathbf{x})$ determinado por ω^*,b^*.

El método MMC relaja las restricciones de máximo margen obteniendo una aproximación de máximo margen "blanda"de modo que a partir de un conjunto $x^1, \dots, x^N$ se consiga un etiquetado que si, posteriormente se utilizara en una SVM se consiguiese el de máximo margen entre todos los posibles.

4. ALGORITMO DE ANÁLISIS DE LA SEÑAL PARAMETRIZADA

Una vez parametrizada la señal de voz aplicaremos el algoritmo de agrupamiento sobre una ventana de análisis que iremos desplazando sobre el conjunto de vectores de parámetros, de modo que los vectores de parámetros de cada ventana queden agrupados en dos clases $y_j \in \{C_1, C_2\}$, $j = 1, \dots, m$ donde m es el número de vectores de parámetros analizado en cada ventana.

Posteriormente analizaremos la información obtenida tras dichos agrupamientos con el fin de realizar la detección de las fronteras entre fonemas. En este paso es crítica, tanto la selección del número de vectores de parámetros analizados en cada ventana como la correcta configuración de los parámetros modificables en el algoritmo de agrupamiento utilizado en la sección 3 ya que de esta selección depende el nivel en el que se analiza la señal de voz.

Tras el deslizamiento de la ventana de análisis sobre la señal parametrizada se obtendrá una matriz, en cuyas columnas se encuentra el etiquetado de cada vector de parámetros. Es decir, para cada ventana formada por los vectores de parámetros

$$\mathbf{x}_i[n]_{i=1}^{k} \quad (2)$$

con $n = N, \dots, N + m$ donde m es el tamaño de la ventana, se obtendrá un vector

$$\mathbf{y}_{j=1}^{m} \quad (3)$$

en el que cada elemento pertenecerá a una clase $y_j \in \{C_1, C_2\}$.

En la figura 2 se muestra una representación del agrupamiento realizado sobre un segmento de señal de voz parametrizada. En este caso, se ha utilizado una ventana de análisis de 19 vectores de parámetros. En cada columna de la figura se puede observar la clasificación obtenida sobre los vectores de parámetros de cada ventana analizada.

Podemos observar a partir de la figura 2 que se detectan estructuras que nos permiten distinguir donde hay un cambio significativo en la señal de voz. Estas estructuras

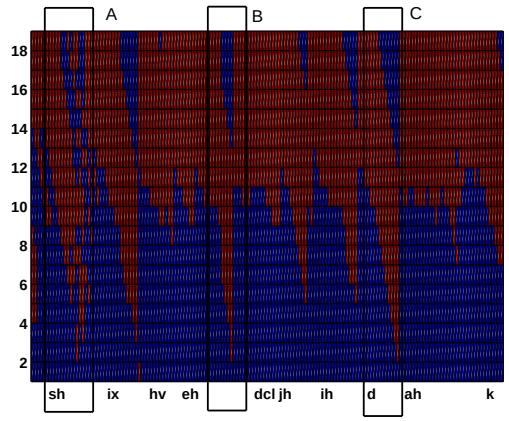

Figura 2. *Representación del agrupamiento por ventanas, en cada ventana vertical las diferencias de color indican la clasificación realizada. En el eje x se representa el comienzo de cada fonema con su nombre.*

son las que analizaremos para realizar la segmentación fonética.

5. SEGMENTACIÓN

En teoría, cuando la ventana quede centrada en una frontera el método de agrupamiento hará que la mitad anterior a la frontera quede agrupada en una clase y la mitad posterior en otra. Según nos separamos de la frontera el número de elementos pertenecientes al fonema sobre el que nos encontramos va creciendo.

Existen fronteras muy claramente definidas mientras que otras se pueden intuir pero no son tan claras. Esto puede ser debido, por ejemplo, a que el fonema sea demasiado breve o al hecho de que hay fonemas cuya separabilidad es mas compleja [9]. En la figura 2 se puede observar que hay estructuras como *C* que están claramente definidas de principio a fin. Sin embargo, hay otras estructuras como *A* que, a pesar de poder apreciarse que existe una frontera, ésta es mucho más ruidosa. En el caso concreto de *A* puede deberse a que al ser un fonema fricativo, la señal es mucho mas ruidosa. Finalmente aparecen estructuras como *B* donde no existe una frontera, lo que nos hace pensar que el algoritmo está detectando más cambios que los meramente fonéticos. Este último tipo de estructura es muy interesante ya que nos sugiere que el algoritmo MMC puede realizar un análisis a un nivel sub-fonético [10] y esto puede ser de gran utilidad para trabajos futuros.

A partir de las clasificaciones realizadas para cada ventana debemos analizar la información que nos permita encontrar la mejor segmentación posible. Como hemos visto, la información obtenida es muy compleja y muy rica, por lo tanto la obtención de un método de automatización no es un problema sencillo de resolver. A continuación presentaremos dos métodos desarrollados para realiar una aproximación a la automatización de la segmentación.

5.1. Método basado en estructuras

Una primera aproximación que se realiza es el análisis a partir de, únicamente, la matriz de vectores clasificados de modo que, a partir de las estructuras observadas y explicadas anteriormente podamos discernir donde se encuentra la frontera y realizar la correcta segmentación.

Para ello, en primer lugar se realiza una selección de aquellas ventanas en las que a cada lado de vector de parámetros central se agrupen los elementos de cada clase. Una vez seleccionadas esas ventanas se realiza un análisis basado en las ventanas siguientes con el fin de observar si aparece una estructura de deslizamiento en las clases.

Este es un problema de comparación con una máscara, es decir, comparar un segmento de ventanas siguientes con una matriz (máscara) que tenga en cada elemento las clases que esperamos. Dependiendo del tamaño de la máscara con la que comparemos, obtendremos resultados diferentes. Cuanto mayor es la máscara, menos patrones coincidirán, pero la probabilidad de que esos patrones no pertenezcan a una frontera será mucho menor. Por lo tanto hay que buscar un compromiso, y para ello se han realizado una serie de simulaciones.

Este método, como veremos en la sección de experimentos, presenta una gran fiabilidad, obteniéndose una tasa de falsa alarma muy baja, sin embargo y debido a que, como ya hemos explicado antes, hay estructuras que no aparecen de un modo claro, la probabilidad de pérdida de fronteras es demasiado elevada.

5.2. Cálculo de las distancias euclídeas

Debido a la riqueza y complejidad de las estructuras, la automatización mediante máscaras es muy compleja por lo tanto decidimos utilizar como ayuda el cálculo de la distancia euclídea entre los puntos medios de cada una de las clases, es decir, para cada trama i,

$$D_i = \sqrt{\sum_{i=1}^{k} (\bar{x}_i^{C_1} - \bar{x}_i^{C_2})^2}, \quad (4)$$

donde

$$\bar{\mathbf{x}}^{C_i} = \frac{1}{N_{C_i}} \sum_{j}^{N_{C_i}} \mathbf{x}_j, \quad (5)$$

siendo C_1, C_2 son las dos clases posibles y N_{C_i} es el número de elementos de la clase C_i.

Los resultados obtenidos, como se puede ver en la figura 3, nos permiten ver que la distancia es mayor en las fronteras, de modo que si aplicamos un sencillo algoritmo de detección de máximos locales obtenemos un método de segmentación automática.

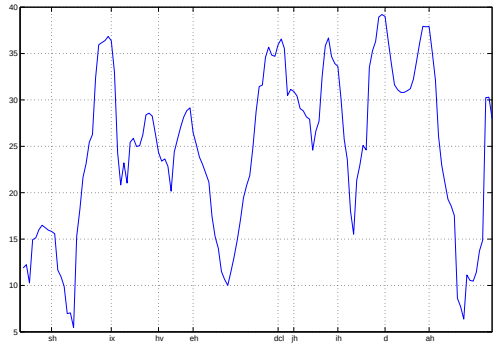

Figura 3. *Representación de las distancias euclídeas entre clases. En el eje x se representa el comienzo de cada fonema con su nombre.*

En la sección de experimentos, veremos que la cantidad de fronteras perdidas es baja, sin embargo al ser la señal de distancias, ruidosa, se introduce una mayor tasa de falsas alarmas.

5.3. Combinación

Para resolver los problemas de los métodos anteriores hemos optado por combinar ambos métodos. En la actualidad estamos desarrollando un método que nos permita combinar ambos métodos de modo que minimicemos tanto la probabilidad de falsa alarma como la de pérdida de fronteras.

6. EXPERIMENTOS

Para este estudio se utilizó un subconjunto de la base de datos Americano-Inglesa DARPA-TIMIT utilizando 4 frases pertenecientes a 4 locutores diferentes, dos masculinos y dos femeninos. Dicha base de datos está muestreada a $16\ KHz$.

El primer paso de la experimentación consistió en optimizar la selección de los diferentes parámetros utilizados tanto en la parte de preprocesado como en el algoritmo de análisis de la señal parametrizada (secciones 2 y 4). En concreto, en lo que se refiere a la parte de preprocesado de la señal de voz, se realizaron pruebas sobre el tamaño de las ventanas de parametrización, así como el solapamiento entre tramas como ya explicamos en la sección 2.

Asimismo inicialmente se realizaron simulaciones modificando el kernel utilizado en el algoritmo MMC, con el fin de optimizar la respuesta del sistema en relación al problema de segmentación de fonemas. Tras estas pruebas decidimos utilizar un kernel *RBF* con $\sigma = 200$. Finalmente realizamos pruebas sobre el tamaño de la ventana de análisis utilizada, decidiendo finalmente utilizar 9 vectores de parámetros a cada lado del vector central, es decir ventanas de 19 vectores de parámetros en total.

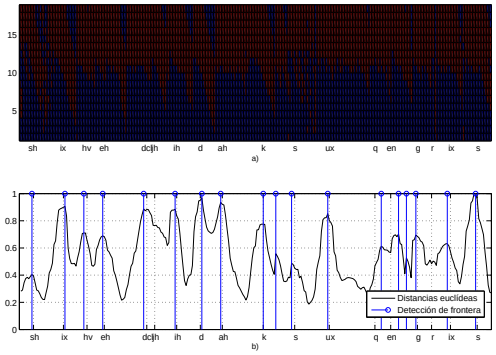

Figura 4. *a) Representación del agrupamiento por ventanas en un segmento de voz. b) Representación de las distancias euclideas entre clases para el mísmo segmento de voz que a) y segmentos detectados aplicando un método de detección de máximos locales.*

En la figura 4 se puede observar a simple vista que la combinación de ámbos métodos permite una segmentación los fonemas bastante precisa, en la parte *a)* se representa la clasificación de las diferentes ventanas, mientras que en la parte *b)* se representan las distancias euclídeas entre cada clase, para cada ventana de análisis, así como los máximos locales detectados.

En la tabla 1 se observan los resultados sobre el método de distancias euclídeas, con una distancia del etiquetado realizado por TIMIT de $\pm 20ms$. En cada columna se obtienen los resultados obtenidos para varias sensibilidades diferentes en la detección de máximos locales. Se puede observar que alcanzanzamos una detección de fronteras razonablemente alta, sin embargo la tasa de falsa alarma aumenta al aumentar la sensibilidad en la detección de máximos debido a lo ruidosa que es la señal.

En la tabla 2 se observan los resultados sobre el método de análisis de estructuras, con una distancia del etiquetado realiazdo por TIMIT de $\pm 20ms$. En cada columna se obtienen los resultados obtenidos con diferentes tamaños de la matríz máscara. Podemos ver que para máscaras muy grandes llegamos a eliminar la falsa alarma, pero debido a que aparecen muy pocas estructuras tan claras, el

	Conf. 1	**Conf.** 2	**Conf. 3**
Front. detect. (%)	84,68	78,99	71, 87
Falsas front. (%)	52,66	35,31	19,52

Tabla 1.
Valores de detección de frontera utilizando el método de las distancias euclídeas entre vectores de parámetros de diferentes clases. Cada columna muestra los resultados para una configuración en la sensibilidad de la detección de máximos locales.

	5×5	4×4	3×3
Front. detect. (%)	8,95	15,65	25,21
Falsas front. (%)	0	5,2	13,27

Tabla 2.
Valores de detección de frontera utilizando el método de detección de estructuras. Cada columna muestra los resultados para una configuración diferente de la matriz máscara.

	$Comb$
Front. detect. (%)	87,1
Falsas front. (%)	32,61

Tabla 3.
Valores de detección de frontera utilizando el método una combinación del método de distancias euclídeas y del método de detección de estructuras.

porcentaje de detección es muy bajo. Según disminuye el tamaño de la máscara, mejora la detección, pero también aumenta la falsa alarma.

Finalmente en la tabla 3 mostramos los resultados de una posible combinación de ambos métodos. Consideramos que hay una frontera donde cualquiera de los métodos haya detectado una frontera y eliminamos aquellas fronteras que estén juntas (en tramas consecutivas). Este es un método muy sencillo de combinación, pero se observa que la probabilidad de pérdida de tramas desciende y aunque se perciba un aumento de la falsa alarma, este aumento se debe a que el método introduce puede introducir en muchos casos como dos fronteras diferenciadas la misma frontera detectada en diferentes lugares (no consecutivos). Esta combinación nos permite intuir que la combinación de ambos métodos de un modo mas complejo puede mejorar los resultados ya que se pueden detectar mas fronteras (como ya veiamos en la figura 4) y la falsa alarma puede disminuirse.

7. CONCLUSIONES

Estamos comenzando una línea de investigación interesante en la que tratamos de estudiar la aplicabilidad del algoritmo de agrupamiento de máximo margen a tareas relacionadas con el procesamiento de la señal de voz. En concreto, en este artículo nos hemos centrado en la segmentación de fonemas pero como ya comentamos en la introducción, esta línea de trabajo abre un amplio abanico de posibilidades.

Los resultados obtenidos en la tarea de segmentación fonética de habla continua nos permiten ser muy optimistas. Se puede observar que tenemos la información necesaria para recuperar las fronteras entre fonemas, pero hay que mejorar la etapa de postprocesado para la extracción automática de dicha información. Podemos decir que, aunque estamos en una fase inicial, los resultados son relativamente buenos dada la sencillez de los algoritmos presentados. En la actualidad, estamos trabajando en la mejora del algoritmo de extracción de las marcas entre fonemas basado en la combinación entre el método de análisis de estructuras y el basado en las distancias euclídeas.

Respecto a otras líneas de trabajo futuro, y puesto que en la experimentación realizada hasta ahora hemos observado que el algoritmo parece capaz de realizar un análisis de voz a nivel subfonético, tenemos planeado su aplicación al análisis de características articulatorias (AF) [11]. Este es un campo menos explorado, pero muy prometedor en el ámbito de reconocimiento automático del habla, en el que la técnica AF aparece como una alternativa más flexible para la modelización de la variación de la voz [12]. Además, en este contexto, todavía no se ha desarrollado un algoritmo de etiquetado automático.

8. BIBLIOGRAFÍA

[1] L. Xu, J.Ñeufeld, B. Larson, y D. Schuurmans, "Maximum margin clustering," *in Advances in Neural Information Processing Systems*, vol. 17, pp. 1537–1544, 2005.

[2] B. Schölkopf y A. Smola, *Learning with kernels*, MIT Press, Cambridge, MA, 2002.

[3] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[4] K. Daoudi y J. Louradour, "Conceiving a new sequence kernel and applying it to svm speaker verification," *Proc. Interspeech*, pp. 3101–3104, 2005.

[5] National Institute of Standards y Technology, "The nist year 2006 speaker recognition evaluation plan," 2006.

[6] Bryan L. Pellom y John H. L. Hansen, "Automatic segmentation of speech recorded in unknown noisy channel characteristics," *Speech Communication*, vol. 25, pp. 97–116, 1998.

[7] Christophe Lévy, Georges Linarès, Pascal Nocera, y Jean-François Bonastre., "Reducing computational and memory cost for cellular phone embedded speech recognition system," *IEEE International Conference on Acoustic, Speech and Signal Processieng*, 2004.

[8] ETSI Standard:ETSI ES 202 050., "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms dsr advanced front end," 2002.

[9] Maria-Barbara Wesenick y Andreas Kipp, "Estimating the quality of phonetic transcriptions and segmentations of speech signals," *Internacional Conference on Spoken Language Processing (ICSLP)*, 1996.

[10] O. Scharenborg, V. Wan, y R. Moore, "Capturing fine-phonetic variation in speech through automatic classification of articulatory features," *ITRW Workshop on Speech Recognition and Intrinsic Variation*, 2006.

[11] P. Taylor y S. King, "Detection of phonological features in continuous speech usin neural networks," *Computer Speech and language*, vol. 14, pp. 333–353, 2000.

[12] M.Ostendorf, "Moving beyond the 'beads-on-a-string model of speech," *Proceedings of the IEEE Automatic Speech Recognition and Understandig Workshop, Keystone, CO*, pp. 79–84, 1999.