FUNDAMENTOS DE MODELOS DE MARKOV ESCONDIDOS (HMM)

*Marco Antônio Rocca de Andrade**

SUMÁRIO

Dentre as várias técnicas de modelagem de fenômenos físicos está a modelagem por máquina de estados finitos em cadeias de Markov. De uma extensão dessa técnica no meio dos processos estocásticos surgiram os modelos de Markov escondidos ou Hidden Markov Models (HMM). *Este artigo expõe de forma genérica os princípios básicos de HMM visando a apresentar mais esta ferramenta de modelagem.*

INTRODUÇÃO

Uma cadeia de Markov[1,2,3] é um conjunto finito de elementos, formando uma máquina de estados. Nesta máquina de estados as transições entre os estados não são governadas por regras determinísticas mas por *probabilidades de transição* entre eles. Porém, se em cada estado uma determinada saída ou observação puder ser gerada de acordo com uma distribuição de probabilidade (ao invés das regras determinísticas normalmente encontradas em máquinas de estados) e se somente a saída ou observação (e não o estado que a gerou) for visível a um observador externo ao processo, então os estados estarão 'escondidos' do exterior. Daí o nome de Modelos de Markov Escondidos[1,4,5,6,7] (*Hidden Markov Model – HMM*).

Diversos fenômenos podem ser modelados por meio de máquinas de estados finitos. E quando os fenômenos possuem características de processos estocásticos, pode-se pensar em usar HMM's como formas de modelá-los.

* Departamento de Engenharia Elétrica, Instituto Militar de Engenharia – IME.

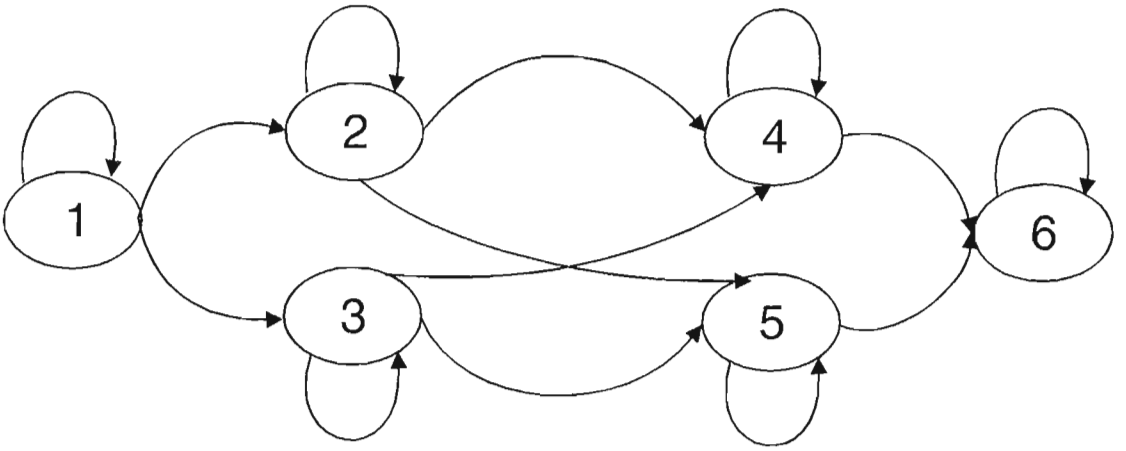

Figura 1: Diagrama genérico de uma máquina de estados com 6 estados

Como exemplo de um fenômeno trivial, em que se pode usar HMM para a modelagem, considere-se o caso de um senhor que recebe da esposa a tarefa de ir à feira e trazer várias frutas relacionadas em uma lista de compras. Após algum tempo, o senhor retorna e apresenta à esposa as frutas compradas conforme a lista. Neste caso, as saídas ou *observações* são as frutas obtidas e é somente o que a dona de casa pode avaliar como observadora externa do processo. Não está explícito nas frutas qual foi o caminho percorrido pelo consumidor através das finitas barracas e o que motivou esse caminho. Aspectos como distâncias, preços, quantidades e qualidades das frutas influenciaram no trajeto, e estes aspectos podem possuir características estocásticas. Ir à feira com uma lista de compras pode ser um problema modelável com HMM.

Uma vez associado o modelo ao fenômeno, pode-se responder a perguntas tais como: qual o melhor trajeto de visita pelas barracas ou se determinado trajeto possibilita a aquisição de toda a lista de compras de forma satisfatória.

ELEMENTOS DE UM HMM

Para definir um HMM de forma completa são necessários os seguintes elementos:

- O número de estados do modelo, N;
- O número de símbolos observáveis em um alfabeto, M. Para símbolos discretos, M pode ser inteiro, para símbolos contínuos, M pode ser infinito;

- O conjunto de probabilidades de transição de estados $A = \{a_{ij}\}$:

$$a_{ij} = p\{q_{t+1} = j \mid q_t = i\}, \qquad 1 \le i,j \le N; \tag{1}$$

onde q_t denota o estado corrente. As probabilidades de transição devem satisfazer as condições estocásticas:

$$a_{ij} \ge 0, \qquad 1 \le i,j \le N \qquad \text{e} \tag{2}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \qquad 1 \le i \le N. \tag{3}$$

- A distribuição de probabilidade de cada estado, $B = \{b_j(k)\}$:

Se a distribuição é discreta,

$$b_j(k) = P\{o_t = v_k \mid q_t = j\}, \qquad 1 \le j \le N, \qquad 1 \le k \le M; \tag{4}$$

onde v_k denota o k-ésimo símbolo do alfabeto, e o_t a observação corrente. Novamente as seguintes condições estocásticas devem ser satisfeitas:

$$b_j(k) \ge 0, \qquad 1 \le j \le N, \qquad 1 \le k \le N, \qquad \text{e} \tag{5}$$

$$\sum_{k=1}^{M} b_j(k) = 1, \qquad 1 \le j \le N. \tag{6}$$

Se a distribuição é contínua, normalmente, são especificados os parâmetros de uma função densidade de probabilidade que é representada por um somatório ponderado de M distribuições gaussianas,[1,5]

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm}\, N(O_t, \mu_{jm}, U_{jm}) \tag{7}$$

onde: c_{jm} são os coeficientes de ponderação das gaussianas N, μ_{jm} são vetores de médias, e U_{jm} são matrizes covariâncias.

As seguintes condições estocásticas devem ser atendidas por c_{jm}:

$$c_{jm} \ge 0, \qquad i \le j \le N, \qquad 1 \le m \le M, \qquad e \tag{8}$$

$$\sum_{m=1}^{M} c_{jm} = 1, \qquad i \le j \le N, \tag{9}$$

- Distribuição do estado inicial, $\pi = \{ \pi_i \}$, onde

$$\pi_j = p\{q_1 = j\}, \qquad i \le j \le N. \tag{10}$$

Com os elementos definidos acima, um HMM com distribuição de probabilidade discreta pode ser representado pela notação compacta:

$$\lambda = (A, B, \pi), \tag{11}$$

onde A é a matriz de elementos a_{ij}, B é o vetor de elementos b_k, e π é o vetor de elementos π_i. E um HMM com distribuição de probabilidade contínua pode ser representado com a notação compacta:

$$\lambda = (A, c_{jm}, \mu_{jm}, U_{jm}, \pi), \tag{12}$$

onde o vetor B é substituído pelos vetores de ponderação c_{jm}, médias μ_{jm} e pela matriz covariância U_{jm}.

Retornando ao exemplo da feira, o número de estados é o número N de barracas existentes. A matriz de probabilidades de transição $A = \{a_{ij}\}$ é uma matriz quadrada de NxN elementos. Os valores destas probabilidades a_{ij} podem ser determinados por um estudo (aprendizado) do comportamento de diferentes fregueses na mesma feira e com a mesma lista de compras. O número de observações ou símbolos do alfabeto é, a princípio, o número de frutas constantes na lista, com M elementos. Porém, se for considerado que cada fruta encontrada possui características únicas e contínuas como tamanho, aparência, preço, qualidade e peso, o número de observações possíveis passa a ser contínuo e incontável. A probabilidade de distribuição $B = \{ b_j(k) \}$ reflete a possibilidade de encontrar na barraca j uma fruta k com boas características de preço ou qualidade ou quantidade etc. A função de distribuição de probabilidade *(fdp)* associada a B pode não ser uma simples gaussiana para um determinado estado ou barraca e pode apresentar vários pontos de máximos locais, tornando-se uma função multimodal na prática. Supondo que os picos da *fdp* da qualidade das frutas fossem para frutas verdes e para frutas 'passadas', pode-se tentar modelar esta *fdp* por uma combinação de no mínimo duas gaussianas. A combinação de gaussianas é uma forma muito popular de obter aproximações para *fdp* complexas.[1,4,5,7] Para o modelo em questão, além do atributo qualidade outros atributos também são relevantes e devem participar da composição de B. Assim, para o atributo preço poderá existir também uma combinação de gaussianas que melhor representará o comportamento da *fdp* associada, o mesmo ocorrendo

para os demais atributos pertinentes ao modelo para cada estado. Assim, ao invés do vetor B para o caso discreto, ter-se-á os vetores c_{jm} e μ_{jm}, e a matriz U_{jm} englobando todas as gaussianas de todos os atributos de todos os estados para representar o modelo λ. Novamente, o levantamento destas probabilidades exige um estudo (aprendizagem) sobre o que é apresentado pelos feirantes em suas barracas. E, finalmente, as barracas que se localizam nas pontas da feira possuem uma maior probabilidade de serem visitadas inicialmente por um freguês recém-chegado. Estas barracas apresentariam valores de π_j maiores que as demais, e um estudo seria necessário para avaliar π para todas as barracas. Ir à feira com uma lista de compras teria um modelo representado pela equação 12.

SIMPLIFICAÇÕES NA TEORIA DE HMM

Com o objetivo de facilitar o trato matemático e computacional, algumas suposições são feitas na teoria de HMM:

• *Suposição de Markov*: o estado seguinte na máquina de estados depende somente do estado atual. A aplicação desta suposição gera um modelo chamado de primeira ordem. Pode-se construir modelos de ordem maiores em que o próximo estado dependa do estado atual e de n estados anteriores, porém o trato matemático e computacional cresce em complexidade muito mais do que a qualidade dos resultados.

• *Suposição de estacionaridade*: as probabilidades de transição de um estado para outro não se alteram no tempo.

• *Suposição de independência de saídas*: uma dada observação de saída é estatisticamente independente da observação da saída anterior.

Embora estas suposições sejam bastante limitadoras, em geral o desempenho dos HMM não é seriamente prejudicado.[4,6]

PROBLEMAS BÁSICOS DOS HMM

Após optar por modelar um fenômeno por meio de HMMs, três problemas são de grande interesse:

• *Problema da avaliação*: dado um modelo λ e uma seqüência de observações $O=o_1,o_2,o_3,...,o_T$, qual a probabilidade desta observação ter sido gerada pelo modelo λ, $p\{O|\lambda\}$?

• *Problema da decodificação*: dado λ e uma seqüência de observações $O=o_1,o_2,...,o_T$, qual a melhor seqüência dentro do modelo capaz de gerar essas observações?

• *Problema do aprendizado*: dado λ e uma seqüência de observações $O=o_1,o_2,...,o_T$, como ajustar os parâmetros $\lambda=\{A, B, \pi\}$ de modo a maximizar o valor $p\{O|\lambda\}$?

Problema da Avaliação

Determinar $p\{O|\lambda\}$ a partir de $O=o_1,o_2,o_3,...,o_T$ e do modelo λ pode ser realizado por meio de processos probabilísticos básicos, mas esse cálculo envolve um número de operações da ordem de N^T, onde T é o número de observações. Mesmo que o número de observações não seja muito grande, o número de operações é elevado. Assim, métodos alternativos foram criados para reduzir a complexidade computacional para encontrar

$$p\{O|\lambda\} = \sum_{i=1}^{N} p\{O, q_t = i \mid \lambda\} . \qquad (13)$$

Um desses métodos faz uso das variáveis auxiliares $\alpha_t(i)$ (chamada de *forward variable* ou *variável progressiva*) e $\beta_t(i)$ (chamada de *backward variable* ou *variável regressiva*). A variável progressiva é definida como a probabilidade da seqüência parcial de observação $\mathbf{o_1, o_2,..., o_T}$, quando esta termina no estado i, ou seja,

$$\alpha_T(i) = p\{ o_1, o_2,..., o_T, q_T = i \mid \lambda\}. \qquad (14)$$

E a variável auxiliar $\beta_t(i)$ é definida como a probabilidade da seqüência parcial de observação $o_{t+1}, o_{t+2}, o_{t+3},..., o_T$ ter ocorrido sendo i o estado atual e λ o modelo,

$$\beta_t(i) = p\{o_{t+1}, o_{t+2}, ..., o_T \mid q_t = i, \lambda\}. \qquad (15)$$

Desta forma, aplicando recursividade, obtém-se as relações:

$$\alpha_{T+1}(j) = b_j(o_{T+1}) \sum_{i=1}^{N} \alpha_T(i)a_{ij}, \qquad i \le j \le N, \qquad 1 \le t \le T-1, \quad \text{e} \qquad (16)$$

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j)a_{ij}b_j(o_{t+1}) , \qquad 1 \le i \le N, \qquad 1 \le t \le T-1, \qquad (17)$$

sendo $\alpha_1(j) = \pi_j b_j(o_1)$, $1 \le j \le N$, e $\beta_T(i) = 1$, $1 \le i \le N$. Disto resulta que a probabilidade procurada pode ser obtida por:

$$p\{O|\lambda\} = \sum_{i=1}^{N} p\{O,q_t = i|\lambda\} = \sum_{i=1}^{N} \alpha_T(i) = \sum_{i=1}^{N} \beta_1(i) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \qquad (18)$$

A complexidade do cálculo reduz-se de N^T para N^2T operações

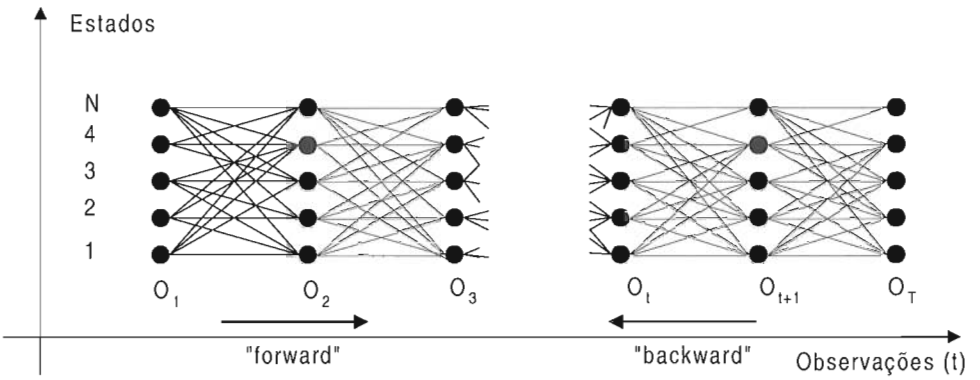

Figura 2: Implementação do cálculo das variáveis *Forward* e *Backward*

Problema da Decodificação

Dado um modelo λ e uma seqüência de observações $O = o_1, o_2, ..., o_T$, deseja-se saber qual a seqüência de estados com maior probabilidade de ter gerado O.

Uma forma de solução seria procurar, na seqüência, qual o estado mais provável como gerador de determinada observação e, depois, encadear todos os estados encontrados. Um problema nessa abordagem é que a solução pode conter seqüências sem significado para o modelo.

Existem vários critérios de otimização que evitam esse problema. Um dos mais usados é o que encontra a melhor seqüência simples de estados, ou seja, maximiza $P(Q/O,\lambda)$, onde Q representa uma seqüência de estados como $q_1, q_2, ..., q_t$. Para encontrar a seqüência Q faz-se uso de um algoritmo de programação dinâmica chamado de algoritmo de Viterbi.[4,5,6,7,8] Nele é definida a variável auxiliar $\delta_t(i)$ como se segue:

$$\delta_t(i) = \max_{q_1,q_2,...,q_{t-1}} p\{q_1, q_2, ..., q_{t-1}, q_t = i, o_1, o_2, ..., o_{t-1} | \lambda\}, \qquad (19)$$

a qual fornece a maior probabilidade que uma dada seqüência parcial de observação ser gerada por uma seqüência de estados até o estado t, quando o estado atual é i.

Fazendo $\delta_1(j) = \pi_j b_j(o_1)$, $1 \le j \le N$ e usando a recursividade, chega-se à relação:

$$\delta_{t+1}(j) = b_j(o_{t+1}) \left[\max_{1 \le i \le N} \delta_t(i) a_{ij}\right], \qquad 1 \le i \le N \qquad 1 \le t \le T-1 \qquad (20)$$

É necessário fazer um acompanhamento do argumento que é maximizado para cada instante t para decodificar a seqüência de estados mais provável de gerar O.

O processo possui uma sistemática semelhante ao utilizado em algoritmos de alinhamento temporal dinâmico[5,7,8] (*Dinamic Time Warp – DTW*).

Problema do Aprendizado

Geralmente, o problema do aprendizado é como ajustar os parâmetros do HMM do modelo λ para que, dado um grupo de observações chamadas de observações de treinamento, o modelo passe a representar estas observações da melhor forma para uma determinada aplicação. Não existe solução ótima para uma quantidade finita de observações de treinamento. O que se faz é tentar maximizar localmente a função $P(O/\lambda)$[7] para um dado modelo λ. Como critério de maximização mais usado cita-se o da *máxima verossimilhança (Maximum Likelihood – ML)*. Entre os algoritmos mais utilizados atualmente estão o Algoritmo de Reestimação de Baum-Welch[5,7] (baseado no processo *Forward & Backward*), o Procedimento de Viterbi (baseado no algoritmo de Viterbi) e o *Segmental K-mean.*[5]

O algoritmo de Baum-Welch faz uso das variáveis auxiliares α e β (progressiva e regressiva) para compor uma terceira variável $\gamma_t(i)$, chamada de *variável de probabilidade a posteriori*, correspondente à probabilidade de estar no estado i no tempo t dada a seqüência de observações O, representada pela relação:

$$\gamma_t(i) = P(q_t = i \mid O, \lambda). \tag{21}$$

De posse das três variáveis auxiliares e dos elementos que compõem o modelo inicial a ser ajustado, os parâmetros do novo modelo discreto λ^* é dado por:

$$\pi_j^* = \text{nº esperado de vezes do estado } q_i \text{ no tempo } t \tag{22}$$

$$a_{ij}^* = \frac{\text{nº esperado de transições do estado } i \text{ para o estado } j}{\text{nº esperado de transições do estado } i} \tag{23}$$

$$b_j(k)^* = \frac{\text{nº esperado de vezes que } o_k \text{ é observado em } q_j}{\text{nº esperado de transições pelo estado } j} \tag{24}$$

Para o caso de λ *ser um modelo contínuo, o novo modelo terá* c_{jm}, μ_{jm}, e U_{jm} *ajustados por:*

$$c_{jm}^* = \frac{\text{nº esperado de ocorrer a Gaussiana } m \text{ no estado } j}{\text{nº esperado de ocorrências do estado } q_j} \tag{25}$$

$$\mu_{jm}^{*} = \frac{\text{nº esp. de ocorrer a Gaussiana } m \text{ em } q_j \text{ ponderada por } o_t}{\text{nº esperado de ocorrer } q_j \text{ e na mistura } m} \quad (26)$$

$$U_{jm}^{*} = \frac{\text{nº esperado de ocorrência da Gaussiana } m \text{ em } q_j \text{ ponderado pela matriz covariância}}{\text{nº esperado de estar no estado } q_j \text{ e na mistura } m} \quad (27)$$

Baseado no procedimento substitui-se λ_{inicial} *por* λ^* *e repete-se as reestimações até que não ocorra melhorias significativas em* $P(O|\lambda^{\infty})$.

No procedimento de Viterbi, em vez de valores esperados, são usadas as somas das transições ocorridas e observações obtidas ao longo da melhor seqüência de estados conseguida para as observações fornecidas. Os parâmetros a_{ij}^* são obtidos pela contagem do número de transições do estado i para o estado j, dividido pelo número de transições feitas a partir de q_i. As médias, covariâncias e coeficientes de misturas são obtidos para cada estado após o agrupamento dos vetores de observação em M grupos pelo algoritmo *K-means modificado*.[4,11] A média é realmente a média de todas as observações associadas a uma determinada Gaussiana, o mesmo ocorrendo para a covariância. O coeficiente de mistura é dado pelo número de observações classificadas no grupo dividido pelo número total de observações classificadas no estado.

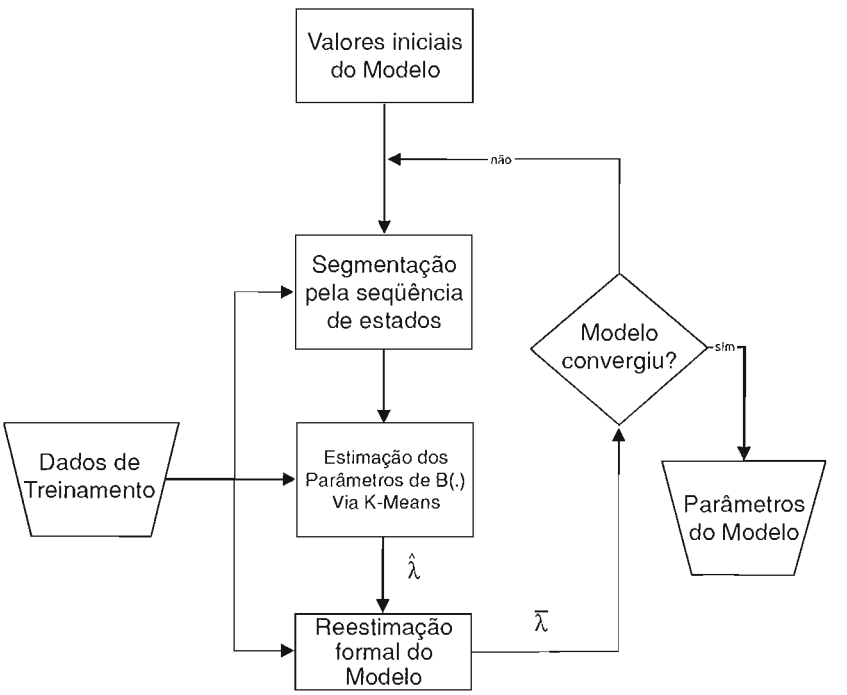

Figura 3: Algoritmo "Segmental K-means"

O algoritmo *Segmental K-means* surgiu como tentativa para solucionar um problema de sensibilidade aos valores iniciais do modelo λ observado nos dois algoritmos citados anteriormente. Neste processo, inicialmente divide-se as observações pelos estados (agrupamento) de forma seqüencial, aplica-se o procedimento de Viterbi para a obtenção de um modelo λ^*, o qual é usado no algoritmo de Baum-Welch para uma nova reestimação de todos os parâmetros. A diferença ou a verossimilhança dos modelos iniciais e reestimados são comparadas. Caso as diferenças estejam dentro de uma tolerância limite, então o processo atingiu uma convergência e finaliza. Caso contrário, o novo modelo é reintroduzido no algoritmo até que ocorra a convergência.

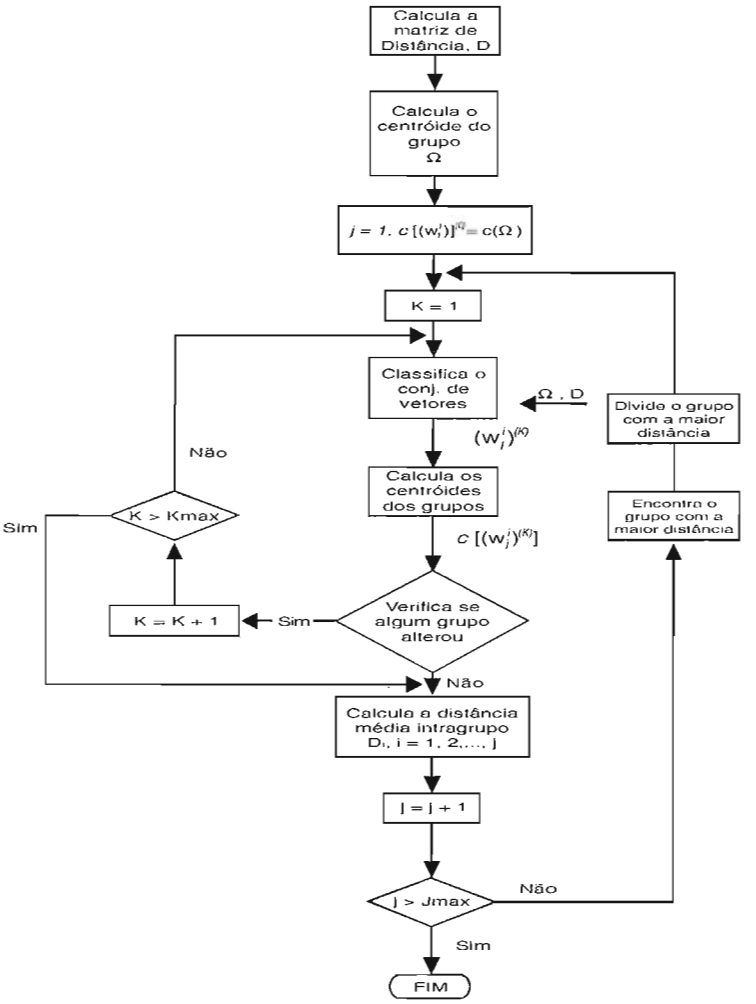

Figura 4: Fluxograma do Algoritmo "K-means" Modificado (MKM)

Uma descrição mais completa das fórmulas para treinamento de um HMM com o método *Segmental K-means* com mais de uma elocução pode ser encontrada em 4, 5 e 7.

APLICAÇÃO DE HMM EM RECONHECIMENTO DE VOZ[4,5,6,7,8,9,10]

Os sinais acústicos da fala podem ser convertidos em uma seqüência de números que represente a variação de amplitude no decorrer do tempo. A seqüência pode ser subdividida em quadros ou janelas contendo amostras vizinhas. E destas janelas pode-se extrair atributos da voz que sejam mais representativos do que puramente uma seqüência de amplitudes. O encadeamento de variações de atributos no tempo pode ser modelado por uma máquina de estados finita.

Na modelagem de voz por HMM, os atributos servem para caracterizar um dado fenômeno acústico como, por exemplo, o início de um fonema. Em geral, a duração da janela (unidade básica de tempo para fim de processamento de voz) fica em torno de 20 milissegundos, de modo que

determinado evento acústico ocorre em um período de algumas janelas. A partir de uma avaliação estatística do comportamento dos atributos pertencentes às janelas de um dado fenômeno acústico (etapa de treinamento), é possível criar um conjunto de valores numéricos – médias, covariâncias, coeficientes de ponderação e de transição – que passará a representar aquele fenômeno, definindo assim um *estado* da cadeias de Markov. Assim, uma dada elocução pode passar a ser representada por uma seqüência de grupos de valores estatísticos, os chamados estados. É comum atribuir três estados representantes do início, meio e fim de cada elocução.

Em uma etapa de decodificação, verifica-se se uma nova elocução possui em seus atributos características estatísticas semelhantes à seqüência de estados treinados. Obtém-se, então, um valor de verossimilhança entre elocução e modelo. Comparando-se valores de verossimilhança entre diferentes modelos e elocuções é possível determinar o par que melhor se adapta, e realizar o reconhecimento de uma palavra ou locutor.

CONCLUSÃO

Foram apresentados a teoria básica de HMM, as restrições que visam a simplificar a implementação e os três problemas básicos de avaliação, decodificação e aprendizagem. Devido às características estocásticas dos sinais de voz, a modelagem por HMM é bastante utilizada na área de reconhecimento de voz. Sendo assim, exemplos práticos da utilização como considerações de ordem computacional podem ser encontradas em trabalhos dessa área[4,8,9,10] dentre outras.

BIBLIOGRAFIA

1 – PAPOULIS, A. *Probability, Random Variables, and Stochastic processes*. McGraw-Hill, 1965.

2 – HOEL, P. G., PORT, S. C. e STONE, C. J. *Introduction to Stochastic Processes*. Houghton Miffin Company, 1972.

3 – ROSS, S. M. *Stochastic Processes*. John Wiley & Sons, 1983.

4 – PARANAGUÁ, E. D. S. *Reconhecimento de Locutores Utilizando Modelos de Markov Escondidos Contínuos*. Tese de Mestrado, IME, 1997.

5 – RABINER, L. R. e JUANG, B. H. *Fundamentals of Speech Recognition*. Prentice Hall, USA, 1993.

6 – WARAKAGODA, N. D. *A Hybrid ANN-HMM ASR system with NN based adaptive preprocessing*. M.Sc. thesis, Intitutt for Teleteknikk Transmisjonsteknikk, 1996.

7 – DELLER, P. e HANSEN, J. *Discrete-Time processing Speech Signals*. McMillan, 1993.

8 – SANTOS, S. C. B. e ALCAIM, A. *Fundamentos de Reconhecimento de Voz, Centro de Estudos em Telecomunicações da Pontifícia Universidade Católica do Rio de Janeiro*. CETUC-DID-01/95, setembro de 1995.

9 – SILVA, D. G. *Comparação entre os Modelos de Markov Escondidos Contínuos e as Redes Neurais Artificiais no Reconhecimento de Voz*. Projeto de Fim de Curso, IME, 1997.

10 – SANTOS, S. C. B. *Reconhecimento de Voz Contínua para o Português Utilizando Modelos de Markov Escondidos*. Tese de Doutorado, Departamento de Engenharia Elétrica, Pontifícia Universidade do Rio de Janeiro, 1997.

11 – WILPON, J. G. e RABINER, L. R. *A Modified K-Means Clustering Algorithm for Use in Isolated Word Recognition*. IEEE Transactions on Acoustics, Speech, and Processing, Vol. ASSP-33, nº 3, junho de 1985.