RESULTADOS PRELIMINARES SOBRE MODELOS MVQHMM

J.C. Segura, A.J. Rubio, J.E. Díaz, P. García M.C. Benítez

Dept. Electrónica y Tecnología de Computadores
Univ. Granada. 18071 Granada (Spain)
Tlf. 34-58-243283
Fax. 34-58-243230

Resumen

En este artículo, se presenta una nueva variante del modelado oculto de Markov discreto. Su principal característica es la utilización de un diccionario VQ específico para cada uno de los modelos HMM. De esta forma, un MVQHMM consta de un diccionario VQ y un modelo HMM discreto. El procedimiento de evaluación de la probabilidad total de generación de una secuencia de vectores incógnita por parte de un modelo MVQHMM, combina la distorsión de cuantización de dicha secuencia con el diccionario VQ del modelo y la probabilidad de generación de la secuencia de símbolos, obtenida en el proveso VQ, por parte del modelo HMM asociado. En estre tabajo, se presentan resultados comparativos frente a dos sistemas de reconocimiento, uno basado en modelos HMM discretos, otro basado en modelos HMM semicontiunos. Los resultados experimentales muestran que la utilización de diccionarios VQ específicos, y la inclusión de las distorsiones de cuantización en el criterio de clasificación, mejoran el rendimiento del sistema de reconocimiento.

1 Introducción

Los modelos ocultos de Markov discretos (DHMM) han sido ampliamente utilizados en sistemas de reconociminto de voz, como modelos acústicos de diferentes tipos de unidades de decisión (palabras, fonemas, etc.), obteniendose buenos resultados. Sus principales ventajas frente a otros tipos de aproximaciones son su moderado coste computacional, y su alta versatilidad. Sin embargo, uno de sus principales inconvenientes es el proceso de discretización del espacio de observaciones, inherente a este tipo de modelado. Esta discretización, usualmente llevada a cabo a través de un proceso cuantización vectorial (VQ), produce pérdidas de información, que pueden deteriorar el rendimiento del sistema [1].

Al menos se han propuesto dos alternativas al modelado DHMM en la bibliografía. La primera de éstas corresponde a los denominados modelos ocultos de Markov continuos (CHMM). En este tipo de modelos, el problema de la discretización del espacio de características se obvia modelando directamente las probabilidades de observación de los modelos como gausianas multivariadas. El principal inconveniente de este tipo de modelado es su alto coste computacional, y el gran número de parámetros que es necesario estimar para la caracterización de las densidades de probabilidad de observación de los modelos (principalmente las matrices de covarianza). De esta forma, cuando el conjunto de datos disponible para el entrenamiento es insuficiente para la adecuada estimación de los parámetros de los modelos, el rendimiento del sistema se ve deteriorado [1].

La segunda alternativa corresponde a los denominados modelos ocultos de Markov semicontinuos (SCHMM), inicialmente propuestos por Huang [2]. En esta aproximación, el diccionario VQ es modelado como un conjunto de gausianas multivariadas, y el proceso VQ es modificado de forma que se generan múltiples candidatos para cada vector observable (uno por cada una de las gausianas). Cada candidato tiene asociada una probabilidad, de acuerdo con la función densidad de probabilidad del centroide correspondiente del diccionario. Estas probabilidades son utilizadas para obtener las probabilidades de observación (ver expresión (5)).

En este trabajo introduciremos una nueva variante del modelado HMM discreto. Esta nueva aproximación utiliza un proceso VQ específico para cada uno de los modelos HMM, en lugar de un proceso

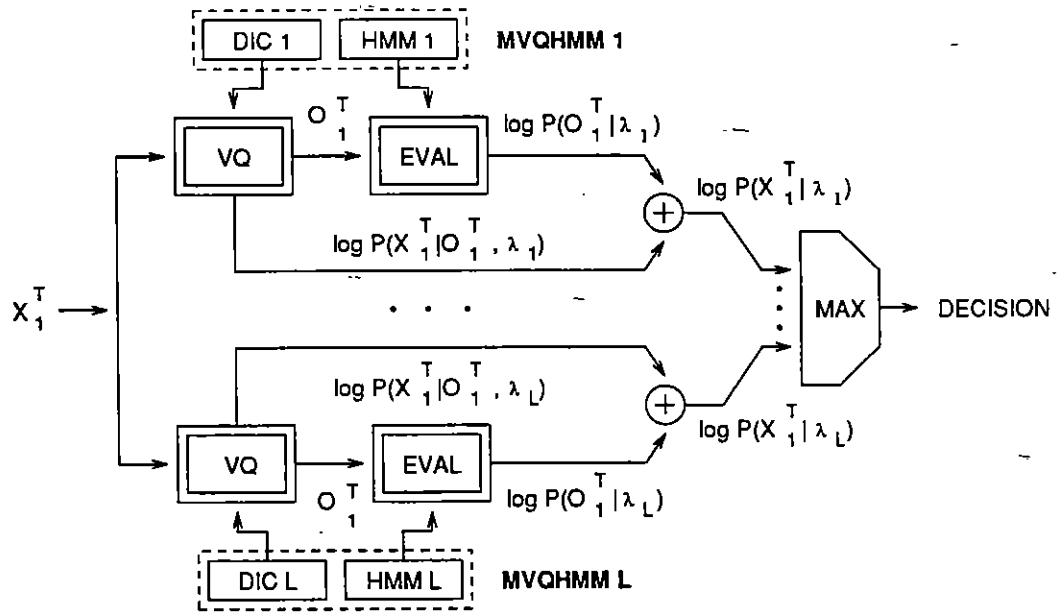

Figura 1: Sistema de reconocimiento basado en modelos MVQHMM

VQ con un diccionario común a todos los modelos. Por este motivo, en adelante denominaremos a éstos modelos HMM con cuantización múltiple (MVQHMM).

Para cada secuencia de vectores observables, el proceso VQ genera una secuencia de símbolos para cada uno de los posibles modelos; estas secuencias de símbolos se obtienen al cuantizar la secuencia de vectores con el diccionario VQ de cada uno de los modelos MVQHMM. Las secuencias de símbolos así obtenidas son utilizadas para evaluar las probabilidades de generación de cada uno de los modelos HMM. Estas probabilidades de generación son combinadas con las distorsiones de cuantización obtenidas en el proceso VQ (una por cada diccionario), obteniendo una probabilidad total de generación que se utiliza como criterio de selección del modelo más adecuado a la secuencia de observaciones.

2 Modelos MVQHMM

Un modelo MVQHMM está compuesto por un diccionario VQ que modela las diferentes producciones acústicas del modelo, y un modelo HMM discreto, que modela el secuenciamiento temporal de los prototipos acústicos del diccionario VQ. De esta forma, cada unidad de decisión (fonema, palabra, etc.) dispone de su propio conjunto de prototipos acústicos, y por lo tanto cabe esperar que el proceso de cuantización con el diccionario VQ adecuado sea más preciso que en el caso de la utilización de un diccionario VQ común a todos los modelos. Además, no se realiza una decisión a priori sobre la cuantización óptima para la secuencia de observaciones, sino que ésta es aplazada y se lleva a cabo conjuntamente con la decisión final sobre el modelo más adecuado a dicha secuencia de observaciones.

Tal y como mostraron Burton y Shore [3] y otros autores [4,5], las distorsiones de cuantización de una secuencia de vectores de características acústicas con diccionarios VQ de diferentes modelos, pueden ser utilizadas como criterio de clasificación. En la evaluación de las probabilidades de generación de los modelos MVQHMM, las distorsiones de cuantización son combinadas con las probabilidades de generación de los HMM para obtener la probabilidad total de generación del modelo MVQHMM, que es la utilizada como criterio de clasificación. En la figura 1 se muestra el esquema de un sistema de reconocimiento de palabras aisladas basado en modelos MVQHMM. En la figura, $\log P(O_1^T|\lambda)$ representa la probabilidad logarítmica de generación de la secuencia de símbolos O_1^T por parte del modelo λ, y $\log P(X_1^T|O_1^T, \lambda)$ es el logaritmo de la probabilidad de cuantización de la secuencia de vectores X_1^T en la secuencia de símbolos O_1^T para el modelos λ. Esta probabilidad logarítmica es esencialmente proporcional a la distorsión media

de cuantización de la secuencia de vectores, si se elige convenientemente la forma de las densidades de probabilidad de los centroides del diccionario VQ (ver sección 2.2).

2.1 Composición de probabilidades

Dada una secuencia de observaciones $X_1^T = x_1 \cdots x_T$, donde x_t es un vector de parámetros acústicos, La probabilidad de generación $P(X_1^T|\lambda)$ para un HMM λ puede expresarse en la siguiente forma:

$$P(X_1^T|\lambda) = \sum_{S^T} P(X_1^T|S_1^T,\lambda)\cdot P(S_1^T|\lambda) \quad (1)$$

$$P(X_1^T|S_1^T,\lambda) = \prod_{t=1}^{T} P(x_t|s_t,\lambda) \quad (2)$$

$$P(S_1^T|\lambda) = P(s_1|\lambda)\cdot\prod_{t=2}^{T} P(s_t|s_{t-1},\lambda) \quad (3)$$

donde $S_1^T = s_1 \cdots s_T$ es una secuencia de estados, $X_1^T = x_1 \cdots x_t \cdots x_T$ es una secuencia de observaciones y λ representa el modelo, $P(x_t|s_t,\lambda)$ es la probabilidad de observación del vector x_t en el estado s_t del modelos, $P(s_1|\lambda)$ es la probabilidad inicial del estado s_1, y $P(s_t|s_{t-1},\lambda)$ es la probabilidad de transición del estado s_{t-1} al estado s_t. La sumatoria en S^T representa la suma sobre todas las posibles secuencias de estados para el modelo. Asumiendo que la densidad de probabilidad $P(x_t|s_t,\lambda)$ puede expresarse como una mezcla de funciones densidad de probabilidad, podemos escribir

$$P(x_t|s_t,\lambda) = \sum_{o_t\in V(s_t,\lambda)} P(x_t|o_t,s_t,\lambda)\cdot P(o_t|s_t,\lambda) \quad (4)$$

donde $V(s_t,\lambda)$ es un conjunto de vectores prototipo pertenecientes al estado s_t del modelo λ. Si $P(x_t|o_t,s_t,\lambda)$ son gausianas entonces la expresión anterior es esencialmente la misma que la utilizada en los modelos HMM contínuos con mezcla de gausianas, siendo $P(o_t|s_t,\lambda)$ los coeficientes de la mezcla.

Asumiendo que el conjunto de clases V es independiente del estado y el modelo considerados, podemos escribir

$$P(x_t|s_t,\lambda) = \sum_{o_t\in V} P(x_t|o_t)\cdot P(o_t|s_t,\lambda) \quad (5)$$

Esta expresión es equivalente a la utilizada por Huang [2] en su formulación del modelado HMM semi-continuo. En la implementación práctica, sólo se consideran los M términos de la sumatoria (5) con mayor valor de $P(x_t|o_t)$.

Además, si asumimos clases disjuntas o levemente solapadas, la expresión anterior puede ser aproximada en la forma

$$P(x_t|s_t,\lambda) = P(x_t|o_t^*)\cdot P(o_t^*|s_t,\lambda) \quad (6)$$

$$o_t^* = \arg\max_{o_t\in V}\{P(x_t|o_t)\} \quad (7)$$

que representa la formulación del modelado HMM discreto. Utilizando las expresiones (1-3) y (6-7) se pueden obtener las siguientes relaciones

$$P(X_1^T|S_1^T,\lambda) = P(X_1^T|O_1^{T^*})\cdot P(O_1^{T^*}|S_1^T,\lambda) \quad (8)$$

$$P(X_1^T|\lambda) = P(X_1^T|O_1^{T^*})\cdot P(O_1^{T^*}|\lambda) \quad (9)$$

Donde $O_1^{T^*} = o_1^* \cdots o_T^*$ es la mejor secuencia de símbolos (en el sentido de la expresión (7)). Nótese que debido a que $P(X_1^T|O_1^{T^*})$ es independiente del modelo actual, sólo es necesario evaluar las probabilidades de generación de la secuencia de símbolos, $P(O_1^{T^*}|\lambda)$.

Finalmente, las fórmulas de evaluación para la probabilidad de generación de un modelo MVQHMM puede obtenerse bajo la suposición de que el conjunto de clases V es disjunto y sólo depende del modelo considerado. Por lo tanto, podemos escribir

$$P(x_t|s_t,\lambda) = P(x_t|o_t^*,\lambda)\cdot P(o_t^*|s_t,\lambda) \quad (10)$$

$$o_t^* = \arg\max_{o_t\in V(\lambda)}\{P(x_t|o_t,\lambda)\} \quad (11)$$

A partir de las expresiones (1-3) y (10-11) es sencillo obtener la siguiente relación

$$P(X_1^T|\lambda) = P(X_1^T|O_1^{T^*},\lambda)\cdot P(O_1^{T^*}|\lambda) \quad (12)$$

En esta expresión $P(X_1^T|O_1^{T^*},\lambda)$ representa la probabilidad de cuantización de la secuencia de observaciones X_1^T en la secuencia de símbolos $O_1^{T^*}$ y $P(O_1^{T^*}|\lambda)$ es la probabilidad de que la secuencia de símbolos $O_1^{T^*}$ sea generada por el modelo λ.

2.2 Modelado de las probabilidades de cuantización

A diferencia de lo que ocurre en el modelado HMM discreto, en el modeloado MVQHMM las probabilidades de cuantización $P(X_1^T|O_1^T,\lambda)$ dependen del modelos considerado y, consecuentemente, deben ser estimadas para obtener la probabilidad de generación del modelo dada por la expresión (12). Para evaluar las probabilidades de cuantización, es preciso establecer un modelo para los centros del diccionario VQ.

En el presente trabajo, cada clase (centroide del diccionario VQ del modelo) se modela como una gausiana con matriz de covarianza identidad Σ_λ. Con esta suposición, la probabilidad de cuantización puede expresarse en la forma siguiente:

$$\Sigma_\lambda = \sigma_\lambda^2\cdot I \quad (13)$$

$$\frac{1}{T}\log P(X_1^T|O_1^{T^*},\lambda) = -\frac{p}{2}\log 2\pi - \frac{p}{2}\log\sigma_\lambda^2 - \frac{D_\lambda(X_1^T)}{2\sigma_\lambda^2} \quad (14)$$

$$D_\lambda(X_1^T) = \frac{1}{T}\sum_{t=1}^{T}||x_t - \mu_{o_t^*,\lambda}||^2 \quad (15)$$

donde $D_\lambda(X_1^T)$ es la distorsión media de cuantización de la secuencia de vectores X_1^T con el diccionario del modelo λ, p es el número de componentes del vector, $\mu_{t,\lambda}^*$ es el centroide más *cercano* al vector, y $p\sigma_\lambda^2$ es el valor esperado de dicha distorsión media. Por lo tanto, la probabilidad de cuantización logarítmica es proporcional a la distorsión media de cuantización.

2.3 Entrenamiento de los modelos MVQHMM

En una aproximación de estimación de máxima probabilidad, la estimación de los parámetros de un modelo MVQHMM debe realizarse maximizando la probabilidad total de generación dada por la expresión (12). Esta aproximación requiere de la optimización conjunta de los centroides del diccionario VQ y de los parámetros del modelo HMM asociado. Sin embargo, debido a la complejidad de esta aproximación, en el presenta trabajo se ha utilizado una aproximación sub-óptima en dos pasos. Esta aproximacion es análoga a la utilizada en la estimaciónd de máxima probabilidad de los parámetros de los modelos HMM discretos.

En el primer paso, se contruyen los diccionarios VQ de los modelos utilizando un algoritmo de agrupamiento. En el presente trabajo se ha utilizado un algoritmo LBG con inicialización por bipartición. Este proceso maximiza las probabilidades de cuantización, de acuerdo con la expresión (15), minimizando la distorsión (distancia) media entre los vectores de la secuencia de entrenamiento y los centroides del diccionario VQ del modelo.

En el segundo paso, los diccionarios VQ previamente construidos se utilizan para cuantizar las secuencias de vectores del conjunto de entrenamiento. Esta cuantización se realiza de forma que los vectores del conjunto de entrenamiento de un determinado modelo son cuantizados con el diccionario VQ de éste. Con las secuencias de símbolos así obtenidas, se estiman los parámetros del modelo HMM discreto asociado a cada MVQHMM utilizando un procedimiento ietrativo tipo Baum-Welch. Los modelos iniciales se construyen a partir de una segmentación lineal uniforme de las secuencias de entrenamiento.

Centroides	DHMM	SCHMM	MVQHMM
4-64	4.37%	3.02%	4.25%
8-128	3.59%	2.13%	2.43%
16-256	2.81%	1.56%	0.95%
32-512	2.03%	1.43%	0.87%

Tabla 1: Errores de reconocimiento

3 Resultados experimentales

Se han realizado experimentos de evaluación comparativos sobre sistemas de reconocimiento de palabras aisladas basados en modelos MVQHMM, y modelos HMM discretos y semicontinuos [6]. El vocabulario está formado por los 10 dígitos castellanos y seis palabras clave (CUERPO, HOMBRO, CODO, MUÑECA, MANO, DEDOS). Exsisten 3 repeticiones de cada palabra del vocabulario pronunciadas por 40 locutores (20 masculinos y 20 femeninos). Debido al reducido número de locutores que componen la base de datos, en la evaluación de los tres sistemas de reconocimiento, se utilizó un método similar al denbominado *leaving-one-out* [7]; 32 locutores (16 masculinos y 16 femeninos) se utilizaron para entrenamiento, y los 8 restantes (4 masculinos y 4 femeninos) para evaluación. Estos experimentos se repitieron para 5 particiones disjuntas de la base de datos, y los resultados se promediaron. Los resultados así obtenidos son equivalentes a los que se obtendrían utilizando una base de datos de 72 locutores, con un conjunto de entrenamiento de 32 locutores y un conjunto de evaluación de 40 locutores diferentes de los anteriores.

Las palabras se muestrearon con una frecuencia de 8 kHz y 12 bits de precisión. Se realizó un análisis LPC de orden de predicción 10, sobre segmentos de señal de 32ms desplazados cada 16ms. Cada segmento de señal se caracteriza como un vector de 33 componentes, formado por 16 coeficienes cepstrum, 16 coeficientes delta-cepstrum, y la delta-energía logarímica [8]. Tanto a los oeficientes cepstrum como a los delta-cepstrum se le aplicó un ventana de *lifter* de longitud 16 [9]. Los coeficientes delta-cepstrum y delta-energía se pesaron con valores 0.925 y 0.728 para rendimiento óptimo de la distancia euclídea utilizada en el proceso VQ [6].

En todos los casos, se utilizaron modelos HMM izquierda-derecha con 10 estados. La información sobre la duración de los estados se tuvo en cuenta a través de un postprocesador que modifica las probabilidades de generación de los modelos HMM [6,10]. En la tabla 1 se muestran los errores de reconocimiento para los sistemas basados en modelos discretos (columna DHMM), semicontinuos (columna SCHMM) y de cuantización dependiente (columna MVQHMM). La primera columna indica el número de centroides para cada uno de los diccionarios VQ de los modelos MVQHMM (primer valor) así como el número de centroides en el diccionario universal utilizado para los modelos DHMM y SCHMM (segundo valor). Como puede observarse, cada una de las filas de la tabla representa una configuración con un número total de centroides equivalente a los tres tipos de modelos utilizados. Para el caso de 4 centroides en modelo MVQHMM, el error de reconocimiento es similar al obtenido con modelos DHMM, pero superior al correspondiente a los modelos SCHMM. Esto es debido a que el número de centroides por modelos es insuficiente para modelar adecuadamente las diferentes producciones acústicas de éstos. Sin embargo, para un número de centroides superior (16 ó 32), el error de reconocimiento obtenido con los modelos MVQHMM es significativemente inferior al correspondiente a los modelos DHMM y MVQHMM.

Referencias

[1] L. R. Rabiner, B. H. Juang, S. E. Levinson and M. M. Sondhi, "Recognition of isolated digits using hidden Markov models with continuous mixture densities," *AT&T Tech. J.*, Vol.64 No.6, pp. 1211–1233, August 1985.

[2] X. D. Huang and M. A. Jack, "Unified techniques for vector quantisation and hidden Markov modeling using semi-continuous models," presented at Proc. ICASSP'89, 1989.

[3] J. E. Shore and D. K. Burton, "Discrete utterance speech recognition without time alignment," *IEEE Trans. Inform. Theory*, IT-29, pp. 473–491, July 1983.

[4] S. Furui, "A VQ-based preprocessor using cepstral dynamic features for speaker-independent large vocabulary word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-7 No.7, pp. 980–987, July 1988.

[5] A. F. Bergh, F. K. Soong and L. R. Rabiner, "Incorporation of temporal structure into a vector-quantization based preprocessor for speaker-independent, isolated word recognition," *AT&T Tech. J.*, Vol.64 No.5, pp. 1047–1063, May-June 1985.

[6] J. C. Segura, "Modelos de Markov con Cuantización Dependiente para Reconocimiento de Voz," Dept. de Electrónica y Tecnología de Computadores. Universidad de Granada, Tesis Doctoral, Noviembre 1991.

[7] R. O. Duda and P. E. Hart, "Estimating the error rate," *Pattern Classification and Scene Analysis*, Vol.1, pp. 211–256, 1973.

[8] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34 No.1, pp. 52–59, February 1986.

[9] B. H. Juang, L. R. Rabiner and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust., Speech., Signal Proccesing*, ASSP-35 No.7, pp. 947–954, July 1987.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications," *Proc. IEEE*, 77 No.2, pp. 257–286, February 1989.