ALISADO DE LOS MODELOS DE MARKOV MEDIANTE REDES FONOTROPICAS. APLICACION AL RECONOCIMIENTO DEL HABLA.

E.Monte
J.B.Mariño
E.LLeida.

Dpt. de Teoría del Senyal i Comunicacions. UPC
Email: enric@tsc.upc.es

I-RESUMEN.

En esta comunicación presentamos un nuevo método para alisar modelos ocultos de Markov en los que la cuantificación vectorial se ha llevado a cabo mediante las redes de Kohonen. Este método está basado en aprovechar la propiedad topológica de las redes fonotrópicas, que consiste en una proyección no ortogonal del espacio de características sobre el mapa fonotrópico. Esta proyección no ortogonal preserva la propiedad de colindancia, es decir puntos colindantes en el espacio de las características son colindantes sobre el mapa fonotrópico, que es de dimensión dos. La propiedad de colindancia hace que ordenando el vector de probabilidad de emisión de cada estado siguiendo el orden del mapa fonotrópico, obtengamos una superficie en la que no tenemos transiciones bruscas de un punto del mapa a otro colindante. Un filtrado paso bajo de dimensión dos de esta matriz realiza el proceso de alisado. En la comunicación comparamos este algoritmo de alisado con otros algoritmos de alisado que han sido propuestos en la literatura y presentamos además las propiedades de este método que lo hacen atractivo como alternativa a los métodos de coocurrencias y de Parzen.

II-INTRODUCCION

En esta comunicación presentamos un nuevo método de alisado basado en el uso de redes fonotrópicas en el contexto de los modelos ocultos de Markov. Este método tal como veremos tiene una serie de ventajas en cuanto a simplicidad y tal vez un atractivo en cuanto a lo intuitivo que resulta, pues la matriz de probabilidad de emisión de los símbolos se puede "ver" y el alisado consiste en un filtrado paso bajo de esta superficie.

Recientemente ha habido varios grupos que han usado las redes fonotrópicas (también conocidas como Self Organizing Maps: SOM) en el contexto de los modelos ocultos de Markov /1/, /2/, /3/, /4/. Los sistemas de reconocimiento que se presentan en estas publicaciones suelen usar como bloque de cuantificación una red fonotrópica y siguen un esquema como el que presentamos en la figura 1. La razón por la que se usa en estos sistemas un cuantificador de este tipo es para aprovechar las propiedades discriminativas de la cuantificación mediante las redes fonotrópicas, lo que les permite reducir la tasa de errores respecto a otros métodos de cuantificación vectorial .

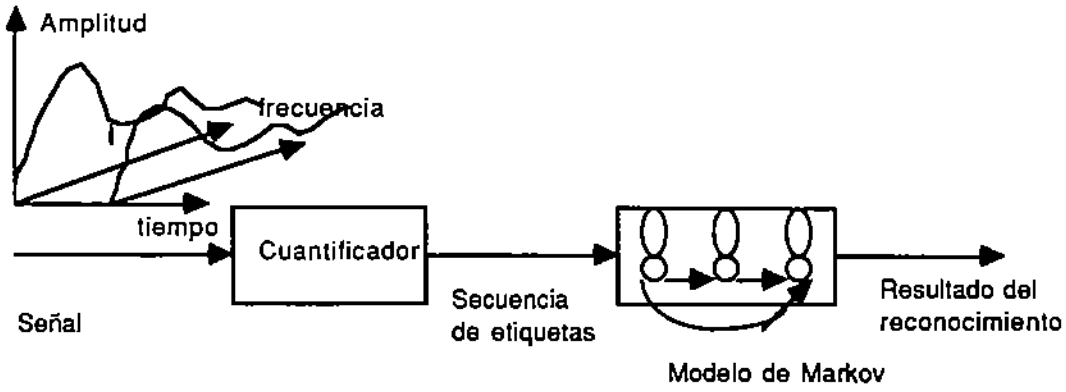


Figura 1. Diagrama de un sistema de reconocimiento basado en Modelos Ocultos de Markov y un cuantificador vectorial. En nuestro caso el cuantificador vectorial es una red fonotrópica.

Por otra parte un tema al que se ha dedicado atención en la literatura reciente es el de solucionar el problema del entrenamiento de los modelos ocultos de Markov cuando la base de datos de entrenamiento es insuficiente. Los modelos de Markov son clasificadores probabilísticos en los que se estima un modelo paramétrico de cada una de las clases que se quiere clasificar. Uno de los inconvenientes que presenta este sistema es el de la estimación correcta de los parámetros a partir de datos insuficientes, o la estimación cuando los datos no son lo suficientemente representativos de la distribución real de las clases. Una manera de paliar en parte este problema es el alisado de los modelos de Markov, que consiste en hacer que las probabilidades de emisión de los símbolos sean una combinación lineal de las probabilidades de símbolos asociados a espectros que sean semejantes según algún criterio. En la literatura se han propuesto diversos métodos que han dado resultados satisfactorios /5/, /6/, /7/. En particular en esta comunicación compararemos los métodos de Coocurrencias y el de Parzen con el método que presentamos.

III-ALISADO MEDIANTE REDES FONOTROPICAS.

La introducción de las redes fonotrópicas en el contexto de los modelos ocultos de Markov trae como consecuencia la aparición de una propiedad muy interesante. Esta propiedad consiste en que las probabilidades de emisión (p.e.) de símbolos generan superficies sin transiciones bruscas entre puntos colindantes, si la matriz de p.e. se ordena siguiendo el orden de las redes fonotrópicas (Ver la figura 2). La razón por la que puntos colindantes sobre el mapa fonotrópico tienen probabilidades similares reside en el hecho de que representan puntos colindantes sobre el espacio de las características. Una manera intuitiva de ver por que se produce este fenómeno, la tenemos observando las trayectorias sobre el mapa fonotrópico que presentamos en la figura 3. Obsérvese que las transiciones de trama a trama sobre el plano fonotrópico se realizan entre puntos colindantes, excepto en las transiciones de un fonema a otro. Dado que habrá regiones asociadas a espectros muy semejantes, las probabilidades asociadas con los centroides de estas zonas serán semejantes porque si los datos de entrenamiento están representan bien la distribución de la clase, los centroides de cada región aparecerán con una frecuencia semejante.

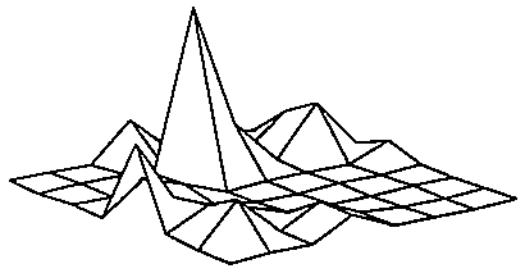


Figura 2. Representación de la matriz de probabilidad de emisión en un estado de un modelo oculto de Markov. La matriz ha sido ordenada siguiendo el orden de el mapa fonotrópico.

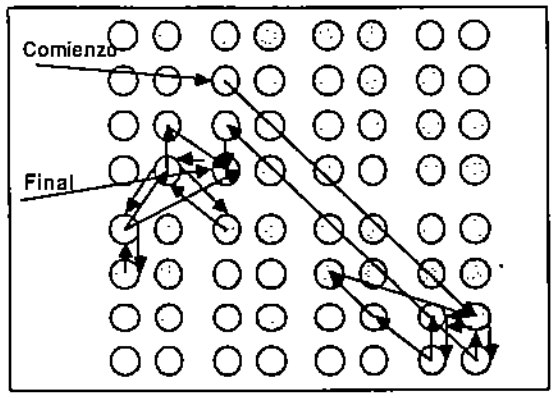
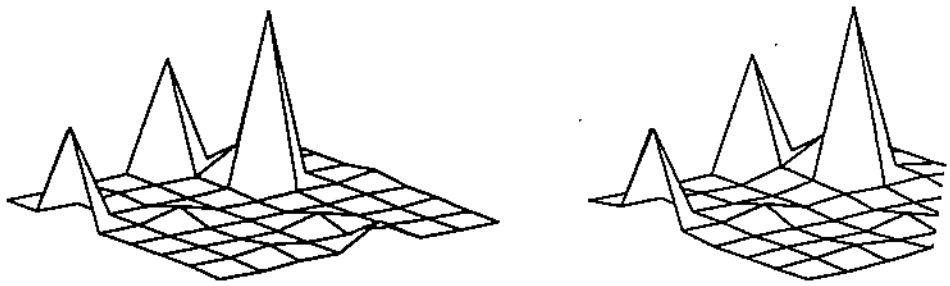
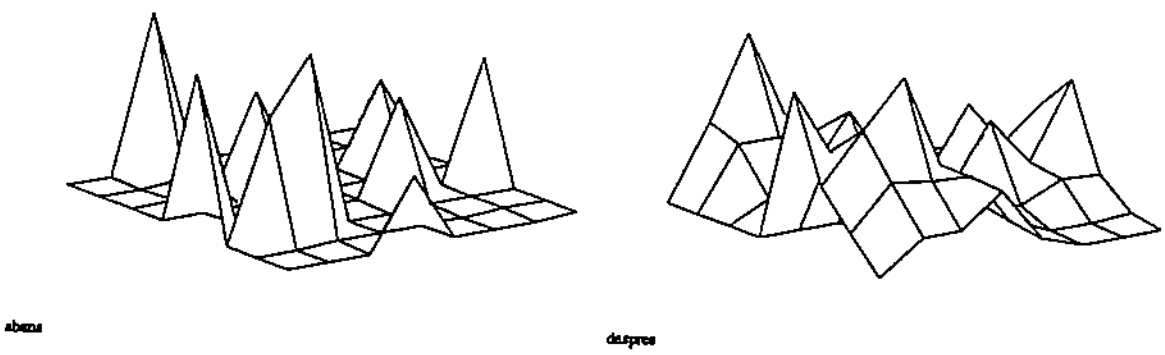


Figura 3. Trayectoria de las tramas de la palabra "tres" sobre el mapa fonotrópico.



(a) (b)



(c) (d)

Figura 4 Matrices de probabilidad de emisión en función de las coordenadas sobre el mapa fonotrópico. (a) Antes del alisado en escala lineal (b) Después del alisado en escala lineal. (c) Antes del alisado en escala logarítmica (-log(.)) ("floor threshold" 1.0e-3) (d) Después del alisado en escala logarítmica (con "floor threshold").

El método de alisado que presentamos consiste en filtrar mediante un filtro bidimensional paso bajo de orden 3 la matriz de p.e. ordenada siguiendo el orden del mapa fonotrópico, de tal manera que si existe algún elemento de la matriz de probabilidad de emisión que está mal entrenado debido a falta de material de entrenamiento, se pueda compensar en parte mediante un promediado, que consiste en hacer que la probabilidad de esta observación sea una combinación lineal de las probabilidades de las observaciones colindantes. En la figura 4 presentamos un ejemplo del efecto del filtrado paso bajo sobre una matriz de probabilidad de emisión.

IV-RESULTADOS EXPERIMENTALES.

A continuación presentaremos una serie de experimentos en los que compararemos la tasa de reconocimiento del método que presentamos con otros métodos de alisado conocidos. La nomenclatura que seguiremos para presentar los resultados es la siguiente:

Sistema "MEDEA": modelos ocultos de Markov en los que la cuantificación vectorial se realiza mediante las redes fonotrópicas.

Sistema "SMOOTH": sistema "MEDEA" en el que los modelos han sido alisados mediante el método del filtrado paso bajo.

Sistema "ULISES": modelos ocultos de Markov en los que el cuantificador vectorial se obtiene mediante el algoritmo LBG.

Sistema "Coocurr.": sistema "ULISES" en el que los modelos han sido alisados mediante el método de las coocurrencias.

Sistema "Parzen": sistema "ULISES" en el que los modelos han sido alisados mediante el método de Parzen.

Los experimentos están pensados para comparar como se comportan los sistemas tras un entrenamiento insuficiente de los modelos. El algoritmo de cuantificación vectorial usado en los sistemas que no utilizan los SOM es el LBG /8/. El filtro paso bajo para realizar el alisado con el

sistema "SMOOTH" se ha elegido de manera empírica, siguiendo el criterio de maximizar la tasa de reconocimiento en unas condiciones experimentales específicas, que son:

Tamaño del codebook :64

Entrenamiento: 9 locutores de la base de datos.

Reconocimiento:1 locutor fuera de la base de entrenamiento.

Filtro paso bajo : kernel de 3*3.

Los experimentos se realizaron con una base de datos formada por los dígitos catalanes. La base de datos tiene las características siguientes:

Número de locutores: 10 (siete hombres y tres mujeres).

Corpus: 10 palabras (los dígitos en catalán)

Repeticiones de cada dígito: 10

Total de elementos de la base: 1000 elementos.

Los experimentos están pensados para ver como se comparan los métodos de alisado en diversas condiciones de entrenamiento:

Experimento 1: Cuatro locutores de entrenamiento: { 1 versión;2 versiones;5 versiones; 10 versiones } por locutor.

Experimento 2: tres locutores de entrenamiento: { 1 versión;5 versiones; 10 versiones } por locutor.

Experimento 3: dos locutores de entrenamiento: { 3 versiones;5 versiones; 10 versiones } por locutor.

Experimento 1:

El experimento consistió en variar el entrenamiento de los modelos y comparar los resultados de reconocimiento para un codebook de dimensión 64.

Los resultados están resumidos en la Tabla I, en la que se ha usado como parámetro el número de versiones de cada dígito usadas en el entrenamiento.

Tabla I

Método	1 versión	2 versiones	5 versiones	10 versiones
Ulises	80,9	91,2	92,67	92,67
Coocurr.	85,4	92,38	92,5	94
Parzen	86,7	94,76	88,67	90,83
Medea	86,6	90,2	88,67	92,17
Smooth	88,7	91,9	89,83	93,5

Para poder interpretar mejor los resultados los presentaremos en forma gráfica. En la figura 4 presentamos los resultados de reconocimiento en función de los algoritmos para un codebook de dimensión 64. Se observa que el método de alisado más consistente es el "SMOOTH", pues en todos los casos mejora la tasa de reconocimiento respecto al sistema "MEDEA", mientras que los métodos de "COOCURRENCIAS" y "PARZEN", se comportan de manera irregular, a veces introducen mejoras y otras reducen la tasa de reconocimiento. Esta consistencia del sistema "SMOOTH" se puede observar también en los experimentos siguientes.

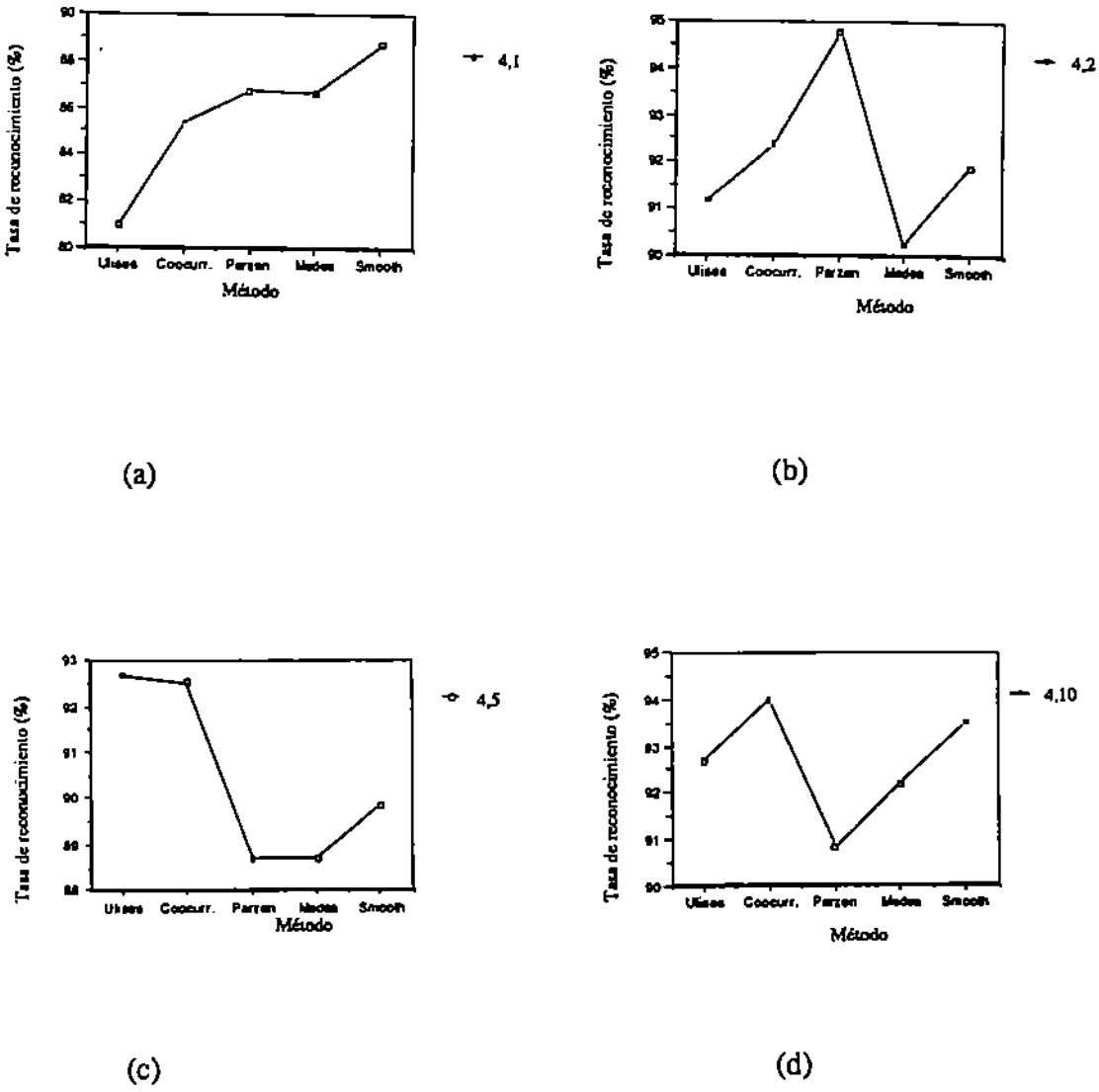


Figura 4. Comparación de las tasas de reconocimiento obtenidas con diversos algoritmos. Los resultados son para codebooks de tamaño 64. (a) Entrenamiento de los modelos con cuatro locutores con una versión por dígito. (b) Entrenamiento con cuatro locutores y 2 versiones (c) Entrenamiento con cuatro locutores y 5 versiones (d) Entrenamiento con 4 locutores y 10 versiones.

Experimento 2: El experimento es análogo al anterior, y el objetivo es ver como se comportan los algoritmos de alisado con un entrenamiento más pobre.

Locutores de entrenamiento: 0,1,2 (Dos hombres y una mujer).

Locutores de reconocimiento: 3,4,5,6,7,8,9 (Cinco hombres y dos mujeres) con diez versiones por dígito. Total de la base de test: 700 palabras.

En la tabla siguiente presentamos los resultados del experimento:

Tabla II

	1 versión	5 versiones	10 versiones
Ulises	87,1	90,14	92,43
Coocurr.	90,4	93,71	92,43
Parzen	88,75	85	84,7
Medea	88,9	89,7	91,57
Smooth	91,25	91,14	94,14

En este experimento se observan unos resultados semejantes al caso anterior, los métodos de alisado "Coocurrencias" y "Parzen" se muestran irregulares en cuanto a prestaciones mientras que el método "Smooth" se muestra consecuente para todos las pruebas, pues la tasa de reconocimiento siempre aumenta respecto al método de base "Medea".

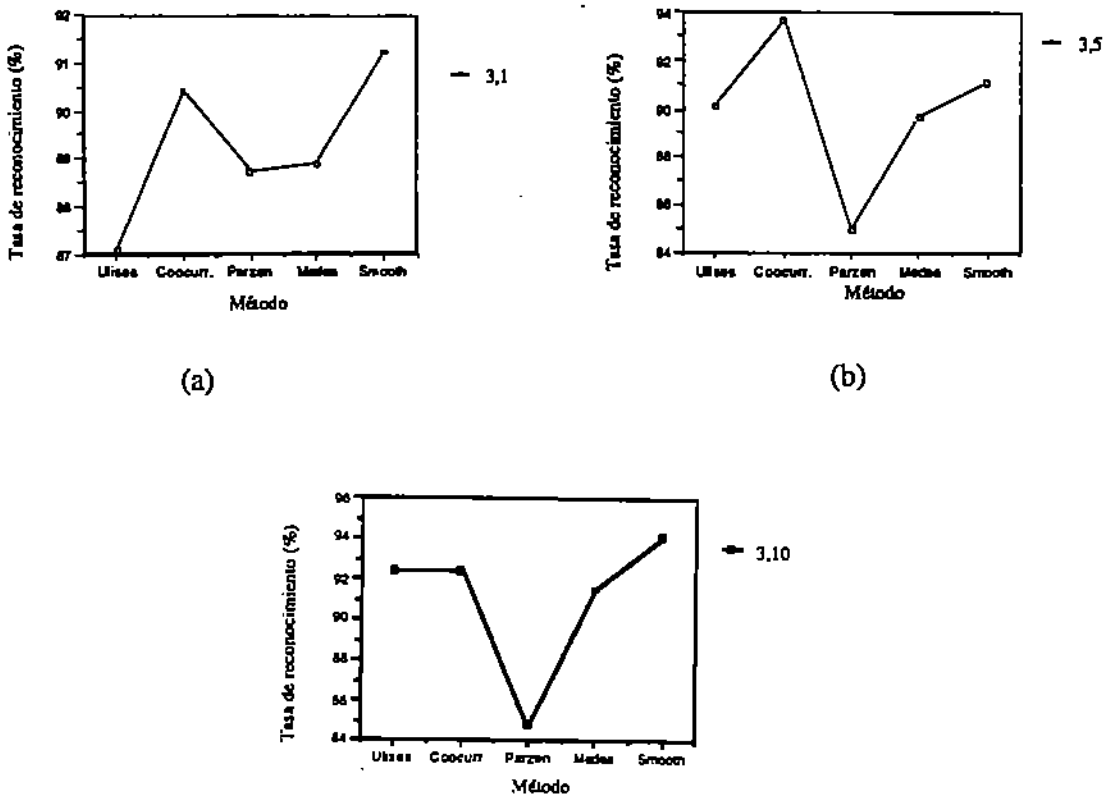


Figura 5. Comparación de los métodos de reconocimiento para los modelos entrenados con tres locutores, con varias versiones (a) 1 versión (b) 5 versiones (c) 10 versiones.

Experimento 3: El experimento siguiente lo realizamos para ver como se comporta el algoritmo al entrenar los modelos de manera insuficiente con dos locutores (uno de cada sexo).

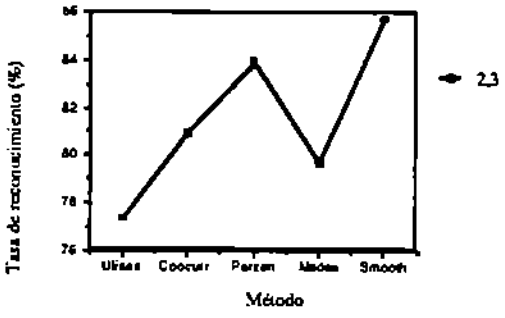
Locutores de entrenamiento: 4,5 (un hombre y una mujer).
Locutores de reconocimiento: 0,1,23,6,7,8,9 (seis hombres y dos mujeres) con diez versiones por dígito. Total de la base de test: 800 dígitos.

A continuación en la tabla IV presentamos los resultados del experimento:

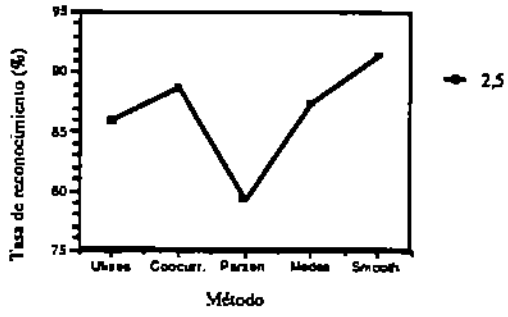
Tabla IV

	1 versión	5 versiones	10 versiones
Ulises	77,32	86	87,8
Coocurr.	80,9	88,8	92,38
Parzen	83,9	79,25	80
Medea	79,64	87,5	87,25
Smooth	85,7	91,5	92,00

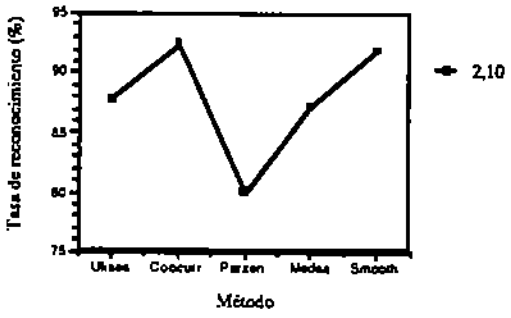
Un aspecto interesante que se desprende de este experimento y se puede confirmar viendo los otros experimentos, es que el sistema "Smooth" es el más robusto cuando en entrenamiento es muy pobre, pues en esos casos presenta la tasa de reconocimiento más elevada o muy similar (< 1% de diferencia) al método de alisado que presenta mejores resultados.



(a)



(b)



(c)

Figura 6. Comparación de los métodos de reconocimiento para los modelos entrenados con dos locutores, con varias versiones (a) 1 versión (b) 5 versiones (c) 10 versiones.

En este caso extremo se observa que la tasa de reconocimiento en general disminuye, sin embargo la tendencia que se ha observado en el caso anterior en cuanto a variabilidad de los métodos de alisado se mantiene.

Comentarios sobre los métodos de alisado que hemos estudiado

El sistema de alisado que presentamos se comporta en general de manera comparable a los sistemas "COOCURRENCIAS" y "PARZEN". Este sistema tiene un buen comportamiento en el sentido de que siempre proporciona resultados mejores que el sistema base, mientras que los métodos de "COOCURRENCIAS" y "PARZEN" tienen una gran variabilidad. En algunos experimentos alguno de los dos sistemas puede dar resultados muy buenos mientras que tras una variación en el tipo de entrenamiento se pueden hundir. Este tipo de comportamiento es muy independiente del entrenamiento de los modelos.

-El método "SMOOTH" que presentamos tiene una complejidad de cálculo inferior al método de coocurrencias, pues el alisado se realiza sobre cada modelo de manera independiente de todos los demás y consiste en una simple convolución. El método de las coocurrencias está basado en una transformación **T** cuyos elementos son probabilidades del tipo $p(k_i/k_j)$ cuyo cálculo necesita del conocimiento de todos los modelos (k_l representa la observación "l" del codebook usado).

-El método de "PARZEN" se asemeja al método que presentamos por el hecho de que usa información sobre distancias para calcular la matriz de transformación **T**. El método de Parzen tal como lo usamos tenía como parámetros: **a**=1 y **s**=1. que usan en /5/.

V-CONCLUSIONES

Hemos presentado en esta comunicación un nuevo método para alisar modelos ocultos de Markov, basada en la propiedad de colindancia que aparece debido al uso de las redes fonotrópicas. Hemos comparado este algoritmo con otros dos algoritmos para alisar modelos de Markov y hemos comprobado que en las condiciones en las que hemos trabajado el algoritmo se comporta de manera satisfactoria. El algoritmo que presentamos es consistente en el sentido que siempre aumenta la tasa de reconocimiento respecto al sistema de base, cosa que no sucede con los otros métodos de alisado que hemos estudiado. Otra ventaja que tiene este sistema es la menor complejidad de calculo respecto al método de las coocurrencias y el atractivo que presenta por el hecho de que se puede "ver" la superficie de probabilidad de emisión y como se ve afectada la superficie al realizar el alisado.

VI-REFERENCIAS

/1/ P.Ramesh, S.Shigeru y Chin-Hui Lee. "A new connected word recognition algorithm based on HMM/LVQ segmentation and LVQ classification". ICASSP-91

/2/ H.Iwamida, S.Katagiri, E.McDermott y Y.Tohkura. "A hybrid speech recognition systema using HMMs with an LVQ-trained codebook". ICASSP-90

/3/ E.Monte y J.B.Mariño. "Modelos de Markov y Cuantificación Vectorial por medio de redes de Kohonen".URSI-91. Cáceres.

/4/ G.Rigoll. Neural network based continuous speech recognition by combining self organizing feature maps and hidden markov modeling. EURASIP 90. SESIMBRA. PORTUGAL

/5/ R.Schwartz, Owen Komball, Francis Kubala, Ming-Whei Feng, Yen-Lu Chow, Chirs Barry, J. Makhoul. "Robust Smoothing Methods for Discrete Hidden Markov Models." ICASSP 89.

/6/ K.F.Lee, Hon. "Speaker Independent Phone Recognition" IEEE, Trans. on ASSP Nov.1989.

/7/ A. Nevot,E.Lleida,A.Bonafonte. "Técnicas de suavizado para la mejora del reconocimiento con modelos ocultos de Markov". IVSimposium nacional de reconocimiento de formas y análisis de imágenes. Granada 90.

/8/ Linde, Buzo, Gray."An Algorithm for Vector Quantizer Design". IEEE Trans. Commun., COM-28,1,pp84-95.1980.