Washington César de Almeida Costa

Reconhecimento de Fala Utilizando Modelos de Markov Escondidos (HMM's) de Densidades Contínuas

Dissertação submetida ao corpo docente da Coordenação dos Cursos de Pós-Gradução em Engenharia Elétrica da Universidade Federal da Paraíba - Campus II como parte dos requisitos necessários para obtenção do grau de Mestre em Engenharia Elétrica.

Benedito Guimarães Aguiar Neto, Dr.-Ing., UFPB
Orientador

Marcos Antônio Gonçalves Brasileiro, D. Sc., UFPB
Orientador

Campina Grande, Paraíba, Brasil

Reconhecimento de Fala Utilizando Modelos de Markov Escondidos (HMMs) de Densidades Contínuas

Washington César de Almeida Costa

Dissertação de Mestrado aprovada em 20/06/1994

Benedito Guimarães Aguiar Neto, Dr.-Ing., UFPB
Orientador

Marcos Antônio Gonçalves Brasileiro - D. Sc., UFPB
Orientador

Mauro Cavalcante Pequeno - D. Sc., UFC
Componente da Banca

Marcelo Sampaio de Alencar, Ph.D., UFPB
Componente da Banca

Campina Grande, Paraíba, Brasil, junho/1994

Dedico este trabalho a minha esposa Silvana,

a minha filha Isabella,

e a minha avó Joaninha (*in memorian*)

"Portanto ofereçamos sempre por ele a Deus sacrifício de louvor ..."- Heb. 13:15.

Agradecimentos

Gostaria de expressar meus sinceros votos de gratidão as seguintes pessoas e instituições:

A minha esposa Silvana, pelo estímulo, ajuda, companheirismo e amor a mim dedicados ao longo de todo este trabalho.

Aos meus pais, Inajá e Lúcia.

Aos meus tios, Terezinha e Flávio, responsáveis maiores por todas as minhas conquistas até aqui alcançadas.

Aos professores, Benedito G. Aguiar Neto e Marcos A. G. Brasileiro, pela orientação e por terem possibilitado a execução deste trabalho.

Ao professor Marcelo Sampaio de Alencar, pelo voto de confiança.

Ao professor Francisco A. Moraes de Souza do DME, pela colaboração prestada.

Aos colegas Allan, Cortez, Adriano Fábio e Wallington, pela ajuda e presteza sempre presentes.

Aos amigos Medeiros, Nilza e Sandro, pelo apoio nos momentos mais difíceis.

A todos que contribuiram de alguma forma para a realização deste trabalho: Rosângela, Glauco, Joseana, Niomar, Kátia, Kíssia, Fabiana, Gleston, Avilez, Fernando, Ricardo e demais companheiros do LAPS.

A Ângela, por sua eficiência e dedicação junto a secretaria executiva da COPELE.

Ao CNPq, órgão financiador deste trabalho.

A Universidade Federal da Paraíba-Campus II, pela oportunidade oferecida.

Resumo

Nesta dissertação é realizado um estudo teórico e a implementação em *software* de um sistema de reconhecimento de fala baseado em Modelos de Markov Escondidos (*Hidden Markov Models* - HMM's).

HMM é uma ferramenta matemática que possibilita um modelamento dos sons da fala em termos de uma estrutura probabilística. Para tanto, utiliza-se, neste trabalho, HMM's do tipo *left-right* de cinco estados e fdp's contínuas, para representar a probabilidade dos vetores de observações em cada estado da cadeia de Markov. Os vetores de observações, de dimensão nove, são formados por oito coeficientes cepstrais e o logaritmo da energia segmental como o nono parâmetro.

O sistema HMM é composto de duas etapas: treinamento e classificação. Na fase de treinamento, o algoritmo de Baum-Welch é utilizado para reestimar os valores finais dos modelos. Por outro lado, na fase de classificação, utiliza-se o algoritmo de Viterbi para fornecer o valor da máxima verossimilhança entre a sentença de teste e os HMM's de referência.

A avaliação do sistema proposto é realizada considerando-se dois diferentes modos de reconhecimento: o reconhecimento independente do locutor e o reconhecimento dependente do locutor. Em ambos os casos, especialmente para o modo de reconhecimento dependente do locutor, as avaliações realizadas levam a resultados bastante satisfatórios, considerando-se as condições gerais de experimentação. Além disso, várias conclusões importantes são obtidas para uma posterior otimização do sistema proposto.

Finalmente, espera-se que este trabalho contribua de forma positiva para a motivação de novos estudos no campo da comunicação vocal homem-máquina.

Abstract

This dissertation presents a theoretical study and the software implementation of a speech recognition system, based on Hidden Markov Models (HMM's).

HMM is a mathematical tool that makes it possible modeling of the speech sounds in terms of a probabilistic structure. In order to do this, use is made in this work of HMM's of the left-right type with five states and continuous fdp's, to represent the observation vectors probability on each state of the Markov chain. The observation vectors, which are nine-dimensional, are formed by eight cepstral coefficients and the logarithm of the segmental energy as the nineth parameter.

The HMM system is divided into two stages: training and classification. In the training stage, the Baum-Welch algorithm is used to reestimate the final values of the models. On the other hand, the classification stage makes use of the Viterbi algorithm to provide the maximum-likelihood value between the test sentence and the reference HMM's.

The evaluation of the proposed system is made considering two different types of voice recognition: the independent speaker recognition and the dependent speaker recognition. In both cases, specially on the speaker dependent mode, the avaluation made given results really satisfactory, account to experimenting general conditions. In addition, some important conclusions are obtained in order to provide a posterior optimization on the proposed system.

Finally, it is expected that this work contributes in a positive way for the motivation of new studies on man-machine voice communication.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Símbolos

N - Número de estados de uma cadeia de Markov

$S_1, S_2, \ldots, S_N$ - Conjunto de estados de uma cadeia de Markov

$Q = q_1, q_2, \ldots, q_T$ - Seqüência de estados assumida em uma cadeia de Markov

$\mathbf{A} = \{a_{ij}\}$ - Matriz transição de estados

Δ - Incremento para o índice dos estados

$\Pi = \{\pi_i\}$ - Vetor de probabilidades inicial

V - Vetor de observações (vetor característico)

D - Dimensão do vetor de observações

T - Número total de vetores de observações de uma palavra

$\mathbf{O} = \{O_1, O_2, \ldots, O_T\}$ - Seqüência de observações de uma palavra

M - Número de símbolos utilizados na quantização vetorial

V - Alfabeto discreto

v_k - Símbolo do alfabeto V

B - Função densidade de probabilidade das observações

x - Vetor de observações contínuo

b_{jk} - fdp discreta $\quad$ $b_j(\mathbf{x})$ - fdp contínua

$\rho\ (\mathbf{x}, \vec{\mu}, \mathbf{U})$ - Função densidade de probabilidade multivariada

$\mathbf{C} = \{c_{jk}\}$ - Matriz ganho das misturas

$\vec{\mu} = \{\mu_{jk}\}$ - Vetor média

$\mathbf{U} = \{U_{jkde}\}$ - Matriz Covariância

M - Número de misturas

$\lambda = (\mathbf{A}, \mathbf{B}, \Pi)$ - Representação compacta dos parâmetros de uma HMM

$\overline{\lambda} = (\overline{\mathbf{A}}, \overline{\mathbf{B}}, \overline{\Pi})$ - modelo reestimado

$\alpha_t(i)$ - Variável progressiva (*forward*)

$\beta_t(i)$ - Variável regressiva (*backward*)

$\delta_t(j)$ - Medida de verossimilhança

$\Psi_t(j)$ - Caminho ótimo

P^* - Máxima verossimilhança

q_t^* - Seqüência de estados ótima

$\log f$ - logaritmo natural da máxima verossimilhança obtido através do algoritmo de Viterbi

$\xi_t(i,j)$ - Probabilidade da cadeia de Markov estar no estado S_i no tempo t, e no estado S_j no tempo t+1

$P(\mathbf{O}/\lambda)$ - Probabilidade da seqüência de observações, **O**, dado o modelo, λ

$\gamma_t(j)$ - Probabilidade da cadeia de Markov estar no estado S_i no tempo t, dada a seqüência de observações e o modelo

$D(\lambda, \overline{\lambda})$ - Distância entre o modelo inicial e o modelo reestimado

$S(n)$ - Sinal de voz discreto no tempo

$S_P(n)$ - Sinal de voz após a pré-ênfase

$W(n)$ - Janela de Hamming

$R_n(i)$ - Função de autocorrelação

p - Ordem do preditor

α_k - Coeficientes do preditor

k_i - Coeficientes de autocorrelação parcial (PARCOR)

$|\mathbf{U}|$ - Determinante da matriz covariância

$\mathbf{U}^{-1}$ - Inversa da matriz covariância

$\mathbf{V} = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_9)$ - Vetor aleatório

$\mathbf{V}'$ - Vetor aleatório transposto

$\mathrm{E}_{\max}$ - Energia máxima do sinal de voz

$\log \mathbf{E}$ - Logaritmo da energia segmental do sinal de voz

$\mathrm{E}^{(i)}$ - Erro de predição para um preditor de i-ésima ordem

c(i) - Coeficientes cepstrais

Lista de Abreviaturas

HMM's - Modelos de Markov Escondidos (*Hidden Markov Models*)

PCM - Modulação por codificação de pulsos (*Pulse Code Modulation*)

Conversão A/D - Conversão analógico-digital

LPC - Codificação preditiva linear (*Linear Prediction Coding*)

DTW - Comparação de padrões por alinhamento dinâmico na escala do tempo (*Dynamic Time Warping*)

PDS - Processador digital de sinais (*Digital Signal Processor*)

fdp - Função densidade de probabilidade

NA - Número de amostras em um quadro de voz

Coeficientes PARCOR - Coeficientes de correlação parcial

CONJ - Conjunto de dígitos

RAM - Memória de leitura e escrita aleatória (*Random Access Memory*)

DOS - Sistema operacional de disco (*Disc Operating System*)

DME - Departamento de Matemática e Estatística - UFPB

LAPS - Laboratório de Automação e Processamento de Sinais - UFPB

COPELE - Coordenação de Pós-Graduação em Engenharia Elétrica

Capítulo 1

Introdução

Embora não se esteja longe do uso da fala para a comunicação com computadores, a comunicação homem-máquina ainda é uma barreira para a maioria das pessoas. Interfaces vocais oferecem a possibilidade de interação com máquinas usando a forma mais natural e bem desenvolvida da comunicação humana - a voz.

A promessa da tecnologia de reconhecimento de fala é que ela seja capaz de remover a barreira da comunicação entre as pessoas e suas máquinas. Isto fará com que o poder dos computadores esteja disponível a qualquer pessoa e auxilie, mais efetivamente, àqueles que já os usam diariamente.

Além da facilidade de comunicação, a voz oferece muitas outras vantagens na comunicação com as máquinas. Velocidade é uma: a maioria das pessoas pode falar facilmente em taxas de 200 palavras por minuto, por outro lado, poucas pessoas podem digitar, em um teclado, mais de 60 palavras por minuto [1]. A voz também pode remover algumas das limitações físicas da interação com os computadores: pode-se interagir com o computador enquanto se trabalha no escuro, ou se utilizar, para a entrada de dados, um microfone em ambientes onde haja incoveniência para o uso de um teclado.

Interfaces vocais não estão limitadas a controlar computadores. Outras áreas de

aplicação envolvem a integração da tecnologia vocal com a telefonia convencional ou móvel, interface visual, fac-símile e controle de sistemas robóticos.

Várias são as técnicas existentes para o reconhecimento automático de fala. Entre essas, a técnica baseada em modelos de Markov escondidos (HMM's - *Hidden Markov Models*) é, atualmente, a tecnologia predominante na maioria dos sistemas de reconhecimento, em pesquisa [1]. Isto se deve, principalmente, à sua grande capacidade de generalização e de abstração das propriedades dos sons da fala em uma estrutura matemática probabilística. Tornando-se, assim, uma das tecnologias mais atrativas para o caso do reconhecimento de fala.

Neste trabalho são estudados os modelos de sinais baseados em HMM's de densidades contínuas para o reconhecimento de dígitos isolados (de zero a nove) falados de maneira dependente e independente do locutor. Os modelos de HMM's utilizados são baseados em cadeias de Markov do tipo *left-right* com cinco estados e funções densidade de probabilidade (fdp's) contínuas, para representar os vetores de observações em cada estado do modelo. Os vetores de observações são formados por parâmetros cepstrais e o logaritmo da energia segmental.

A importância deste trabalho de pesquisa, inédito do ponto de vista da utilização de HMM's de densidades contínuas para o reconhecimento automático de fala na língua portuguesa do Brasil, deve-se principalmente à aquisição do *know-how* da tecnologia em desenvolvimento nos institutos de pesquisa por todo o mundo.

A organização desta dissertação está disposta como segue:

O capítulo 2 apresenta os aspectos básicos da comunicação vocal homem-máquina. As áreas da comunicação vocal são descritas e uma maior ênfase é dada para os sistemas de reconhecimento de fala, objeto desse trabalho. Ainda nesse capítulo, são mencionadas as técnicas mais usuais para o reconhecimento de palavras isoladas e uma breve explanação é feita sobre o método de reconhecimento de palavras isoladas baseado na técnica DTW ("Dynamic Time Warping" - Comparação de vetores de padrões por alinhamento dinâmico na escala do tempo). Esse capítulo termina com a apresentação do estado da arte e perspectivas futuras para a tecnologia do reconhecimento de fala.

O capítulo 3 discute os aspectos teóricos dos modelos de Markov escondidos aplicados ao reconhecimento automático de fala colocados em termos de três problemas fundamentais: 1) O cálculo da probabilidade de uma seqüência de observações dado um HMM específico. 2) A determinação de uma seqüência ótima de estados do modelo. 3) O ajuste dos parâmetros do modelo de modo que melhor representem o sinal que está sendo modelado. As soluções para esses três problemas são também apresentadas. Ainda nesse capítulo são abordados temas como: Tipos de HMM's, número de estados do modelo, a matriz transição de estados, a matriz ganho das misturas, a função densidade de probabilidade das observações, a utilização de fdp's contínuas para representar as observações em cada estado, seqüências de observações múltiplas, estimativas iniciais e finalmente as limitações dos HMM's.

No capítulo 4 são apresentados os detalhes de implementação de um reconhecedor de palavras isoladas utilizando HMM's com densidades contínuas. São então relacionadas e discutidas todas as etapas que compõem cada uma das duas fases (treinamento e classificação) presentes no sistema de reconhecimento utilizado.

O capítulo 5 relaciona todas as avaliações e resultados experimentais realizados com o sistema proposto. Essas avaliações são feitas, inicialmente, considerando-se um sistema de reconhecimento independente do locutor. Posteriomente, outras avaliações são levadas a efeito para um reconhecimento do tipo dependente do locutor.

No sexto e último capítulo, são apresentadas as conclusões gerais a respeito do sistema proposto, como também são dadas algumas sugestões para continuidade deste trabalho e para trabalhos futuros a serem realizados na área da comunicação vocal homem-máquina.

Capítulo 2

Comunicação vocal homem-máquina

2.1 Introdução

Dentre as várias áreas que compõem o campo da comunicação por voz, a área da comunicação vocal homem-máquina é uma das mais interessantes e estimulantes. O desejo, bem como a necessidade, das pessoas se comunicarem com as máquinas na maneira mais natural de comunicação, isto é, a voz humana, tem dado um grande impulso ao crescimento desta área. A capacidade de fornecer comunicação vocal permitiria a máquinas complexas serem acessadas e controladas por grandes grupos de pessoas não treinadas.

Aplicações industriais de sistemas de entrada vocal geralmente envolvem dados que devem ser capturados em suas fontes, ou um processo que deve ser controlado em tempo real. Entrada vocal é particularmente vantajosa nessas situações quando uma ou mais das seguintes condições se aplicam:

- As mãos do usuário estão ocupadas;

- Mobilidade é exigida durante o processo de entrada de dados;
- Os olhos do operador devem permanecer fixos sobre um *display*, um instrumento ótico, ou algum objeto a ser rastreado;
- Não seria conveniente o uso de um teclado no ambiente.

Entrada vocal é adequada para essas aplicações por não requerer nem as mãos, nem os olhos do usuário para sua operação. Algumas das aplicações de sistemas de entrada vocal são [2]:

- Controle de simuladores de alto desempenho para cabines de aviões;
- Controle de tráfego aéreo;
- Entrada de dados cartográficos e barométricos;
- Auxílio a deficientes físicos.

Outras aplicações de cunho industrial são:

- Controle de qualidade e inspeção;
- Movimentação automática de materiais de um ponto a outro;
- Entrada vocal direta para computadores.

2.2 Áreas da comunicação vocal homem-máquina

De uma forma geral, a área da comunicação vocal homem-máquina inclui os três modos de comunicação indicados abaixo:

- Sistemas de Resposta Vocal;
- Sistemas de Reconhecimento de Locutor;
- Sistemas de Reconhecimento de Fala.

2.2.1 Sistemas de resposta vocal

Os sistemas de resposta vocal são aqueles capazes de responder a algum pedido de informação usando mensagens faladas. Assim, esses sistemas comunicam voz em uma única direção, ou seja, da máquina para o usuário.

Para gerar a saída acústica para um vocabulário de várias centenas de palavras, é geralmente suficiente usar elementos de texto armazenados digitalmente, consistindo de frases, palavras, fonemas ou certos parâmetros chaves (codificação paramétrica) que podem ser concatenados para formarem a saída desejada.

Todos os métodos de codificação de forma de onda conhecidos (PCM[1], PCM diferencial, PCM diferencial adaptativo, etc.) e métodos de análise-síntese (técnicas de codificação preditiva linear) usando taxas em torno de 2,4 kbits/s até 64 kbits/s podem ser usados para armazenar os elementos de texto. A escolha do método a ser utilizado é uma função da qualidade da reprodução das mensagens e da capacidade de armazenamento exigidos pelo sistema. A qualidade da voz depende, essencialmente, do método de codificação utilizado. Entretanto, para um mesmo método, por exemplo, em codificação de forma de onda, aumentando-se a taxa de bits por amostra obtém-se uma melhor qualidade, contudo às custas de uma maior capacidade requerida para o armazenamento das mensagens. Baixas taxas de bits são obtidas, em geral, com técnicas de codificação paramétrica (métodos de análise-síntese de voz). Os valores dos parâmetros derivados dessa representação são, então, usados para controlar um sintetizador de voz que modela a produção da voz humana. De forma análoga, nesses codificadores, a qualidade do sinal sintetizado é uma função do número de bits por parâmetro.

A desvantagem do armazenamento de voz por codificação paramétrica reside no fato dos valores dos parâmetros, no processo de análise, serem obtidos pelo fabricante. Assim, o usuário não pode, em geral, modificar as mensagens de saída para adequá-las às suas exigências. Alguns problemas que ainda devem ser solucionados quanto à

[1]Modulação por Codificação de Pulsos

síntese de voz incluem, entonação incorreta de frases e pronúncia errônea de palavras mais complexas, ou de combinações de palavras [3].

2.2.2 Sistemas de reconhecimento de locutor

Para o problema de reconhecimento de locutor, a tarefa do sistema é verificar se um dado locutor é quem ele alega ser, ou identificar um determinado locutor dentre um conjunto pré-estabelecido de possíveis locutores. No primeiro caso, o locutor é dito cooperativo, uma vez que ele deseja sempre ser reconhecido pelo sistema. No segundo caso o locutor é dito não cooperativo, uma vez que, em geral, não deseja ser reconhecido pelo sistema, como é o caso de aplicações voltadas para criminalística.

A tecnologia de verificação de locutor é similar a de reconhecimento de fala, tornando-se assim, atrativa a combinação de ambas as técnicas no mesmo *hardware* para aplicações específicas.

2.2.3 Sistemas de reconhecimento de fala

Reconhecimento de fala é o passo decisivo em direção à simplificação da comunicação homem-máquina. É o processo através do qual, o usuário pode usar simplesmente comandos falados que podem ser reconhecidos e interpretados por um sistema de reconhecimento de fala automático.

A tarefa básica de um sistema de reconhecimento de fala é reconhecer exatamente toda a sentença falada, ou ainda, "entender" a expressão falada (isto é, responder corretamente de alguma maneira ao que foi falado). O conceito de entendimento, ao invés de reconhecimento, é de grande importância para sistemas que tratam com entrada de voz contínua com grande vocabulário, enquanto que o conceito de reconhecimento exato é de maior importância para sistemas de palavras isoladas, vocabulário limitado e pequeno número de usuários [4].

A tecnologia de reconhecimento de fala ainda não permite o entendimento automático de voz fluente, de qualquer locutor, usando a mesma linguagem. Os problemas de reconhecimento de fala por máquinas estão relacionados à estrutura complexa da voz humana, que depende de fatores tais como características vocais, entonação, velocidade da fala, estado emocional do usuário, etc.

Os métodos de reconhecimento de fala podem ser classificados em função de fatores como, tamanho e flexilibilidade do vocabulário, número de usuários, condições de fala, etc. De uma forma geral, todos os tipos de sistemas de reconhecimento automático de fala podem ser considerados como pertencentes a uma das seguintes categorias:

- Sistemas de Reconhecimento de Palavras Isoladas;
- Sistemas de Reconhecimento de Palavras Conectadas;
- Sistemas de Reconhecimento Dependente do Locutor;
- Sistemas de Reconhecimento Independente do Locutor.

2.2.3.1 Sistemas de reconhecimento de palavras isoladas

Os sistemas de reconhecimento de palavras isoladas podem ser definidos como aqueles sistemas que exigem uma pausa curta antes e depois das sentenças que devem ser reconhecidas [5].

A duração mínima de uma pausa que separa padrões independentes deve ser da ordem de 100 ms. Qualquer intervalo de tempo menor do que 100 ms pode ser confundido com as pequenas pausas produzidas pela presença de uma consoante oclusiva no meio de uma palavra [2].

A taxa de emissão de comandos falados que pode ser obtida com sistemas de reconhecimento de palavras isoladas é naturalmente muito menor do que para sistemas de palavras conectadas. Uma taxa de 120–150 dígitos/minuto foi obtida por T. B. Martin [2], para testes com locutores treinados, lendo dígitos em ordem aleatória. Outros testes forneceram taxas de fala média entre 30 e 70 palavras isoladas por minuto, em ambiente de fábrica [2].

2.2.3.2 Sistemas de reconhecimento de palavras conectadas

O modo de entrada de palavras conectadas pode ser conveniente para o usuário porque se assemelha à maneira mais natural de se falar, contudo este tipo de comunicação tem algumas limitações em vista do presente estágio da tecnologia de reconhecimento de fala. Além disso, a entrada de palavras conectadas não é equivalente à fala contínua, usando um vocabulário de milhares de palavras, assim, o usuário não se comunica com o sistema da mesma maneira como com outra pessoa.

Seria impossível restringir um grande número de usuários às limitações da tecnologia atual para palavras conectadas - vocabulário limitado com uma sintaxe restrita e cada frase de entrada falada sem pausas. Por outro lado, este tipo de sistema pode ser usado efetivamente em aplicações onde um pequeno grupo de usuários pode ser orientado em como usar o sistema [6, 7]. Taxas de fala acima de 300 palavras por minuto podem ser obtidas para curtos intervalos de palavras conectadas [2].

2.2.3.3 Sistemas de reconhecimento dependente do locutor

Os sistemas dependentes do locutor são caracterizados por serem treinados para obedecerem às características específicas da fala dos seus usuários.

As principais vantagens dos sistemas de reconhecimento de fala dependente do locutor são os vocabulários de várias centenas de palavras que podem ser definidas e produzidas pelo usuário e a facilidade para modificação e atualização desse vocabulário.

As principais desvantagens desses sistemas são que eles devem ser treinados para cada usuário, o que pode ser proibitivo para grandes vocabulários e grande número de usuários. O desempenho é sensível à mudanças na voz do usuário causadas por *stress*, fadiga ou rouquidão. Também o custo por usuário é alto, quando um dispositivo separado deve ser alocado para cada pessoa, ou pela necessidade de grande quantidade de memória para armazenar o vocabulário de cada pessoa.

O desempenho dos sistemas de reconhecimento de alta qualidade (voz contínua) está em torno de 90%, sob condições bem definidas de laboratório [8]. Na prática, contudo, a avaliação do equipamento de reconhecimento de fala é difícil porque envolve fatores

que afetam as variáveis a serem medidas, tais como a ocorrência de palavras similares, o estado físico e emocional do usuário, o tipo e localização do microfone e o ruído ambiental.

2.2.3.4 Sistemas de reconhecimento independente do locutor

Os sistemas de reconhecimento independente do locutor, ou sistemas "insensíveis" ao locutor, podem ser definidos como aqueles que não estão presos às características específicas da fala do usuário.

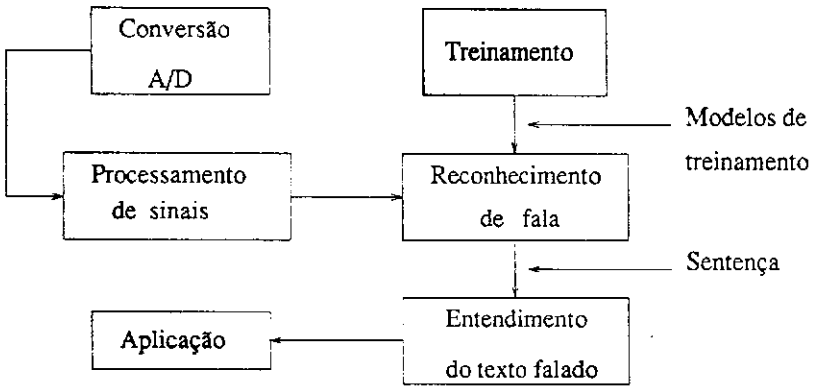
Os sistemas independentes do locutor oferecem um desempenho de reconhecimento aceitável para um grande número de usuários. Contudo, eles não desempenham igualmente bem para todos os locutores independente de sexo, dialeto, sotaque, ou comportamento da fala. O vocabulário de tais sistemas é fixo e muito menor do que aquele dos sistemas dependentes do locutor. Além disso, como o vocabulário tem de ser gerado pelo fabricante, a atualização deste é bastante dispendiosa.

A geração de palavras de referência independentes do locutor consome tempo, uma vez que as amostras da voz de um grande número de locutores devem ser colecionadas em um banco de dados para cada sentença, para cobrir as variações das pronúncias dos locutores para cada palavra falada.

Em geral, se a identificação do locutor e a modificação do vocabulário de aplicação são necessárias, um sistema de reconhecimento dependente do locutor deve ser utilizado [9].

2.3 Sistemas de comunicação vocal homem-máquina no reconhecimento de fala

A Figura 2.1, dada a seguir, mostra uma organização comum para sistemas de interface vocal. Todos eles têm, de uma forma ou de outra, três componentes básicos: processamento, reconhecimento e entendimento da fala [1].



Fonte: Lee (1990) [1], pag.226.

Figura 2.1: Sistema de entrada vocal consistindo de três componentes básicos: processamento, reconhecimento e entendimento da fala.

Inicialmente, um microfone converte as variações que a fala causa na pressão do ar em variações de tensão. Em seguida, o sistema amostra e digitaliza essas variações usando um conversor A/D. Os sistemas comumente utilizam freqüências de amostragem que variam de 8 khz a 20 khz. A seqüência de amostras resultante é chamada de forma de onda digital do sinal de voz.

Em princípio, pode-se tentar reconhecer diretamente a forma de onda representada digitalmente. Porém, uma sentença de apenas 1 segundo de fala pode produzir uma taxa de até 160 kbits/s (amostragem a 20 khz), resultando em um tempo de processamento proibitivo. Além disso, a forma de onda contém variações redundantes e irrelevantes. O processamento dessas informações seria também redundante e ineficiente. Assim, sistemas de voz utilizam técnicas de processamento digital de sinais para reduzir a redundância e enfatizar as características mais importantes do sinal de voz, reduzindo consideravelmente o volume de dados a ser processado [1].

Técnicas de redução mais conhecidas incluem bancos de filtros, como em codificação por sub-bandas [10], e transformadas rápidas de cosseno (DCT), como em codificação por transformação adaptativa [10] que, para cada quadro de amostras do sinal de voz, determinam o nível de energia em diferentes faixas de freqüências. Também é comum

o uso da codificação preditiva linear (LPC), que gera um vetor de coeficientes com informação espectral do sinal de voz [1].

Essas técnicas de redução produzem uma taxa de bits em torno de 8 kbits/s a 16 kbits/s, ou seja, uma redução por volta de 1/8 a 1/4 em relação à taxa obtida em sistemas PCM.

Reconhecimento de fala envolve a comparação de uma sentença - uma seqüência de vetores característicos - de entrada, com modelos de fala previamente armazenados, sujeitos a restrições léxicas e gramaticais. Em todos os casos, é necessário o treinamento dos modelos antes de usar o sistema para o reconhecimento. Estes modelos podem ser baseados em várias unidades da fala, dependendo do método utilizado. Alguns deles incluem, palavras completas ou frases, sílabas, ou fonemas.

Para algumas aplicações, tais como ditado automático, discagem telefônica vocal, e entrada de dados, necessita-se apenas recuperar a seqüência de palavras faladas. Mas para outras, o sistema deve entender o significado do que foi dito. Ou seja, o sistema de entendimento de fala tenta interpretar a intenção do locutor. Isso, muitas vezes, exige processamento complexo para levar em conta não apenas a estrutura da própria sentença, mas também os resultados de entradas anteriores e o conhecimento geral sobre o domínio da aplicação [1].

Reconhecimento de fala contínua é muito mais complexo do que reconhecimento de palavras isoladas por várias razões: o vocabulário é maior, as palavras são pronunciadas com menos cuidado, e as sentenças têm uma estrutura semântica-sintática que deve ser levada em consideração pelo sistema de reconhecimento [11, 12].

Entrada de voz sem restrições, gramaticais ou léxicas e entendimento da linguagem (algumas vezes realizadas sem esforço por pessoas) é muito difícil e permanece um problema sem solução, salvo para aplicações de pequenos vocabulários [1].

A partir da década de 70 tem havido um grande avanço na meta de se construir sistemas capazes de entender a voz humana. A razão principal desse progresso tem sido o desenvolvimento e a aplicação de métodos matemáticos que permitam modelar o sinal de voz como um código complexo com vários níveis de estrutura [13].

Entre os métodos matemáticos desenvolvidos para o reconhecimento automático de fala, os métodos que têm alcançado maiores resultados são: o método de comparação de padrões DTW; e o método de modelagem estocástica baseado em modelos de Markov escondidos [13].

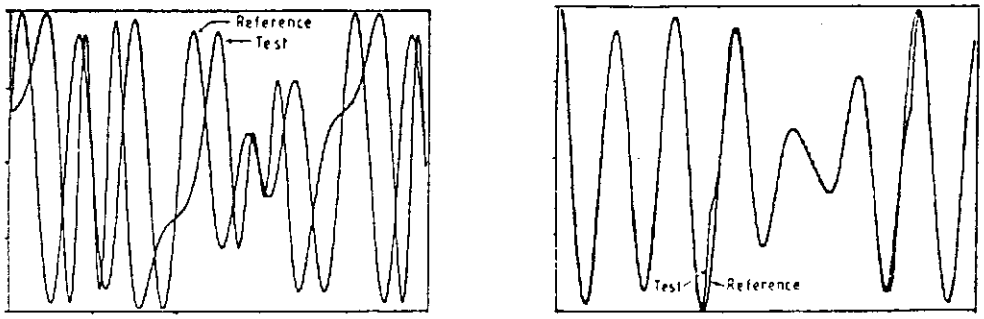
2.4 Técnicas utilizadas para o reconhecimento de palavras isoladas

Uma grande variedade de técnicas tem sido proposta para o reconhecimento de palavras isoladas. Dentre essas, a técnica de reconhecimento baseada no alinhamento dinâmico na escala do tempo dos vetores de padrões (DTW), que usa Programação Dinâmica para a comparação de padrões, é a que tem obtido maior sucesso [14]. DTW é largamente utilizada em reconhecedores de fala comerciais [1].

2.4.1 O reconhecedor de palavras DTW convencional baseado na análise LPC

O reconhecedor LPC/DTW tem como base o alinhamento de sinais distorcidos por pertubações lineares ou não lineares nos instantes de amostragem. Exemplos dessa classe de pertubações surgem devido ao uso de janelas ou espelhos não uniformes, variações na velocidade de gravação ou de transmissão, e variações no ritmo da voz [15].

A Figura 2.2, dada a seguir, mostra um sinal de referência e outro obtido a partir desse por uma distorção na escala do tempo (Fig. 2.2a), e o resultado da aplicação do processo de alinhamento aos dois sinais (Fig. 2.2b).

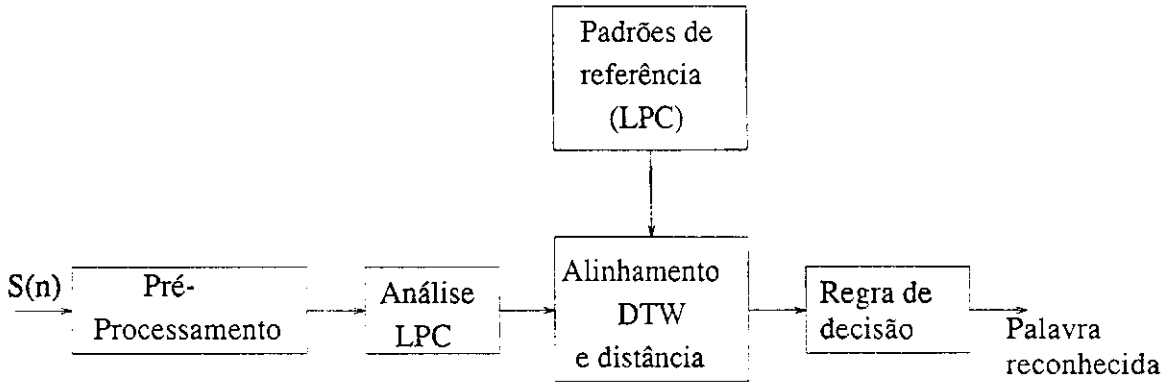


(a) (b)

Fonte: Maheswaran (1990) [15], pag.1648.

Figura 2.2: Alinhamento dinâmico no eixo do tempo (DTW). (a) Sinal de referência e sua versão distorcida. (b) Sinais após o alinhamento no eixo do tempo.

A Figura 2.3, dada abaixo, mostra um diagrama de blocos do reconhecedor de palavras isoladas LPC/DTW convencional [16].



Fonte: Rabiner (1983) [16], pag.1077.

Figura 2.3: Reconhecedor de palavras isoladas LPC/DTW convencional

O padrão de teste é obtido a partir de uma análise LPC realizada em blocos de

amostras do sinal de voz de entrada, S(n). Esse padrão de teste é comparado com cada padrão de referência usando um algoritmo de alinhamento DTW, que fornece simultaneamente uma medida de distância associada a esse alinhamento. As medidas de distância para todos os padrões de referência são enviadas a uma regra de decisão, que fornece uma classificação da palavra falada, e possivelmente um conjunto ordenado (pela distância) das melhores candidatas.

Os padrões de referência da palavra para o reconhecedor da Figura 2.3, são criados por um algoritmo de treinamento. Para aplicações dependentes do locutor, basicamente, um único padrão de referência é criado para cada palavra do vocabulário usando um algoritmo de treinamento robusto. Para aplicações independentes do locutor, um conjunto com vários padrões de referência é criado para cada palavra do vocabulário usando um procedimento de agrupamento ("clustering"). Tipicamente em torno de 12 vetores de padrões por palavra são suficientes para o reconhecimento das palavras [16].

No reconhecedor LPC/DTW o procedimento de treinamento trata-se de um processo de armazenagem e coleção de dados computacionalmente simples. O cálculo da probabilidade, por outro lado, é muito dispendioso já que se deve medir a distância para todo padrão de referência no conjunto de treinamento.

Apesar do sucesso obtido pela técnica DTW, métodos alternativos de reconhecimento têm sido estudados devido principalmente aos seguintes fatores [14]:

1. O alto custo computacional do método usando programação dinâmica;
2. As dificuldades em estender esse método para problemas mais difíceis como, por exemplo, o reconhecimento de palavras conectadas ou de voz contínua;
3. O desejo de usar-se um modelo paramétrico robusto, ao invés dos vetores de padrões não paramétricos, para a representação do sinal de voz;
4. O desejo de usar-se unidades de voz que não sejam as palavras, como por exemplo, sílabas ou fonemas.

Devido a uma ou mais das razões acima, vários métodos diferentes têm sido propostos, tais como o uso da quantização vetorial no cálculo da programação dinâmica;

o uso da quantização vetorial em substituição à própria programação dinâmica; o uso de um pré-processador *front-end* baseado na quantização vetorial; e o uso de Modelos de Markov Escondidos (HMM's) para representar o sinal de voz.

2.4.2 Modelos de Markov escondidos (HMM's)

Modelos de Markov escondidos, representam muito bem as propriedades da fala em uma estrutura probabilística. Usando algoritmos automáticos, o sistema estima as propriedades estatísticas dos eventos da fala representando-os através de modelos probabilísticos.

HMM's têm propriedades superiores de generalização, ou seja, podem representar vários eventos da voz (palavras, sílabas, fonemas, etc). Além disso, eles têm sido adequados para grandes vocabulários e fala contínua, como também para aplicações independentes do locutor. HMM é a tecnologia predominante na maioria dos sistemas de pesquisa [1].

Embora os reconhecedores baseados em quantização vetorial tenham obtido um desempenho muito bom no reconhecimento de palavras isoladas, e tenham reduzido significantemente os custos computacionais, eles têm contribuido muito pouco para diminuir as dificuldades na extenção dos métodos baseados em vetores de padrões para as aplicações do reconhecimento de voz contínua e do reconhecimento de palavras conectadas, com grandes vocabulários.

Dessa forma, o reconhecedor HMM tem sido de grande interesse devido ao seu baixo custo computacional na fase de classificação, e por possibilitar um modelamento de padrões da fala na forma probabilística, reduzindo assim, o volume de dados a ser armazenado para servir como padrão de referência de cada modelo [14].

A técnica de reconhecimento de fala baseada em modelos de Markov escondidos será exaustivamente tratada nos capítulos seguintes.

2.4.3 Sistemas de reconhecimento baseados em redes neuronais

Uma técnica mais recente que vem sendo também usada em sistemas para reconhecimento automático de fala baseia-se em Redes Neuronais [6].

Os sistemas baseados em Redes Neuronais, constituem uma promessa de uma nova tecnologia que codifica propriedades da fala em uma representação distribuída. Redes Neuronais têm muitas propriedades adequadas, tais como generalização e capacidade discriminativa. Embora não existam sistemas de reconhecimento de fala baseados em Redes Neuronais em larga escala, eles têm sido usados como componentes de sistemas de pesquisa [1].

2.5 O presente e o futuro da tecnologia de reconhecimento de fala

Embora o reconhecimento automático de voz falada fluentemente permaneça sem solução, uma grande quantidade de conhecimento fundamental tem sido obtido. Três avanços significativos no desenvolvimento de algoritmos ilustram o rápido progresso obtido nos últimos anos:

- Independência do locutor: reconhecer palavras sem treinamento específico para um dado locutor;
- Reconhecimento de fala contínua: entender sentenças faladas fluentemente ao invés de palavras isoladas;
- Localização de palavras: focalizar sobre palavras chaves contidas em uma sentença falada.

Em adição ao progresso na pesquisa de reconhecimento de fala, tem havido muitos avanços na tecnologia de Processadores Digitais de Sinais (PDS). Os PDS mais rápidos permitem que algoritmos avançados sejam implementados em tempo real e que se obtenha um *hardware* bastante eficiente para aplicações comerciais.

Com os recentes melhoramentos na precisão em reconhecimento de fala, espera-se que essa tecnologia encontre várias aplicações significativas dentro dos próximos anos. Exemplos de aplicações que estão em fase de desenvolvimento no mundo afora incluem: Transações bancárias via telefone, serviços de operadores e pedidos da bolsa de valores [17]. O crescimento dos sistemas de resposta vocal, podem naturalmente ser extendidos para permitirem entrada vocal ao invés de sinais digitados em um teclado.

A tecnologia de reconhecimento automático de fala tem avançado vertiginosamente nos últimos anos devido ao surgimento de novas técnicas para modelagem das seqüências de sons da fala, à possibilidade de caracterização de grande número de locutores e ao desenvolvimento de técnicas de computação mais avançadas que permitem tratar com palavras naturalmente conectadas [18].

O progresso tem, portanto, ido dos primeiros sistemas para reconhecimento de vetores de padrões dependentes do locutor usando umas poucas dezenas de palavras isoladas, para os sistemas emergentes que são independentes do locutor e capazes de tratar com palavras conectadas em uma conversação interativa.

Para alta confiabilidade, estes sistemas ainda são limitados a pequenos vocabulários (centenas de palavras), mas podem servir perfeitamente a uma grande população de usuários. Tipicamente, eles também usam gramáticas finitas que são projetadas para fins específicos. Entrada de voz é, portanto, restrita à costrução de sentenças permitidas neste sub-conjunto limitado da linguagem natural.

Pesquisas recentes utilizam vocabulários de várias centenas de palavras, baseadas em reconhecimento de unidades de comprimento de sub-palavras, para o reconhecimento automático de fala [19, 20]. A grande maioria dos modelos estatísticos baseados em sub-palavras usam técnicas de modelos de Markov escondidos para estimar as seqüências da palavra falada [18, 21].

Avanços substanciais em cálculos de alta velocidade, e econômicos, serão necessários

para suportar sistemas com grandes vocabulários e entrada de linguagem natural. Estes avanços podem ser realisticamente esperados para a próxima década [17].

A tecnologia de reconhecimento de fala está agora emergindo dos laborátorios de pesquisa de forma prática e robusta. Sua confiabilidade e capacidade têm sido estabelecidas através de novas pesquisas em independência do locutor e técnicas de reconhecimento de palavras chaves [18]. Essa tecnologia embora computacionalmente dispendiosa, é economicamente suportada pelos recentes avanços em microprocessadores. Consequentemente, um número significativo de novas aplicações comerciais para reconhecimento de fala estão rapidamente se desenvolvendo.

É importante ter em mente que reconhecimento de fala apesar dos avanços em confiabilidade, permanece propenso a erros. Por esta razão, os produtos e serviços que obterão maior sucesso serão aqueles que tenham as seguintes características: simplicidade, crescimento evolucionário - os primeiros sistemas serão extensão dos sistemas existentes, e tolerância a erros.

Olhando para um futuro mais distante, pode-se fazer algumas especulações sobre novas facilidades, baseadas em pesquisas futuras, que abrirão aplicações mais ambiciosas [17]:

- Unidades de sub-palavras: será possível construir um dicionário composto de modelos fonéticos, primeiro para vocabulários pequenos e facilmente distinguíveis, depois para grandes vocabulários. Assim, o esforço e o tempo gastos para reunir fala de muitos locutores para cada palavra do vocabulário será eliminado;

- Imunidade ao ruído: algoritmos de melhoramento de voz mais eficientes farão reconhecedores de fala mais precisos em ambientes ruídosos;

- Entendimento da linguagem: a habilidade para localizar palavras chaves em uma frase é o primeiro passo para o entendimento da essência de uma sentença, mesmo se algumas palavras não são reconhecidas;

- Adaptação ao locutor: pessoas podem adaptar-se rapidamente aos dialetos e sotaques na fala. As máquinas poderiam responder mais precisamente se elas tivessem uma capacidade de aprendizagem similar.

A meta final é conversar fluentemente com um computador.

Sistemas práticos para conversação com computadores, sobre tarefas elementares, são uma realidade emergente. Com o progresso da tecnologia de reconhecimento de fala ocorre, consequentemente, a evolução dos sistemas - simples no início mas tornando-se mais poderosos - que levarão as pessoas a se comunicarem com as máquinas, usando as facilidades inerentes à comunicação vocal.

A conversação interativa com máquinas e computadores, com fala conectada de maneira natural, independência do locutor, e alta confiabilidade, deverá permanecer restrito à aplicações de fins específicos. Isto implica em vocabulários da ordem de várias centenas de palavras, com modelos de gramática que são restritos a sub-conjuntos da linguagem natural. Apesar disso, o usuário capaz de entender e aceitar as limitações não humanas das máquinas, encontrará um sistema de informações extraordinariamente útil.

Vocabulários para reconhecimento de fala aumentarão de tamanho através da pesquisa de unidades de sub-palavras. Sistemas com mais de 1.000 palavras já estão em testes de laboratório, e uma meta de vocabulários de 10.000 palavras parece realística [18].

Implementações práticas da tecnologia de fala dependem do processamento digital de sinais. O futuro continua encorajador e a economia e capacidade de processadores de sinais continuarão a se expandir. Implementações práticas e econômicas de sistemas de reconhecimento com grandes vocabulários, dependerão diretamente deste progresso computacional. Com o avanço da microeletrônica e a disponibilidade de computação de baixos custos, a próxima década verá o rápido surgimento de sofisticados sistemas vocais interativos.

Capítulo 3

Modelos de Markov escondidos (HMM's)

3.1 Introdução

Embora inicialmente introduzidos e estudados no final dos anos 60 e início dos anos 70, os métodos estatísticos de fontes de Markov, ou Modelos de Markov Escondidos, tornaram-se mais populares nos últimos anos. Existem duas fortes razões para que isto tenha ocorrido. Primeiro, os modelos são muito ricos em estrutura matemática e podem assim formar a base teórica para o uso em uma grande faixa de aplicações. Segundo, os modelos quando aplicados adequadamente, trabalham muito bem na prática para várias aplicações importantes [22]. Este capítulo apresenta os aspectos teóricos deste tipo de modelagem estatística voltados para o reconhecimento automático de fala.

3.2 Modelos de sinais

Um problema de fundamental importância no estudo dos sinais presentes no mundo real é a caracterização desses sinais reais em termos de modelos. Uma classe de modelos de sinais de bastante interesse agrupa os modelos estatísticos, onde tenta-se caracterizar as propriedades estatísticas do sinal. Exemplos de tais modelos estatísticos incluem processos Gaussianos, processos de Poisson, processos de Markov, e processos de Markov escondidos, dentre outros. A hipótese levantada pelos modelos estatísticos é que o sinal pode ser bem caracterizado como um processo aleatório paramétrico, e que os parâmetros do processo estocástico podem ser determinados (estimados) de uma maneira precisa e bem definida [22].

Para aplicações na área de processamento de voz, os modelos de sinais estocásticos têm apresentado bons resultados. Neste trabalho, em particular, estuda-se os modelos de sinais baseados em funções probabilísticas de cadeias de Markov, ou ainda denominados, modelos de Markov escondidos (HMM - *Hidden Markov Models*) para o reconhecimento de dígitos isolados falados de maneira dependente e independente do locutor.

A teoria básica dos modelos de Markov escondidos foi publicada em uma série de artigos clássicos por Baum et al [23]-[27] no final dos anos 60 e início dos anos 70, e foi aplicada para o propósito do reconhecimento de voz por Baker [28], e independentemente, por Jelinek et al [29] nos anos 70. Contudo, foi apenas na década de 80 que o entendimento e a aplicação dos HMM's para o processamento de voz tornaram-se comuns [22].

3.3 Processos discretos de Markov

Considere-se um sistema que pode ser descrito em qualquer tempo como estando em um dos seus N estados distintos, $S_1, S_2, \ldots, S_N$. Em tempos discretos, regularmente espaçados, o sistema realiza uma mudança de estado de acordo com um conjunto de probabilidades associado ao estado. Denotando-se os instantes discretos de tempo

associados com as mudanças de estado como t = 1, 2, ..., T, o estado atual no tempo t será representado por q_t. Uma descrição probabilística de um sistema será obtida a partir de uma cadeia de Markov de primeira ordem a qual requer apenas a informação a respeito do estado atual e do estado imediatamente anterior. Analiticamente, tem-se:

$$P[q_t = S_j / q_{t-1} = S_i, q_{t-2} = S_k, ...] = P[q_t = S_j / q_{t-1} = S_i] \tag{3.1}$$

$$1 \leq i, j, k \leq N \qquad 1 \leq t \leq T$$

O processo estocástico descrito acima poderia ser chamado de modelo de Markov observável, uma vez que a saída do processo é o conjunto de estados em cada instante de tempo, onde cada estado corresponde a um evento físico (observável) [22].

Embora se possa considerar modelos de Markov nos quais cada estado corresponda a um evento físico observável, este tipo de modelo é muito restrito para ser aplicado a muitos problemas de intesse prático [22]. Como conseqüência imediata deste fato, surge a extensão do conceito dos modelos de Markov para incluir os casos onde a observação é uma função probabilística do estado. O modelo resultante (o qual é chamado de modelo de Markov escondido) é um processo estocástico duplo, ou seja, existe um processo que não é observável (escondido), mas que apenas pode ser observado através de um outro conjunto de processos estocásticos que produzem a seqüência de observações.

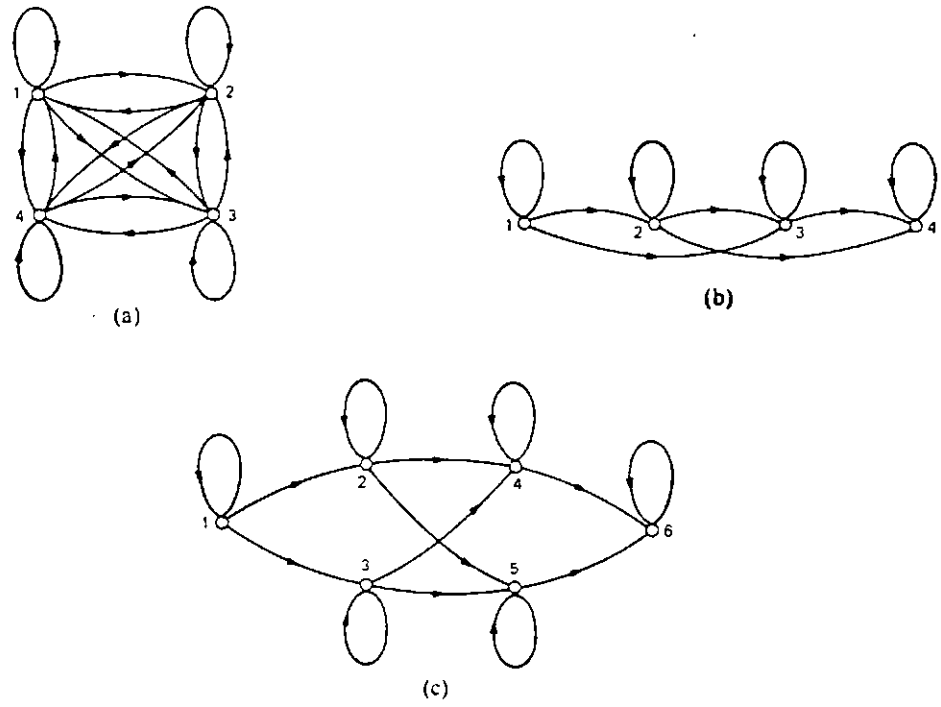
3.4 Tipos de HMM's

Para o caso do reconhecimento de fala utilizando modelos de Markov escondidos, a primeira questão envolvida na determinação dos HMM's ótimos[1] para cada palavra do vocabulário é a estrutura do modelo.

Entre as diversas variações e combinações possíveis para representar a forma dos HMM's, é comum considerar-se três tipos de estruturas para os modelos, ou sejam: o modelo sem restrição (ergódico), o modelo serial restrito, e o paralelo restrito [16].

[1]HMM ótimo é o conjunto de parâmetros de um HMM que melhor representa uma determinada palavra do vocabulário.

Exemplos de cada um destes modelos são mostrados na Figura 3.1 dada a seguir. Para os modelos sem restrição (Fig. 3.1a) pode ocorrer uma transição de um estado para qualquer outro. Os modelos do tipo serial restrito (Fig. 3.1b) e paralelo restrito (Fig. 3.1c) são modelos *left-right* ou modelos de Bakis [30]. Neste tipo de modelo, uma transição do estado q_i para o estado q_j é possível apenas se $j \geq i$.



Fonte: Rabiner (1983) [16], pag.1081.

Figura 3.1: Diagramas de estados para cadeias de Markov. (a) Modelo de Markov sem restrição com quatro estados, (b) Modelo de Markov serial restrito com quatro estados, e (c) Modelo de Markov paralelo restrito com seis estados.

Os modelos seriais geralmente seguem seqüencialmente através dos estados (embora estados individuais possam ser evitados). Por outro lado, os modelos paralelos permitem caminhos múltiplos, com cada caminho pulando um ou mais estados.

Os modelos seriais restritos apresentam melhor desempenho do que aqueles sem

restrição para o caso do reconhecimento de palavras isoladas. Este resultado advém da própria estrutura da fala ser inerentemente seqüencial, a liberdade adicional de transição de estados presente nos modelos sem restrição não refletem as variações dos parâmetros da fala caracterizados por vetores de padrões. Foi verificado, portanto que o uso dos modelos sem restrição apresentam um menor desempenho quando comparados com os modelos restritos [16].

A idéia principal para o uso de uma estrutura paralela é que ela poderia modelar os efeitos do uso de modelos de palavras múltiplos da mesma forma que os vetores de padrões múltiplos são usados no reconhecedor LPC/DTW convencional. Contudo, experimentos mostraram que não existe qualquer vantagem real no uso de uma estrutura paralela [16]. Isto indica que um modelo equivalente de HMM's múltiplas não é obtido diretamente pela simples mudança da estrutura do modelo.

Para o propósito de reconhecimento de palavras isoladas, é comum considerar-se modelos do tipo *left-right* [30]. Estes modelos têm as seguintes propriedades:

1. A primeira observação é produzida quando a cadeia de Markov encontra-se em um estado bem determinado, chamado de estado inicial, e designado por q_1;
2. A última observação é gerada enquanto a cadeia de Markov está em um outro estado bem determinado, chamado de estado final ou estado de absorção, e designado por q_N;
3. Uma vez que em uma cadeia de Markov se deixa um estado, aquele estado não pode ser mais visitado num tempo posterior.

3.5 A matriz transição de estados

O conjunto de probabilidades que rege a transição de estados no modelo HMM é representado pela matriz transição de estados, e simbolizado por $\mathbf{A} = \{a_{ij}\}$. Onde a_{ij} é a probabilidade de ocorrer uma transição do estado i para o estado j.

Considerando-se apenas os processos para os quais o lado direito da Equação (3.1) é independente do tempo, deriva-se um conjunto de probabilidades de transição de estados a_{ij} da forma:

$$a_{ij} = P[q_t = S_j / q_{t-1} = S_i] \qquad 1 \leq i, j \leq N \tag{3.2}$$

com os coeficientes de transição de estados obdecendo às seguintes restrições estocásticas [22]:

$$a_{ij} \geq 0 \tag{3.3}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{3.4}$$

Para o caso especial onde qualquer estado pode ser alcançado a partir de qualquer outro em um único passo (modelo ergódico), tem-se $a_{ij} > 0$ para todo i, j. Para outros tipos de HMM's (*lef-right*, por exemplo), tem-se $a_{ij} = 0$ para um ou mais pares (i, j).

A propriedade fundamental de todos os modelos de HMM's do tipo *left-right* é que os coeficientes a_{ij} da matriz transição de estados, $\mathbf{A}$, obedecem à propriedade:

$$a_{ij} = 0, \quad \text{para } j < i \qquad 1 \leq i, j \leq N \tag{3.5}$$

Para modelos do tipo *left-right*, restrições adicionais são frequentemente colocadas sobre os coeficientes da matriz de transição de estados, para assegurar que não ocorram mudanças muito grandes nos índices dos estados; assim, uma restrição da forma:

$$a_{ij} = 0, \qquad j > i + \Delta \qquad \Delta = 1, 2, \ldots \tag{3.6}$$

onde Δ é um incremento no índice do estado, é frequentemente utilizada [22].

Além disso, as probabilidades de estado inicial apresentam a propriedade:

$$\pi_i = \begin{cases} 0, & \text{para i} \neq 1 \\ 1, & \text{para i} = 1 \end{cases} \tag{3.7}$$

3.6 O número de estados do modelo

No caso do reconhecimento de fala, um outro fator que afeta a determinação dos HMM's ótimos para cada palavra do vocabulário é o número de estados do modelo [16]. O número de estados é representado pela letra **N**.

Cada uma das estruturas mostradas anteriormente na Figura 3.1 pode ser generalizada para incluir um número arbitrário de estados. Contudo, o número de parâmetros livres do modelo de Markov é da ordem de $\mathbf{N}^2$ para a matriz **A**. Assim, se **N** torna-se muito grande, as dificuldades para a determinação precisa e confiável dos valores ótimos de **A**, como também dos demais parâmetors do HMM, pode se tornar difícil para um conjunto de treinamento de tamanho limitado. Apesar disso, dentro dessas restrições tem-se investigado modelos com apenas dois estados, como também modelos com até 20 estados [16]. Os resultados parecem indicar a não existência de uma maneira teórica ótima para a escolha do número de estados necessários para um modelo da palavra, desde que os estados não estão necessariamente fisicamente relacionados com qualquer fenômeno simples observável [16].

De uma maneira bastante simplificada, e para sinais de voz derivados de um pequeno vocabulário de palavras isoladas, pode-se imaginar com razoável grau de segurança, o trato vocal como estando em uma de suas várias configurações articulatórias ou estados num determinado instante de tempo. Em cada estado um sinal de curta-duração (no tempo) é produzido, o qual tem um número finito de configurações espectrais. Assim, os espectros de potência de curtos intervalos de tempo do sinal de voz são determinados apenas pelo estado atual do modelo, enquanto que a variação da composição espectral do sinal ao longo do tempo é governada predominantemente por uma lei de transição de estados probabilística de uma cadeia de Markov [30].

Na realidade, existe uma interação altamente complexa entre a taxa de erros e o tamanho do modelo. Assim não pode ser afirmado, por exemplo, que palavras com duas sílabas necessitam de mais estados em seus modelos que palavras monossílabas. Também é certo que não existe qualquer relação simples entre a acuracidade da palavra,

número de sons (sílabas, etc.) da palavra, e o número de estados necessários no HMM da palavra [16].

3.7 A função densidade de probabilidade das observações

A partir da distribuição de probabilidades para o estado inicial, representada por $\Pi = \{\pi_i\}$, é possível se determinar qual o estado do modelo é o responsável pelo início da cadeia de Markov. Analiticamente tem-se:

$$\pi_i = P[q_1 = S_i] \qquad 1 \leq i \leq N \tag{3.8}$$

Para o caso do reconhecimento de fala utilizando HMM's, em um dado estado, q_j, a saída observada do modelo é um vetor aleatório com uma função densidade de probabilidade (fdp) b_j [31].

Assumindo-se que o sinal a ser representado pelo HMM consiste de uma seqü ência de T vetores de observações $\mathbf{O} = \{O_1, O_2, \ldots, O_T\}$, onde cada i-ésimo vetor O_i caracteriza o sinal no tempo t = i, então podem ser considerados dois tipos de funções densidade de probabilidade de observações, ou seja, discreta e contínua [16].

Para o tipo discreto, cada vetor O_i é trocado por um dos M possíveis símbolos $v_k \in V, 1 \leq k \leq M$, onde V representa um alfabeto discreto obtido através de algum tipo de quantização vetorial tal que a *distorção na quantização*[2] de O_i seja mínima. Seja q_j o estado no tempo t, então $B = [b_{jk}]$, $1 \leq j \leq N$, é a probabilidade de observação do k-ésimo símbolo v_k no estado q_j [16].

O problema resultante do uso da quantização vetorial, pelo menos para algumas aplicações, é que as observações são sinais contínuos (ou vetores contínuos). Embora seja possível quantizar esses sinais contínuos através de dicionários obtidos da

[2]Erro inerente ao processo de quantização vetorial.

quantização vetorial, sempre existe degradação associada com tal quantização. Assim seria vantajoso o uso de HMM's com funções densidade de probabilidade contínuas [22].

No caso contínuo, tem-se a função densidade de probabilidade B = $b_j(\mathbf{x})$, 1≤j≤N, onde $\mathbf{x}$ é um vetor de observação de dimensão D e $b_j(\mathbf{x})d\mathbf{x}$ é a probabilidade de observação do vetor O_t entre $\mathbf{x}$ e $\mathbf{x} + d\mathbf{x}$ no estado q_j. Os tipos de funções densidades permitidas para $b_j(\mathbf{x})$, para os quais existe um algoritmo de reestimação, incluem densidades log-côncavas, densidades elipticamente simétricas e misturas de densidades log-côncavas ou elipticamente simétricas [14, 32].

3.8 HMM's com densidades contínuas

A representação mais geral para a fdp contínua, para a qual existe um procedimento de reestimação, é uma mistura finita da forma:

$$b_j(\mathbf{x}) = \sum_{k=1}^{M} c_{jk}\rho(\mathbf{x}, \vec{\mu}_{jk}, \mathbf{U}_{jk}) \qquad 1 \leq j \leq N \tag{3.9}$$

onde $\mathbf{x}$ é o vetor que está sendo modelado e $\rho(\mathbf{x}, \vec{\mu}, \mathbf{U})$ representa uma função densidade de probabilidade D-dimensional, geralmente Gaussiana, de vetor média $\vec{\mu}_{jk}$ e matriz covariância $\mathbf{U}_{jk}$ para a k-ésima componente da mistura no estado q_j. A representação da Equação(3.9) pode ser usada para aproximar, com grau de precisão arbitrário, qualquer função densidade contínua finita [31].

Os coeficientes , c_{jk} , são os ganhos das componentes da mistura e satisfazem a seguinte restrição estocástica [31]:

$$\sum_{k=1}^{M} c_{jk} = 1 \qquad 1 \leq j \leq N \tag{3.10}$$
$$c_{jk} \geq 0, \qquad 1 \leq k \leq M$$

tal que

$$\int_{-\infty}^{+\infty} b_j(\mathbf{x})d\mathbf{x} = 1, \qquad 1 \leq j \leq N \tag{3.11}$$

Em resumo, a especificação completa de um HMM como uma mistura de densidades contínuas, exige a escolha de valores (e/ou a estimação dos parâmetros) para o seguinte [22]:

N - Número de estados do modelo;

M - Número de misturas;

D - Dimensão do vetor de observação;

$\mathbf{A} = \{a_{ij}\}$ $\quad 1 \leq i, j \leq N$ - Matriz de transição de estados;

$\mathbf{C} = \{c_{ik}\}$ $\quad 1 \leq k \leq M$ - Matriz ganho das misturas;

$\vec{\mu} = \{\mu_{jkd}\}$ $\quad 1 \leq d \leq D$ - Vetores médias das componentes da mistura;

$\mathbf{U} = \{U_{jkde}\}$ $\quad 1 \leq e \leq D$ - Matrizes covariâncias das componentes da mistura.

$\mathbf{B} = \{b_j(\mathbf{x})\}$ - Função densidade de probabilidade

$\Pi = \{\pi_i\}$ - Probabilidades para o estado inicial

É comum a utilização de uma notação compacta para representar os parâmetros de um HMM específico. Essa notação é dada por [22]:

$$\lambda = (\mathbf{A}, \mathbf{B}, \Pi) \tag{3.12}$$

O problema com o reconhecedor HMM de densidades contínuas é o cálculo de $b_j(\mathbf{x})$ que é extremamente dispendioso, especialmente para valores de $M > 1$ [14]. O custo computacional do reconhecedor HMM de densidades contínuas em relação ao reconhecedor DTW está em torno de um quarto para o caso de uma única fdp por estado do modelo. No entanto, estes custos chegam a ser equivalentes para o caso de cinco fdp's por estado [14]. Por outro lado, para o reconhecedor HMM de densidades discretas, o custo computacional é uma ordem de magnitude inferior ao exigido pelo reconhecedor DTW [16].

3.9 Os três problemas fundamentais dos HMM's e suas soluções

A idéia de caracterizar os aspectos teóricos dos modelos de Markov escondidos em termos da solução de três problemas fundamentais é atribuída a Jack Ferguson [22]. Esses três problemas são os seguintes :

1. O cálculo da probabilidade (ou verossimilhança) de uma seqüência de observações dado um HMM específico;
2. A determinação de uma seqüência ótima de estados do modelo;
3. O ajuste dos parâmetros do modelo de modo a que melhor representem o sinal que está sendo modelado.

Esses três problemas podem ser colocados de maneira mais explícita da seguinte forma:

O problema 1 é essencialmente uma questão de cálculo, ou seja, dado um modelo $\lambda = (\mathbf{A}, \mathbf{B}, \Pi)$ e uma seqüência de observações $\mathbf{O} = (O_1, O_2, \ldots, O_T)$, deseja-se calcular de maneira eficiente a probabilidade de que essa seqüência de observações (vetores paramétricos do sinal de voz) tenha sido produzida pelo modelo de referência (parâmetros do HMM).

No problema 2 tenta-se descobrir a parte escondida do modelo, isto é, encontrar a seqüência de estados que melhor represente essas observações.

O problema 3 trata da otimização dos parâmetros do modelo, tal que possam representar de maneira específica cada palavra do vocabulário de interesse.

3.9.1 Solução ao primeiro problema

Deseja-se calcular a probabilidade da seqüência de observações, $\mathbf{O} = (O_1, O_2, \ldots, O_T$, dado o modelo λ, ou seja, $P(\mathbf{O}/\lambda)$. A maneira mais direta de se resolver este problema é dada por [22]:

$$P(\mathbf{O}/\lambda) = \sum_{\forall Q} P(\mathbf{O}/Q, \lambda) P(Q/\lambda) \tag{3.13}$$

A equação dada acima, indica a necessidade de se calcular a probabilidade de todas as possíveis seqüências de estados de tamanho T (número total de observações). Considerando-se uma dessas possíveis seqüências de estados, tem-se:

$$Q = q_1, q_2, \ldots, q_T \tag{3.14}$$

onde q_1 é o estado inicial. A probabilidade da seqüência de observações **O** dada a seqüência de estados acima, é representada por:

$$P(\mathbf{O}/Q, \lambda) = \prod_{t=1}^{T} P(\mathbf{O}_t/q_t, \lambda) \tag{3.15}$$

Na expressão acima assume-se que as observações são estatisticamente independentes. Desse modo tem-se:

$$P(\mathbf{O}/Q, \lambda) = b_{q_1}(O_1) \cdot b_{q_2}(O_2) \ldots b_{q_T}(O_T) \tag{3.16}$$

A probabilidade de ocorrência da seqüência de estados Q dado o modelo λ pode ser escrita como:

$$P(Q/\lambda) = \pi_{q_1} \cdot a_{q_1 q_2} \cdot a_{q_2 q_3} \ldots a_{q_{T-1} q_T} \tag{3.17}$$

A probabilidade conjunta de **O** e Q, isto é, a probabilidade de que **O** e Q ocorram conjuntamente, é simplesmente o produto dos dois termos acima, ou seja:

$$P(\mathbf{O}, Q/\lambda) = P(\mathbf{O}/Q, \lambda)P(Q, \lambda) \tag{3.18}$$

A probabilidade de **O** (dado o modelo) é obtida somando esta probabilidade conjunta sobre todas as seqüências de estados possíveis:

$$\begin{aligned} P(\mathbf{O}/\lambda) &= \sum_{\forall Q} \pi_{q_1} \cdot b_{q_1}(O_1) \cdot a_{q_1 q_2} \cdot b_{q_2}(O_2) \ldots \\ &\quad \ldots a_{q_{T-1} q_T} \cdot b_{q_T}(O_T) \end{aligned} \tag{3.19}$$

Inicialmente (no tempo t = 1) tem-se o estado q_1 com probabilidade π_{q_1}, que produz a observação O_1 (neste estado) com probabilidade $b_{q_1}(O_1)$. Para a transição do instante t para o instante t + 1 (t = 2) tem-se, então, a transição do estado q_1 para o estado q_2 com probabilidade $a_{q_1 q_2}$, e a geração da observação O_2 com probabilidade $b_{q_2}(O_2)$. O processo continua sucessivamente até que a última transição seja realizada (no tempo T) do estado q_{T-1} para o estado q_T com probabilidade $a_{q_{T-1} q_T}$ e produz a observação O_T com probabilidade $b_{q_T}(O_T)$.

O cálculo de $P(\mathbf{O}/\lambda)$ de acordo com a Equação (3.19) envolve cerca de $2TN^T$ operações (adições e multiplicações) o que torna este método proibitivo [22].

Um método prático para resolver este problema é conhecido como o Algoritmo Progressivo-Regressivo (*Forward-Backward Procedure*) de Baum-Welch [27].

O algoritmo progressivo-regressivo baseia-se no emprego de duas variáveis auxiliares, a variável progressiva $\alpha_t(i)$ e a variável regressiva $\beta_t(i)$.

A variável progressiva (*forward*), $\alpha_t(i)$, é definida como [22]:

$$\alpha_t(i) = P(O_1 O_2 \ldots O_t, q_t = S_i/\lambda) \tag{3.20}$$

isto é, a probabilidade de ocorrência da seqüência de observações parcial, O_1, O_2, ..., O_t, (até o tempo t) e estado S_i no tempo t, dado o modelo λ.

Para a solução do problema 1 em questão necessita-se apenas da parte progressiva do algoritmo. O cálculo de P($\mathbf{O}/\ \lambda$) é realizado recursivamente da seguinte forma:

1. Inicialização:

$$\alpha_1(i) = \pi_i b_i(O_1) \qquad 1 \leq i \leq N \tag{3.21}$$

2. Recursividade:

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] b_j(O_{t+1}) \quad 1 \leq t \leq T-1 \quad 1 \leq j \leq N \tag{3.22}$$

3. Finalização:

$$P(\mathbf{O}/\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{3.23}$$

O cálculo de P($\mathbf{O}/\lambda$) através da Equação (3.23) requer cerca de N^2T operações (multiplicações e adições), o que é significamente inferior à quantidade $2TN^T$ exigida pela Equação (3.19) [22].

A variável regressiva (*backward*), β_t(i), não se faz necessária para a solução do problema 1, no entanto, ela será agora definida pois a mesma será utilizada na solução do problema 3.

A definição de β_t(i) é dada por:

$$\beta_t(i) = P(O_{t+1} O_{t+2} \ldots O_T / q_T = S_i, \lambda) \tag{3.24}$$

isto é, a probabilidade de ocorrência da seqüência de observações parcial do instante t + 1 até o instante T, dado o estado S_i no tempo t e o modelo λ. A solução da Equação (3.24) é obtida de forma recursiva, da seguinte maneira:

1. Inicialização:

$$\beta_T(i) = 1 \qquad 1 \leq i < N \tag{3.25}$$

2. Recursividade:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \quad t = T-1, T-2, \ldots, 1 \quad 1 \leq i \leq N \tag{3.26}$$

O cálculo de β_t(i) requer cerca de N^2T operações (adições e multiplicações) [22].

3.9.2 Solução ao segundo problema

Diferentemente do que acontece no problema 1, para o qual existe uma solução exata, existem várias maneiras possíveis de se resolver o problema 2. A dificuldade de se encontrar a seqüência de estados ótima, associada com uma dada seqüência de observações, está exatamente na definição do que seja essa seqüência ótima, isto é, existem vários critérios de otimização.

Um possível critério de otimização está baseado na escolha dos estados q_t que são individualmente mais prováveis. Esse critério de otimização maximiza o número esperado de estados individualmente corretos [22]. A implementação dessa solução ao problema 2, está baseada em termos das variáveis progressiva e regressiva, $\alpha_t(i)$ e $\beta_t(i)$, anteriormente definidas.

Embora esse critério maximize o número esperado de estados corretos (escolhendo o estado mais provável para cada t), poderão existir alguns problemas com a seqüência de estados resultante. Por exemplo, quando o modelo HMM tem probabilidade de transição de estados igual a zero ,$a_{ij} = 0$ para algum par (i, j), a seqüência de estados "ótima" poderá até mesmo não ser válida. Isto deve-se ao fato de que a solução desse problema através desse critério simplesmente determina o estado mais provável em cada instante de tempo, sem se importar com a probabilidade de ocorrência da seqüência de estados.

Uma possível solução para o problema descrito acima é modificar o critério de otimização. Por exemplo, pode-se resolvê-lo considerando-se a seqüência de estados que maximize o número esperado de pares de estados corretos (q_t, q_{t+1}) ou triplas de estados (q_t, q_{t+1}, q_{t+2}) etc. Embora esses critérios sejam possíveis para algumas aplicações, o critério mais largamente utilizado é o de se encontrar a seqüência de estados ótima de maneira única (caminho ótimo), isto é, maximizar $P(Q/\mathbf{O}, \lambda)$ que é equivalente a se maximizar $P(Q, \mathbf{O}/ \lambda)$ [22].

Existe uma técnica formal para se encontrar esta seqüência de estados ótima única, baseada em métodos de programação dinâmica, ela é chamada de Algoritmo de Viterbi [33].

3.9.2.1 O algoritmo de Viterbi

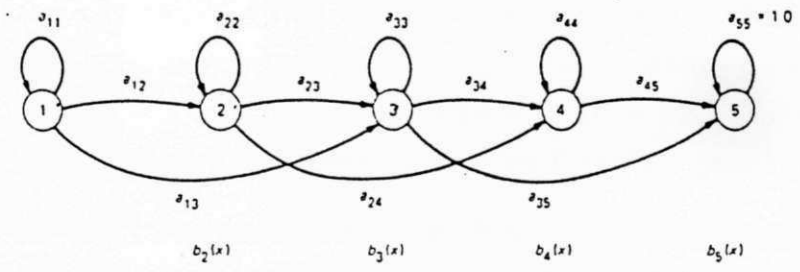
O algoritmo de Viterbi é uma solução ótima recursiva ao problema de estimar a seqüência de estados de um processo de Markov discreto no tempo [34].

O algoritmo de Viterbi foi proposto em 1967 como um método de decodificação de códigos convolucionais [33]. Desde então ele tem sido reconhecido como uma solução atrativa para uma variedade de problemas de estimação digital.

Em sua forma mais geral, o algoritmo de Viterbi pode ser visto como uma solução ao problema de maximizar a estimação da probabilidade *a posteriori* (probabilidade condicional) da seqüência de estados de um processo de Markov discreto no tempo [34]. Em outras palavras, dada uma seqüência de observações de um processo de Markov discreto no tempo, $\mathbf{O} = \{O_1, O_2, \ldots, O_T\}$, o algoritmo de Viterbi fornece a seqüência de estados, $Q = \{q_i, q_j, \ldots\}$, para a qual a probabilidade *a posteriori*, $P(Q/\mathbf{O})$, seja máxima.

O problema da estimação da seqüência de estados ótima é essencialmente igual ao problema de se encontrar o caminho mais curto, ou mais provável, através de um certo grafo [34]. O algoritmo de Viterbi aparece como um meio natural para a solução desse problema.

Considerando-se a cadeia de Markov do tipo *left-to-right* com 5 estados, mostrada na Figura 3.2 dada a seguir, o algoritmo de Viterbi é melhor entendido associando-se a essa cadeia de Markov uma descrição mais redundante chamada de treliça (Fig. 3.3).



Fonte: Rabiner (1985) [14], pag.1213.

Figura 3.2: Cadeia de Markov do tipo *left-to-right* com 5 estados.

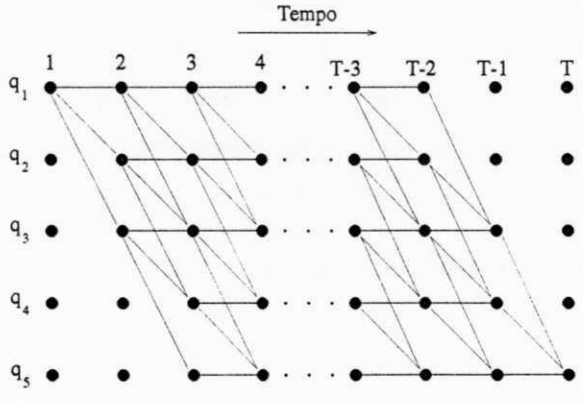


Figura 3.3: Estrutura de treliça associada à cadeia de Markov da Figura 3.2.

Na estrutura de treliça cada nó corresponde a um estado distinto da cadeia de Markov em um dado instante de tempo e cada "braço" representa uma transição para um novo estado no instante de tempo imediatamente posterior. A treliça começa e termina em estados bem definidos, estados inicial e final da cadeia de Markov, respectivamente. A propriedade mais importante, inerente a essa estrutura, é que para cada seqüência de estados possível Q corresponde um único caminho através da treliça e vice-versa [34].

Dada uma seqüência de observações **O**, a cada caminho pode ser associada uma distância proporcional a - log P(Q, **O**), onde Q é a seqüência de estados associada com esse caminho. Logo, deve-se encontrar a seqüência de estados para a qual P(Q/**O**) seja máxima ou, equivalentemente, para a qual P(Q, **O**) = P(Q/**O**)P(**O**) seja máxima, encontrando-se o caminho cujo comprimento - log P(Q, **O**) seja mínimo.

Observando a treliça da Figura 3.3, pode-se notar que para vários instantes de tempo diferentes, existe mais de um caminho parcial chegando em cada nó (estado), cada um com determinado comprimento (valor de probabilidade). O segmento de caminho mais curto ou seja, aquele que apresenta maior valor de probabilidade, é chamado de "sobrevivente" correspondente a cada nó. Em outras palavras, para cada instante de tempo existe um número de sobreviventes igual ao número de nós na treliça.

No último instante de tempo deve existir apenas um único sobrevivente, pois a cadeia de Markov deve terminar em um estado bem determinado. Nesse ponto, o caminho total (de t = 1 até t = T) representa o menor caminho percorrido, ou seja, apresenta o maior valor de probabilidade. Percorrendo de volta a seqüência de estados desse caminho, determina-se a seqüência de estados associada que fornece o caminho mais provável, ou seja, a seqüência de estados ótima.

Definindo-se a variável δ_t(i) como o maior valor de probabilidade ao longo de um único caminho até o instante de tempo t ou seja, considerando-se as t primeiras observações que terminam no estado S_i, tem-se por indução que:

$$\delta_{t+1}(j) = [\underbrace{\max}_{i} \delta_t(i)a_{ij}]b_j(O_{t+1}) \qquad 1 \leq i \leq N \tag{3.27}$$

Para se obter a seqüência de estados, é necessário reter a trilha do argumento que maximiza a Equação (3.27), para cada t e j. Para fazer isto define-se a variável Ψ_t(j). O procedimento completo para se encontrar a seqüência de estados ótima é dado por [22]:

1. Inicialização

$$\delta_1(i) = \pi_i b_i(O_1) \qquad 1 \leq i \leq N \tag{3.28}$$

$$\Psi_1(i) = 0 \tag{3.29}$$

2. Recursividade

$$\delta_t(j) = \underbrace{\max}_{1\leq i\leq N}[\delta_{t-1}(i)a_{ij}]b_j(O_t) \qquad 2 \leq t \leq T \qquad 1 \leq j \leq N \tag{3.30}$$

$$\Psi_t(j) = \underbrace{\text{argmax}}_{1\leq i\leq N}[\delta_{t-1}(i)a_{ij}] \qquad 2 \leq t \leq T \qquad 1 \leq j \leq N \tag{3.31}$$

3. Término

$$P^* = \underbrace{\max}_{1\leq i\leq N}[\delta_T(i)] \tag{3.32}$$

(3.33)

$$q_T^* = \underbrace{\text{argmax}}_{1\leq i\leq N}[\delta_T(i)] \tag{3.34}$$

4. Seqüência de estados ótima

$$q_t^* = \Psi_{t+1}(q_{t+1}^*) \qquad t = T-1, T-2, \ldots, 1 \tag{3.35}$$

Pode-se notar que o algoritmo de Viterbi é similar (exceto na determinação da seqüência de estados ótima) ao procedimento *forward* das Equações (3.20)-(3.22). A maior diferença é a maximização na Equação (3.30) sobre os estados anteriores que é usada em lugar do somatório da Equação (3.23).

3.9.3 Solução ao terceiro problema

O terceiro, e mais difícil, problema dos HMM's é determinar um método para ajustar os parâmetros do modelo $\lambda = (\mathbf{A}, \mathbf{B}, \Pi)$ de modo a maximizar a probabilidade da seqüência de observações, dado o modelo.

Infelizmente não existe qualquer maneira analítica para encontrar o modelo que maximiza a probabilidade da seqüência de observações. Na realidade, dada qualquer seqüência de observações finita como dados de treinamento, não existe qualquer maneira ótima de se estimar os parâmetros do modelo. É possível, no entanto, escolher-se λ tal que $P(\mathbf{O}/\ \lambda)$ seja localmente maximizada usando um procedimento iterativo como, por exemplo, o método de Baum-Welch [22].

A fim de descrever o procedimento para reestimação dos parâmetros dos HMM's, é necessário a definição da variável ξ_t(i, j), que representa a probabilidade da cadeia de Markov estar no estado S_i no tempo t, e no estado S_j no tempo t+1, dado o modelo e a seqüência de observações. Isto é,

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j/O, \lambda) \qquad 1 \leq i,j \leq N \qquad 1 \leq t \leq T \tag{3.36}$$

A partir da definição das variáveis progressiva, Eq.(3.20), e regressiva, Eq.(3.24), pode-se escrever ξ_t(i, j) na forma

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O/\lambda)}$$

ou,

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \tag{3.37}$$

onde o numerador é P(q_t = S_i, q_{t+1} = S_j/O, λ) e a divisão por P(**O**/ λ) fornece a medida de probabilidade desejada.

A probabilidade da cadeia de Markov estar no estado S_i no tempo t, dada a seqüência de observações e o modelo, é representada pela variável γ_t(i) a qual está relacionada à ξ(i, j) da seguinte forma:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{3.38}$$

Somando-se γ_t(i) ao longo do tempo t, tem-se uma quantidade que pode ser interpretada como o número esperado de vezes que o estado S_i é visitado ou, equivalentemente, o número esperado de transições feitas do estado S_i. Da mesma forma, o somatório de ξ_t(i, j) sobre t (de t = 1 até t = T-1) pode ser interpretado como o número esperado de transições do estado S_i para o estado S_j. Isto é,

$$\sum_{t=1}^{T-1} \gamma_t(i) = \text{número esperado de transições a partir de } S_i \tag{3.39}$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \text{número esperado de transições de } S_i \text{ para } S_j \tag{3.40}$$

A partir das fórmulas acima um método para a reestimação dos parâmetros de um HMM é obtido [22]. O conjunto de fórmulas de reestimação para Π, $\mathbf{A}$, e $\mathbf{B}$ é dado por:

$\overline{\pi_i}$ = freqüência esperada (número de vezes) no estado S_i no tempo (t=1), ou seja:

$$\overline{\pi_i} = \gamma_1(i) \tag{3.41}$$

$$\overline{a_{ij}} = \frac{\text{número esperado de trans. do estado } S_i \text{ para o estado } S_j}{\text{número esperado de transições do estado } S_i}$$

ou seja:

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{3.42}$$

As fórmulas de reestimação para os coeficientes da fdp de cada componente da mistura, isto é, c_{jk}, $\vec{\mu}_{jk}$, e $\mathbf{U}_{jk}$, são da forma:

$$\overline{c_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{M} \gamma_t(j,k)} \tag{3.43}$$

$$\overline{\mu_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) O_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.44}$$

$$\overline{\mathbf{U}_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)(O_t - \mu_{jk})(O_t - \mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.45}$$

onde o apóstrofo indica a matriz transposta do vetor e γ_t(j,k) é a probabilidade do vetor O_t está no estado q_j no tempo t com a k-ésima componente da mistura, isto é,

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}\right]\left[\frac{c_{jk} N(O_t, \vec{\mu}_{jk}, \mathbf{U}_{jk})}{\sum_{m=1}^{M} c_{jm} N(O_t, \vec{\mu}_{jk}, \mathbf{U}_{jm})}\right] \tag{3.46}$$

A interpretação das Equações (3.43)-(3.45) é bastante simples. A fórmula de reestimação para c_{jk} é a relação entre o número esperado de vezes que o sistema está no estado q_j usando a k-ésima componente da mistura e o número esperado de vezes que o sistema está no estado q_j. Similarmente, a fórmula de reestimação para o vetor média $\vec{\mu}_{jk}$ pondera cada termo do numerador da Equação (3.43) pelo vetor de observação, resultando no valor esperado da porção do vetor de observação presente na k-ésima componente da mistura. Uma interpretação similar pode ser dada para o termo de reestimação para a matriz covariância $\mathbf{U}_{jk}$.

Definindo-se o modelo inicial como $\lambda = (\mathbf{A}, \mathbf{B}, \Pi)$, e utilizando as fórmulas de reestimação dadas acima, obt ém-se um modelo reestimado $\overline{\lambda} = (\overline{\mathbf{A}}, \overline{\mathbf{B}}, \overline{\Pi})$ com as seguintes propriedades: 1) O modelo inicial λ define um ponto critico da função de verossimilhança, no caso de $\overline{\lambda} = \lambda$; ou 2) O modelo λ é mais representativo do que o modelo λ, onde tem-se $P(O/\ \overline{\lambda}) > P(O/\ \lambda)$, ou seja, encontra-se um novo modelo $\overline{\lambda}$ para o qual a seqüência de observações apresenta maior probabilidade de ter sido gerada.

Com a aplicação sucessiva do algoritmo de reestimação, com $\overline{\lambda}$ no lugar de λ, tem-se um aumento no valor da probabilidade de geração da seqüência de observações até que seja atingido um ponto limite. O resultado final desse procedimento de reestimação é uma estimativa de máxima verossimilhança do HMM. Vale a pena ressaltar que o algoritmo *forward-backward* leva a um máximo local embora se possa ter vários máximos locais [22].

Um aspecto importante do procedimento de reestimação é que as restrições estocásticas dos parâmetros do HMM são automaticamente satisfeitas em cada iteração, ou seja:

$$\sum_{i=1}^{N} \overline{\pi}_i = 1 \tag{3.47}$$

$$\sum_{j=1}^{N} \overline{a}_{jk} = 1 \qquad 1 \leq i \leq N \tag{3.48}$$

$$\sum_{k=1}^{M} \overline{c}_{jk} = 1 \qquad 1 \leq j \leq N \tag{3.49}$$

3.10 Seqüências de observações múltiplas

O maior problema associado ao modelo de HMM do tipo *left-right*, é que não se pode usar uma única seqüência de observações para treinar o modelo (isto é, para a reestimação dos parâmetros do modelo) [22]. Isto se deve à natureza transitória dos estados dentro do modelo, permitindo apenas um pequeno número de observações para qualquer estado (até que uma transição seja feita para um estado sucessor). Assim, a fim de se obter dados suficientes para se fazer estimativas confiáveis de todos os parâmetros do modelo, deve-se usar seqüências de observações múltiplas.

A modificação do procedimento de reestimação é direto e é mostrado a seguir.

Seja o conjunto de Q seqüências de observações representado por

$$\mathbf{O} = [\mathbf{O}^1, \mathbf{O}^2, \ldots, \mathbf{O}^Q] \tag{3.50}$$

onde $\mathbf{O}^q = [O_1^q O_2^q ... O_{T_q}^q]$ é a q-ésima seqüência de observação.

Assume-se que as seqüências de observações são independentes e o objetivo é o ajuste dos parâmetros do modelo λ que maximizam a expressão:

$$P(O/\lambda) = \prod_{q=1}^{Q} P(\mathbf{O}^q/\lambda) = \prod_{q=1}^{Q} P_q \tag{3.51}$$

Uma vez que as fórmulas de reestimação são baseadas em freqüências de ocorrências de eventos, as fórmulas de reestimação para as seqüências de observações múltiplas são modificadas adicionando-se as freqüências de ocorrências individuais de cada seqüência. Assim, as fórmulas de reestimação modificadas para os elementos da matriz de transição, a_{ij}, e para os coeficientes das fdps das componentes da mistura, isto é, c_{jk}, $\vec{\mu}_{jk}$, e $\mathbf{U}_{jk}$, são dadas por:

$$\overline{a}_{ij} = \frac{\sum_{q=1}^{Q}\sum_{t=1}^{T-1}\xi_t^q(i,j)}{\sum_{q=1}^{Q}\sum_{t=1}^{T-1}\gamma_t^q(i)} \tag{3.52}$$

$$\overline{c}_{jk} = \frac{\sum_{q=1}^{Q}\sum_{t=1}^{T}\gamma_t^q(j,k)}{\sum_{q=1}^{Q}\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t^q(j,k)} \tag{3.53}$$

$$\overline{\mu}_{jk} = \frac{\sum_{q=1}^{Q}\sum_{t=1}^{T}\gamma_t^q(j,k)O_t}{\sum_{q=1}^{Q}\sum_{q=1}^{Q}\sum_{t=1}^{T}\gamma_t^q(j,k)} \tag{3.54}$$

$$\overline{\mathbf{U}}_{jk} = \frac{\sum_{q=1}^{Q}\sum_{t=1}^{T}\gamma_t(j,k)(O_t-\vec{\mu}_{jk})(O_t-\mu_{jk})'}{\sum_{q=1}^{Q}\sum_{t=1}^{T}\gamma_t(j,k)} \tag{3.55}$$

3.11 Estimativas iniciais para os parâmetros dos HMM's

Teoricamente, as equações de reestimação devem fornecer valores para os parâmetros dos HMM's que correspondam a um máximo local da função de verossimilhança. Um ponto fundamental é, portanto, a escolha das estimativas iniciais desses parâmetros tal que o máximo local seja igual ao máximo global da função de verossimilhança.

Basicamente, não existe qualquer maneira simples nem direta de resolver este problema. Por outro lado, experiências mostram que estimativas iniciais uniformes ou mesmo aleatórias (sujeitas às restrições estocásticas) para os valores de Π e $\mathbf{A}$ apresentam, na maioria dos casos, resultados satisfatórios no processo de reestimação desses parâmetros. Contudo, para os parâmetros de $\mathbf{B}$, experiências têm mostrado que estimativas iniciais "ótimas" são essenciais quando se trata de misturas múltiplas no caso da distribuição contínua [16]. Essas estimativas iniciais podem ser obtidas de várias maneiras, incluindo a segmentação manual da seqüência de observações em estados e o posterior cálculo de médias a partir das observações dentro de cada estado, a segmentação de máxima verossimilhança, o cálculo de médias, e a segmentação "*k-means*" segmental com técnicas de agrupamento ("clustering") [22].

3.12 Limitações dos HMM's

Embora o uso de modelos de HMM tenha contribuído bastante para os recentes avanços em reconhecimento de voz, existem algumas limitações inerentes a estes modelos estatísticos para a voz. A maior limitação é a hipótese que observações sucessivas (blocos do sinal de voz) sejam independentes, e portanto a probabilidade de ocorrência de uma seqüência de observações pode ser escrita como um produto de probabilidades de observações individuais.

Outra limitação é a hipótese de que as distribuições dos parâmetros de observação individuais podem ser bem representadas como uma mistura de densidades Gaussianas. Finalmente a própria hipótese de Markov, ou seja, que a probabilidade da cadeia de Markov estar em um dado estado no tempo t, depende apenas do estado no tempo t-1, é claramente imprópria para sons vocais, onde as dependências frequentemente extendem-se através de vários estados. Contudo, apesar dessas limitações, este tipo de modelagem estatística tem sido usada com sucesso para certos tipos de problemas de reconhecimento de voz [22].

Capítulo 4

Reconhecimento de palavras isoladas utilizando HMM's com densidades contínuas

4.1 Introdução

Trabalhos realizados na IBM [35], e mais recentemente na Phillips [36], têm usado HMM's de densidades contínuas para o reconhecimento de fala onde se assume que todos os parâmetros de interesse possuem distribuições Gaussianas [37]. Uma forma alternativa para o uso de HMM's é a combinação com a quantização vetorial onde os parâmetros de interesse (vetores dos coeficientes LPC) são transformados em um conjunto de observações discretas. Tem-se, então, os chamados HMM's de densidades discretas [16].

Neste trabalho, em particular, estuda-se como aplicar os HMM's de densidades contínuas em aplicações para o reconhecimento de dígitos isolados dependente e independente do locutor.

O reconhecimento de palavras isoladas usando HMM consiste em duas fases: treinamento e reconhecimento (ou classificação). Na fase de treinamento, um conjunto de observações é usado para obter-se um conjunto de padrões de referência, representados por modelos, um para cada palavra do vocabulário. Na fase de classificação, a probabilidade de ocorrência da seqüência de observação de teste é calculada para cada modelo de referência. A seqüência de teste é, então, classificada como a palavra cujo modelo de referência fornece o valor de probabilidade mais alto [38].

4.2 Fase de treinamento dos HMM's

Para cada palavra do vocabulário um HMM é concebido; ou seja, o conjunto de parâmetros que compõem os HMM's é estimado de um conjunto de dados de treinamento representando ocorrências múltiplas da palavra, produzidas por um mesmo locutor ou por um grande número de locutores.

Para o caso do reconhecimento de fala utilizando HMM's de densidades contínuas, teoricamente é possível escolher-se valores iniciais aleatórios para cada um dos parâmetros do modelo (sujeito às restrições iniciais) e deixar o procedimento de reestimação determinar os valores ótimos (máxima verossimilhança). Contudo, experimentos com o procedimento de reestimação mostram que as estimativas de máxima verossimilhança das médias, $\vec{\mu}$, são muito sensíveis às estimativas iniciais [14].

Em geral, os HMM's são na realidade muito mais sensíveis a pequenos erros nos valores de $\vec{\mu}$ que a pequenos erros nos valores da matriz ganho das misturas, **C**, ou da matriz de transição de estados, **A**, a menos que essas diferenças sejam extremamente grandes [39]. O grau de sensibilidade depende das relações entre as médias e as variâncias associadas [39]. Com uma boa estimativa inicial para as médias, o procedimento de reestimação dos parâmetros é capaz de gerar bons modelos mesmo que outros parâmetros tenham estimativas iniciais pobres. Dessa forma, faz-se necessário um procedimento para fornecer boas estimativas iniciais de $\vec{\mu}$ para cada mistura.

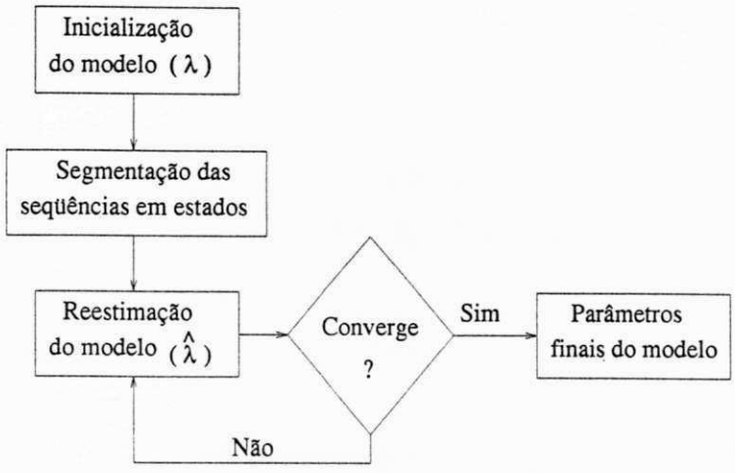


Figura 4.1: Diagrama de blocos da fase treinamento dos HMM's.

Como mostrado na Figura 4.1, dada acima, o primeiro passo no procedimento de treinamento é a escolha de uma estimativa inicial do modelo. Para as estimativas iniciais dos parâmetros dos HMM's foi utilizada uma única ocorrência de cada um dos dígitos. A estimativa inicial foi dada da seguinte forma: para a matriz ganho das misturas, $\mathbf{C} = \{c_{ij}\}$, inicialmente foi considerada apenas uma função densidade de probabilidade associada a cada estado do modelo. Ou seja:

$$c_{i1} = 1 \qquad\qquad 1 < i < N \tag{4.1}$$

Em um segundo caso, considerando-se agora uma mistura de duas fdp's por estado, a estimativa inicial para a matriz ganho das misturas foi dada como:

$$c_{ij} = 1/2 \qquad 1 < i < N \;\; e \;\; 1 \leq j \leq 2. \tag{4.2}$$

Para a matriz de transição de estados, $\mathbf{A} = \{a_{ij}\}$, uma mesma matriz foi considerada para todas as palavras do vocabulário. Uma vez que, teoricamente, é possível a escolha aleatória dos parâmetros de um HMM, os valores de probabilidade dos elementos da matriz transição de estados foram inicialmente estimados como a média aritmética dos elementos não nulos do modelo em cada linha. Ou seja,

$$A = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.3}$$

Para os vetores médias, $\vec{\mu} = \{\mu_j\}$, as estimativas iniciais foram obtidas através de uma segmentação manual e linear dos vetores característicos de cada palavra do vocabulário em estados. Em outras palavras, o número total de vetores característicos da palavra é dividido igualmente para cada um dos estados do modelo. A segmentação em estados resulta em um conjunto de vetores característicos pertencentes a cada um dos estados. Após a segmentação da palavra de treino em estados, um vetor média foi calculado para cada estado, onde cada uma de suas componentes é dada pela média das respectivas componentes dos vetores pertencentes àquele estado.

Para as matrizes covariâncias, $\mathbf{U} = \{U_j\}$, os elementos das matrizes covariâncias (diagonais) foram inicialmente estimados de forma análoga aos elementos dos vetores médias. Ou seja, através do cálculo das variâncias entre os elementos dos vetores característicos pertecentes aos seus respectivos estados.

Para o caso de duas componentes de mistura, as estimativas iniciais, para os vetores médias e para as matrizes covariâncias, foram realizadas de forma análoga às anteriores, utilizando-se nesse caso duas pronúncias de cada dígito para extração dos parâmetros das misturas.

O segundo passo no processo de treinamento é segmentar cada uma das seqüências de observações em estados. Esta segmentação é levada a efeito dividindo-se o número total de vetores de observação de cada palavra em estados, de forma que cada um dos estados contenha o mesmo número de vetores, com excessão do último estado que poderá conter um número maior.

A Figura 4.2, dada a seguir, mostra um gráfico que ilustra a segmentação inicial em estados para uma locução do dígito zero.

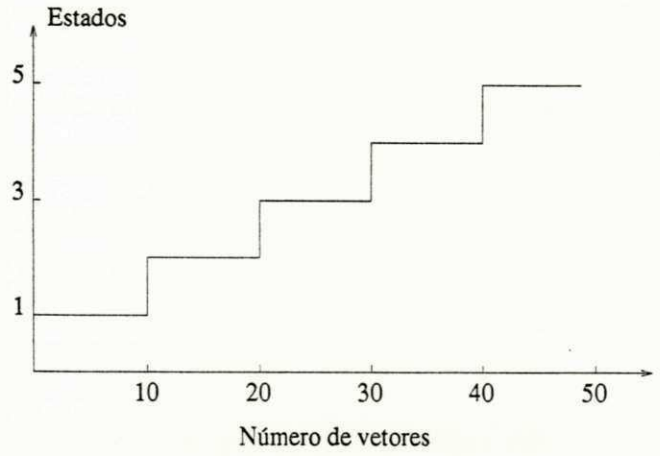


Figura 4.2: Segmentação inicial em estados para uma locução do dígito zero.

Após a segmentação em estados de todas as seqüências de observações de treinamento, o algoritmo de reestimação iterativo de Baum-Welch [22], é utilizado para a obtenção dos parâmetros finais de cada modelo. Para o caso do reconhecimento independente do locutor, o procedimento de reestimação é realizado sobre um banco de dados formado por 10 locuções de cada um dos dígitos (de zero a nove) isolados e fornecidos por 10 locutores distintos. Para o caso do reconhecimento dependente do locutor, o treinamento dos HMM's é levado a efeito considerando-se um outro banco de dados formado por 10 locuções de cada um desses mesmos 10 dígitos, porém adquiridos a partir de um único locutor.

O algoritmo de Baum-Welch é um procedimento iterativo que deve ser aplicado várias vezes até que se tenha atingido um certo critério de convergência, ou um determinado número de iterações. O modelo resultante, após cada iteração, é comparado ao modelo anterior através de uma medida de distância que reflete a similaridade estatística das HMM's. Se a medida de distância excede um determinado limiar, o modelo antigo, λ, é trocado pelo novo modelo, $\hat{\lambda}$ (o resultado da reestimação), e todo o processo é repetido. Se a medida de distância cai abaixo desse limiar, a convergência do modelo é assumida e os parâmetros do modelo são salvos [14].

A partir dos resultados obtidos através de testes de simulação realizados durante

este trabalho, foram considerados dois pontos de parada para o procedimento de reestimação dos parâmetros dos HMM's. Ou seja, o algoritmo deve iteragir até que a distância entre o modelo corrente e o modelo anterior seja menor do que 10^{-5}, ou caso este valor não seja atingido, o número de iterações seja igual a 10.

A medida de distância utilizada para medir a dissimilaridade estatística entre dois modelos de HMM's está baseada na medida de máxima verossimilhança e é dada por [40]:

$$D(\lambda, \hat{\lambda}) = \frac{1}{T}[\log P(\mathbf{O}/\lambda) - \log P(\mathbf{O}/\hat{\lambda})] \tag{4.4}$$

A medida de distância dada acima, fornece a distância normalizada em termos do logaritmo da verossimilhança das seqüências de observações de treinamento, onde a normalização é o número total de vetores, T, contidos nessas sequências.

As Figuras 4.3 e 4.4, dadas a seguir, mostram os gráficos com os valores do logaritmo da verossimilhança (Fig. 4.3), e os valores de distância (Fig. 4.4), obtidos na fase de treinamento para o dígito 0 (zero).

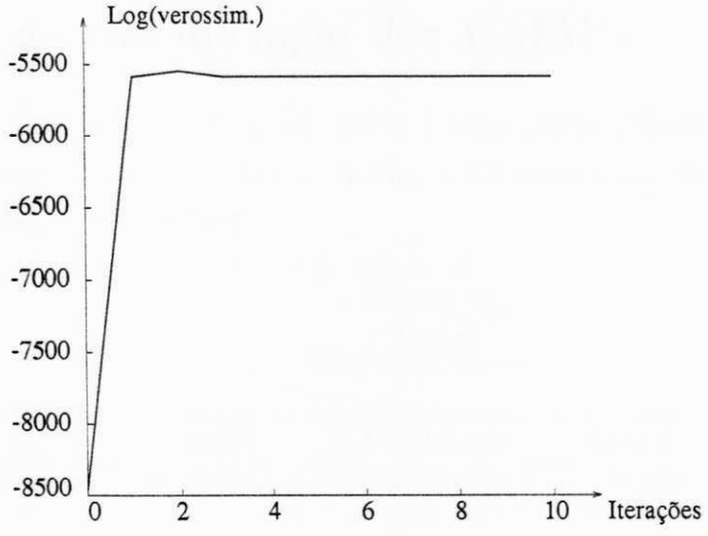


Figura 4.3: Valores do logaritmo de verossimilhança para o dígito zero.

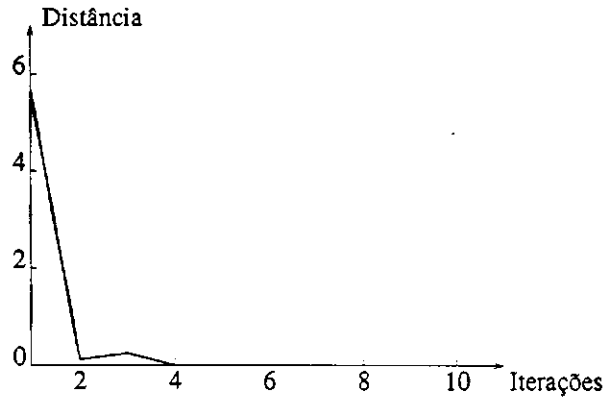
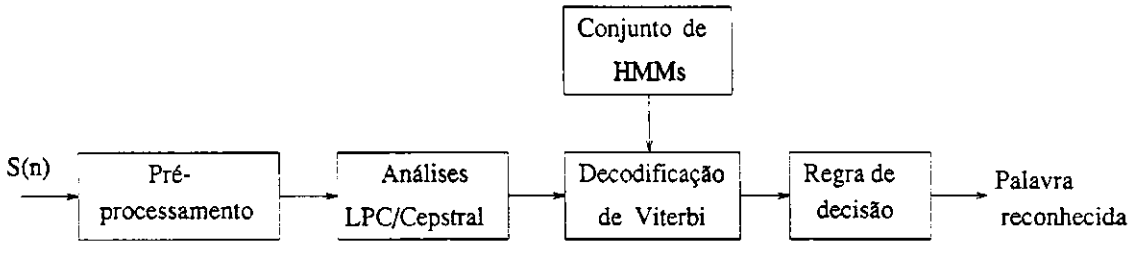


Figura 4.4: Valores de distância entre modelos no processo de reestimação dos parâmetros do HMM para o dígito zero.

A partir da análise dos gráficos mostrados nas Figuras 4.3 e 4.4, pode-se observar que, a partir da quarta iteração a distância entre os modelos reestimados é praticamente desprezível. Ou seja, o número de iterações utilizado neste trabalho (10 iterações), garante a convergência do modelo. Estes resultados são ainda válidos para todos os outros dígitos.

4.3 Fase de classificação dos HMM's

Uma vez que os HMM's tenham sido treinados para cada palavra do vocabulário, a estratégia de reconhecimento é direta. A Figura 4.5 mostra um diagrama de blocos simplificado do classificador HMM.



Fonte: Rabiner (1985) [14], pag.1219.

Figura 4.5: Diagrama de blocos do classificador HMM

Todas as etapas presentes no classificador HMM implementado neste trabalho são descritas a seguir:

4.3.1 Pré-processamento do sinal

O processo de classificação, conforme mostrado na Figura 4.5, é iniciado por um pré-processamento do sinal que consiste das seguintes etapas:

1. Detecção de início e fim de palavras;
2. Pré-ênfase;
3. Segmentação e janelamento.

4.3.1.1 Detecção de início e fim de palavras

O reconhecimento de palavras isoladas está baseado na hipótose de que o sinal, dentro de um intervalo de gravação, consiste de uma sentença isolada, isto é, existe um intervalo de silêncio ou qualquer outra base de ruído antes e depois da ocorrência dessa palavra. Assim, quando uma palavra é falada, é assumido que os segmentos de voz podem ser separados dos segmentos de pausas de uma maneira segura. O processo de separação dos segmentos de voz de uma sentença de sua base de silêncio (ou ruído) é chamado de detecção de início e fim de palavra ("*endpoint detection*").

Em sistemas de palavras isoladas, a detecção correta dos pontos de início e fim de palavras é importante por duas razões [38]:

1. A classificação correta da palavra é criticamente dependente da precisão dessa detecção.
2. Os cálculos necessários para o processamento do sinal de voz são minimizados quando os *endpoints* são localizados com precisão.

Os problemas na detecção dos pontos limites (início e fim) de uma palavra estão associados aos ruídos produzidos pelo locutor, ambiente de gravação e/ou pelo sistema de transmissão. Esses ruídos dificultam a tarefa de detecção consideravelmente, pois se confundem com segmentos de baixa energia do sinal de voz e mascaram os intervalos de silêncio.

Uma maneira de minimizar os efeitos do ruído ambiental é a utilização de microfones de alta diretividade em ambientes acusticamente isolados para gravar o sinal de voz. Contudo, este método não é adequado para várias aplicações reais (por exemplo, transmissão via linha telefônica). Portanto, um método de detecção confiável dos pontos limites de uma palavra é um componente essencial para o bom desempenho de um reconhecedor de palavras isoladas.

Embora o desempenho de todo reconhecedor de fala esteja diretamente relacionado à precisão do detector dos pontos limites das palavras, pouco tem sido publicado a respeito de algoritmos específicos para realizar essa tarefa [37, 38].

O algoritmo utilizado para a detecção dos pontos limites das palavras neste trabalho foi proposto por Rabiner e Sambur [37]. Esse algoritmo está baseado em duas medidas temporais do sinal de voz de entrada, ou seja, a energia segmental e a taxa de cruzamentos por zero (Fig. 4.6).

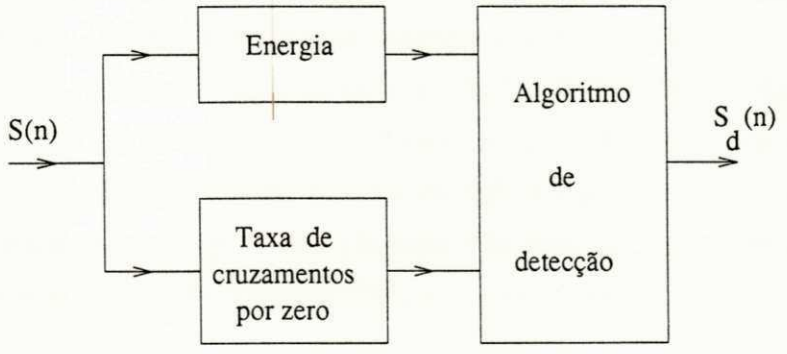


Figura 4.6: Detecção dos *endpoints* baseada no cálculo da energia e da taxa de cruzamentos por zero.

Vários limiares fixos de energia do sinal são utilizados para fornecer uma estimativa inicial dos pontos limites da palavra, e a taxa de cruzamentos por zero (tcr) é utilizada para um refinamento posterior na discriminação dos intervalos de fala de baixa energia (sons surdos) (Fig. 4.7).

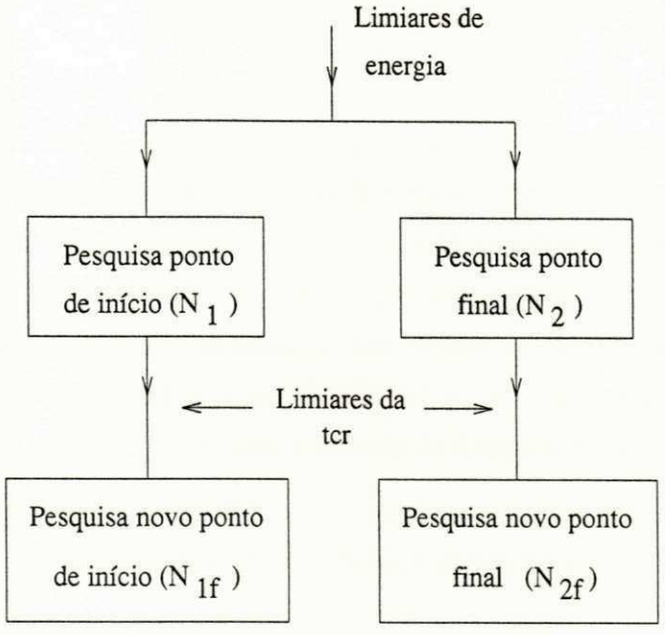


Figura 4.7: Diagrama de blocos generalizado do algoritmo de detecção implementado.

As principais características desse algoritmo são: processamento simples, rápido e eficiente; localização confiável de eventos acústicos significativos e a facilidade de poder ser aplicado a várias bases de silêncio onde a relação sinal-ruído seja da ordem de 30 dB ou mais [37]. Na prática, contudo, foi observado que o algoritmo implementado não apresenta desempenho uniforme para todos os dígitos quando a gravação é realizada em ambiente ruídoso, como por exemplo uma sala com computadores em operação. Apresentando, eventualmente, erros facilmente observáveis na determinação dos pontos de início e fim das palavras.

Um algoritmo para detecção de início e fim de palavras otimizado é proposto por Lamel et al. [38], onde os pontos limites da palavra em questão são inicialmente estimados através de medidas de energia e ajustados a partir de resultados fornecidos

na saída do reconhecedor (*feedback*).

Existe ainda, um algoritmo para a classificação de cada segmento do sinal de voz como um som sonoro, um som surdo ou ausência de voz, baseado em um levantamento estatístico da ocorrência dos sons da fala, utilizando parâmetros temporais [41].

4.3.1.2 Pré-ênfase

As componentes de freqüência mais altas do sinal de voz são caracterizadas por apresentarem baixas amplitudes e por isso facilmente corrompidas pelo ruído. Apesar do sinal de voz ter a energia mais concentrada nas freqüências mais baixas, as componentes de freqüência mais altas são responsáveis pela geração dos sons surdos (fricativos). Por isso, após a determinação dos pontos de início e fim das palavras, é levado a efeito uma pré-ênfase no sinal de voz a fim de tornar mais plano o espectro desse sinal [39]. A pré-ênfase é realizada através da fórmula usual:

$$S_P(n) = S(n) - 0,95 \times S(n-1) \tag{4.5}$$

4.3.1.3 Segmentação em blocos e janelamento

Uma vez feita a pré-ênfase do sinal de entrada, este é segmentado em blocos de 40 ms (NA = 320 amostras) e é analisado a cada 10 ms. Para cada bloco de amostras é feito um janelamento, isto é, cada bloco é multiplicado por uma janela de Hamming, $w(n)$, para minimizar os efeitos adversos resultantes da segmentação abrupta que causa descontinuidades no espectro do sinal de voz.

$$W(n) = 0,54 - 0,46 \times \cos(2\pi n/(NA-1)) \quad 0 \leq n \leq NA-1 \tag{4.6}$$

4.3.2 Análise preditiva linear (Análise LPC)

Uma das mais poderosas técnicas de análise de sinais de voz é o método da análise preditiva linear. Esse método tem se tornado a técnica predominante para a estimação dos parâmetros básicos da fala, como por exemplo, *pitch*, formantes, espectro, funções do trato vocal, e para representar a fala para a transmissão ou armazenamento a baixa taxa de bits. A importância deste método está ligada à sua capacidade de fornecer estimativas extremamente eficientes dos parâmetros da voz, além de sua alta velocidade de cálculo [39].

O princípio básico associado a análise preditiva linear é que o sinal de voz é modelado como uma combinação linear de seus valores passados e de valores presentes e passados de uma entrada hipotética de um sistema cuja saída é o sinal dado. No domínio da freqüência isto é equivalente a modelar o espectro do sinal por um espectro de polos e zeros.

Através da minimização da soma das diferenças quadradas (sobre um intervalo finito) entre as amostras reais da fala e as outras amostras obtidas através da combinação linear das primeiras, um único conjunto de coeficientes do preditor pode ser determinado. Os coeficientes do preditor são os coeficientes de ponderação usados na combinação linear.

A idéia básica de predição linear leva a um conjunto de técnicas de análise que podem ser usadas para estimar parâmetros de um modelo de fala. Este conjunto geral de técnicas de análise preditiva linear é frequentemente chamada de Análise por Codificação Preditiva Linear ou Análise LPC [39].

O principal problema associado a análise de predição linear é determinar um conjunto de coeficientes do preditor diretamente a partir do sinal de voz, a fim de se obter uma boa estimativa das propriedades espectrais do sinal de voz. Devido à natureza variante no tempo do sinal de voz, os coeficientes do preditor devem ser estimados em segmentos de curtos intervalos de tempo.

Vários são os métodos conhecidos para a determinação dos coeficientes do preditor,

dentre esses, o método da autocorrelação é o mais utilizado [39].

Para cada bloco do sinal de entrada, após o janelamento, a função de autocorrelação, $R_n(i)$,

$$R_n(i) = \sum_{k=0}^{NA-k} S_p[k] \times S_p[k+1] \tag{4.7}$$

é calculada a curtos intervalos de tempo (10 ms) , obtendo-se um vetor de dimensão nove.

Para o método da autocorrelação, a equação matricial para se obter a solução dos coeficientes do preditor é da forma [39],

$$\sum_{k=1}^{\rho} \alpha_k R_n(\mid i-k \mid) = R_n(i) \qquad 1 \leq i \leq \rho \tag{4.8}$$

onde ρ é a ordem do preditor, α_k são os coeficientes do preditor e R_n(i) é a função de autocorrelação do sinal de voz.

O método mais eficiente para se resolver esse sistema de equações é o método recursivo de Durbin [39], conhecido também como o Algoritmo de Durbin, que pode ser estabelecido como segue:

$$\begin{cases} E^{(0)} = R(0) & \\ k_i = [R(i) - \sum_{j=1}^{i-1} \alpha_j(i-1)R(i-j)]/E^{(i-1)} & 1 \leq i \leq \rho \\ \alpha_i^{(i)} = k_i & \\ \alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} & 1 \leq j \leq i-1 \\ E^{(i)} = (1-k_i^2)E^{(i-1)} & \end{cases} \tag{4.9}$$

as equações acima são resolvidas recursivamente para $i = 1, 2, \ldots, \rho$ e a solução final é dada por,

$$\alpha_j = \alpha_j^{\rho} \qquad 1 \leq j \leq \rho \tag{4.10}$$

onde, $E^{(i)}$ é denominado de erro de predição para um preditor de i-ésima ordem. As quantidades intermediárias $k_i, 1 \leq i \leq \rho$, são conhecidas como os coeficientes de reflexão, ou ainda, como coeficientes de correlação parcial (Parcor).

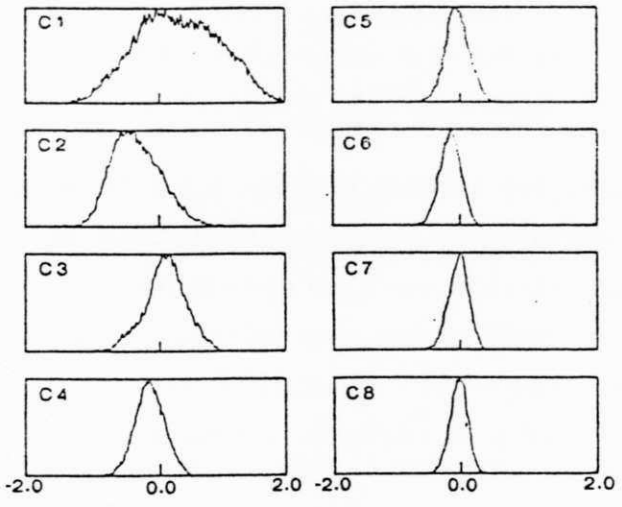
No trabalho ora em questão, foi realizada uma análise LPC onde um vetor de 8 coeficientes é calculado a partir de um vetor de autocorrelação usando-se o algorítmo recursivo de Durbin.

4.3.3 Análise cepstral

Os coeficientes cepstrais, c(i), são calculados recursivamente a partir dos coeficientes do preditor linear, $\alpha(i)$, através das seguintes relações [42]:

$$\begin{cases} c(1) = -\alpha(1) \\ c(i) = -\alpha(i) - \sum_{k=1}^{i-1}(1 - k/i)\alpha(k)c(i-k) \quad 1 < i \leq \rho \end{cases} \tag{4.11}$$

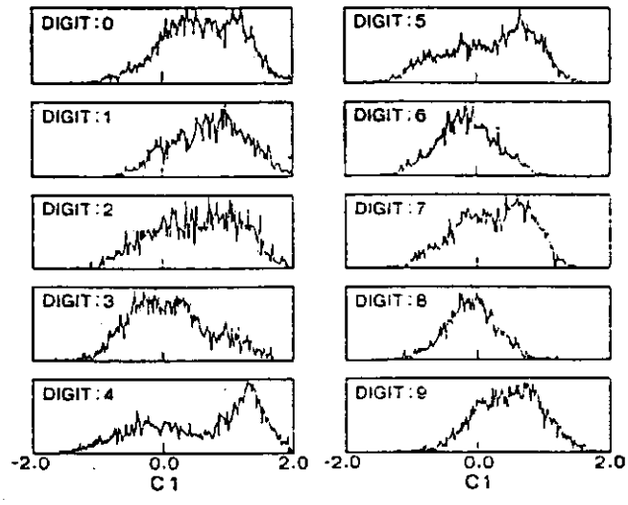
A Figura 4.8, mostra as distribuições estatísticas dos coeficientes cepstrais obtidos a partir de um banco de dados de dígitos, da língua inglesa, fornecidos por 100 locutores [42].



Fonte: Yoh'ichi (1987) [42], pag.1415.

Figura 4.8: Distribuições estatísticas dos coeficientes cepstrais.

As distribuições individuais do primeiro coeficiente cepstral, para cada dígito, são mostradas na Figura 4.9.



Fonte: Yoh'chi (1987) [42], pag.1415.

Figura 4.9: Distribuições estatísticas dos coeficientes cepstrais de primeira ordem para cada um dos dez dígitos.

A partir das Figuras 4.8 e 4.9 , pode-se observar que as distribuições de cada coeficiente cepstral, para qualquer um dos dígitos, são aproximadamente Gaussianas, embora a distribuição de cada dígito seja um pouco diferente uma da outra. Também pode ser visto da Figura 4.8, que existe uma grande diferença na variância entre os coeficientes cepstrais. Isto é, a variância dos coeficientes cepstrais de ordem mais alta é muito menor do que as variâncias dos coeficientes de ordem mais baixa [42].

4.3.4 Cálculo da energia segmental

A energia presente em um quadro do sinal voz, chamada de energia segmental, é definida como a soma dos quadrados das amostras que fazem parte desse quadro. A

energia é um parâmetro temporal do sinal de voz bastante utilizado em processamento digital de fala para a distinção entre intervalos de sons surdos e sons sonoros.

Para a utilização da energia segmental como o nono parâmetro do vetor característico, uma transformação logaritmica foi aplicada ao valor da energia normalizada para aproximá-lo à escala perceptual de audição [43]:

$$\log \mathbf{E} = \log[(\sum_{n=1}^{NA} S^2(n))/E_{\max}], \tag{4.12}$$

onde S(n) representa a i-ésima amostra do sinal de voz em um quadro de tamanho NA e E_{max} é a energia máxima desse sinal. O vetor característico resultante, de dimensão 9, representará o sinal de entrada em cada instante de tempo.

4.3.5 O vetor de variáveis aleatórias

O vetor característico, $V(x_1, x_2, ..., x_9)$, formado pelos 8 coeficientes cepstrais (c_1, c_2, $c_3..., c_8$) e o logaritmo da energia segmental , $\log \mathbf{E}$, constitui-se, em cada instante de tempo, em um dos possíveis eventos gerados por um vetor aleatório formado por 9 variáveis aleatórias.

A função densidade de probabilidade conjunta do vetor característico, assumindo-se variáveis aleatórias Gaussianas, pode ser expressa de maneira mais concisa usando-se uma notação vetorial. Representando-se esse vetor aleatório por $\mathbf{V}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_9)$ como um vetor linha, então $\mathbf{V}'$, o transposto de V, é um vetor coluna:

$$\mathbf{V}' = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_8 \\ \mathbf{X}_9 \end{bmatrix} \tag{4.13}$$

A matriz covariância **U** é dada por:

$$\mathbf{U} = \begin{bmatrix} U_{11} & U_{12} & \dots & U_{19} \\ U_{21} & U_{22} & \dots & U_{29} \\ \vdots & \vdots & & \vdots \\ U_{91} & U_{92} & \dots & U_{99} \end{bmatrix} \tag{4.14}$$

onde U_{ij} é a covariância entre a variável aleatória $\mathbf{X}_i$ e a variável aleatória $\mathbf{X}_j$, a qual é dada por, $\mathrm{U}_{ij} = \overline{(\mathbf{X}_i - \mu_i)(\mathbf{X}_j - \mu_j)}$. Onde μ_i é o valor médio da variável aleatória $\mathbf{X}_i$ e a barra sobre a expressão indica o seu valor médio.

$\mid \mathbf{U} \mid$ é o determinante e $\mathbf{U}^{-1}$ é a inversa da matriz covariância **U**.

Assim, a função densidade de probabilidade conjunta das 9 variáveis conjuntamente Gaussianas pode ser expressa por:

$$b_{\mathbf{v}}(V) = \frac{1}{(2\pi)^{9/2}\sqrt{\mid \mathbf{U} \mid}}[-\frac{1}{2}(V - \vec{\mu})\mathbf{U}^{-1}(V - \vec{\mu})'] \tag{4.15}$$

onde $\vec{\mu}$ representa o vetor média das variáveis aleatórias.

Considerando-se a matriz covariância, **U**, duas formas para essa matriz podem ser consideradas, ou seja, matrizes diagonais (correlação nula entre as componentes da representação vetorial) e matrizes covariâncias completas.

Se as variáveis aleatórias contidas no vetor aleatório **V** são descorrelacionadas, então $\mathrm{U}_{ij} = 0$ para $\mathrm{i} \neq \mathrm{j}$ e a matriz covariância **U** torna-se uma matriz diagonal.

A vantagem da matriz covariância diagonal é que o cálculo de $\mathrm{b}_j(V)$, ou seja, a probabilidade de vetor característico V pertencer ao estado q_j da cadeia de Markov, reduz-se a uma simples soma de produtos dos parâmetros das Gaussianas. Por outro lado, para uma matriz covariância completa, o cálculo de $\mathrm{b}_j(V)$ requer uma multiplicação de matrizes. A desvantagem da matriz covariância diagonal é que, em geral, para componentes vetoriais correlacionadas é necessário um valor de **M** (o número de misturas) maior do que o exigido para a representação da matriz covariância completa para dar um modelo adequado. Nenhuma representação tem qualquer vantagem particular em termos da facilidade de tomar-se estimativas iniciais ou facilidades de reestimação.

Experimentos têm mostrado que é possível modelar um processo aleatório correlacionado de dimensão D por uma mistura de M processos aleatórios Gaussianos, D-dimensionais, descorrelacionados [39].

Considerando-se, então, as variáveis aleatórias Gaussianas que compõem o vetor característico como sendo descorrelacionadas, elas também são por sua vez independentes. Daí, a distribuição de probabilidade conjunta dada pela Equação 4.15, reduz-se ao produto de 9 funções densidade de probabilidade normais unidimensional (Eq. 4.16) [44].

$$b_{\mathbf{v}}(V) = \prod_{i=1}^{9} \frac{1}{\sqrt{2\pi \mid \mathbf{U} \mid}} \exp(-(x_i - \mu_i)^2 / 2U_{ii}) \tag{4.16}$$

4.3.6 Decodificação de Viterbi

Após a determinação dos vetores característicos da palavra de entrada, o próximo passo no reconhecimento é encontrar a seqüência de estados ótima correspondente ao HMM para cada palavra do vocabulário e calcular o valor da medida log-verossimilhança para o caminho ótimo. O **Algoritmo de Viterbi** - Equações (3.28)-(3.34), determina a seqüência de estados ótima e o logaritmo da máxima verossimilhança.

Considerando-se o modelo de HMM do tipo *left-right* dado pela Figura 3.2 (capítulo 3), o algoritmo de Viterbi pode ser compactamente estabelecido como [37]:

1) Inicialização:

$$\begin{cases} \delta_1(1) = \log[b_1(O_1)] \\ \delta_1(i) = -\infty \qquad i \neq 1 \end{cases} \tag{4.17}$$

$$\Psi_1(i) = 0 \qquad 1 \leq i \leq N \tag{4.18}$$

2) Recursividade:

$$\delta_t(j) = \underbrace{\max}_{j-2 \leq i \leq j} \{\delta_{t-1}(i) + \log a_{ij}\} + \log[b_j(O_t)] \quad 2 \leq t \leq T, \quad 1 \leq j \leq N \tag{4.19}$$

$$\Psi_t(j) = \underbrace{\operatorname{argmax}}_{j-2 \leq i \leq j} [\delta_{t-1}(i) + \log a_{ij}] \quad 2 \leq t \leq T, \quad 1 \leq j \leq N \tag{4.20}$$

3) Término:

$$\log f = \delta_T(N) \tag{4.21}$$

$$q_T^* = N \tag{4.22}$$

4) Seqüência de estados ótima

$$q_t^* = \Psi_{t+1}(q_{t+1}^*) \qquad t = T-1,\ T-2, \ldots,\ 1 \tag{4.23}$$

A Figura 4.10, dada a seguir, mostra a seqüência de estados ótima para uma locução do dígito 0 (zero), cuja segmentação inicial em estados é mostrada na Figura 4.2. Como pode ser observado a partir dessas figuras, a seqüência inicial de estados difere de forma bastante significativa da seqüência de estados "ótima" obtida através do algoritmo de Viterbi. Assim, o uso do algoritmo de Viterbi para segmentar as seqüências de observações em estados, na fase de treinamento, poderá proporcionar melhores resultados para a reestimação dos parâmetros dos HMM's.

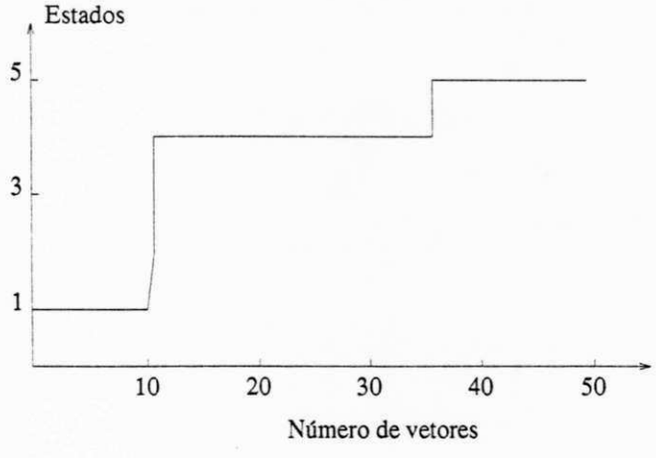


Figura 4.10: Seqüência ótima de estados , determinada pelo algoritmo de Viterbi, para uma locução do dígito zero.

4.3.7 A regra de decisão

A regra de decisão associa a palavra desconhecida à palavra do vocabulário cujo modelo tem o valor mais alto da medida log-verosimilhança, fornecida pelo algoritmo de Viterbi.

Capítulo 5

Avaliações e resultados experimentais

5.1 Condições experimentais

As avaliações experimentais deste trabalho foram realizadas utilizando-se um microcomputador compatível com IBM PC, Cx80486DLC, 40 MHz, com 4 Mbytes de memória RAM. Para a aquisição dos dados foi utilizado um conversor A/D, parte integrante de um cartão contendo o processador TMS320C25, com a digitalização sendo realizada com 16 bits/amostra. A freqüência de amostragem utilizada foi de 8 khz.

Para avaliar o desempenho do reconhecedor HMM utilizando funções densidade de probabilidade contínuas, uma série de experimentos foi realizada nos quais alguns parâmetros dos modelos variaram. As avaliações iniciais foram realizadas sobre um banco de dados formado por 10 locuções de cada um dos dígitos (de zero a nove) isolados e fornecidos por 10 locutores distintos (reconhecimento independente do locutor). Outras avaliações foram levadas a efeito considerando-se um outro banco de dados formado ainda pelos mesmos 10 dígitos, porém adquiridos a partir de um único locutor (reconhecimento dependente do locutor).

Quatro conjuntos de dígitos foram usados. Esses consistiam do seguinte:

CONJ 1 - 10 locutores (5 masculinos, 5 femininos) e uma pronúncia de cada dígito por locutor.

CONJ 2 - 10 novos locutores (5 masculinos e 5 femininos) e uma pronúncia de cada dígito por locutor.

CONJ 3 - Um único locutor masculino e 10 pronúncias de cada dígito.

CONJ 4 - O mesmo locutor do conjunto CONJ 3; 10 pronúncias de cada dígito e gravações feitas várias semanas após aquelas do banco de dados CONJ 3.

Assim, cada um dos quatro conjuntos acima continha 100 dígitos. Para o treinamento dos modelos apenas os bancos de dados CONJ 1 ou CONJ 3 foi utilizado; para teste e avaliação de desempenho do reconhecedor os conjuntos CONJ 2 ou CONJ 4 foram usados.

5.2 Reconhecimento HMM independente do locutor sem treinamento do modelo

O primeiro experimento teve como objetivo a avaliação de desempenho do reconhecedor HMM independente do locutor sem nenhuma espécie de treinamento. Isto é, a fase de treinamento consistiu apenas da estimativa inicial para os parâmetros dos HMM's.

Para as estimativas iniciais dos parâmetros dos HMM's foi utilizada uma única ocorrência de cada um dos dígitos extraídas do conjunto de dígitos CONJ 3. Neste primeiro experimento, foi considerada apenas uma única função densidade de probabilidade associada a cada estado do modelo.

O conjunto de dígitos utilizado na fase de classificação para o primeiro experimento foi o CONJ 2. Os resultados obtidos a partir desse experimento podem ser observados através do gráfico mostrado na Figura 5.1 dada a seguir, que fornece a taxa de erro para cada um dos locutores.

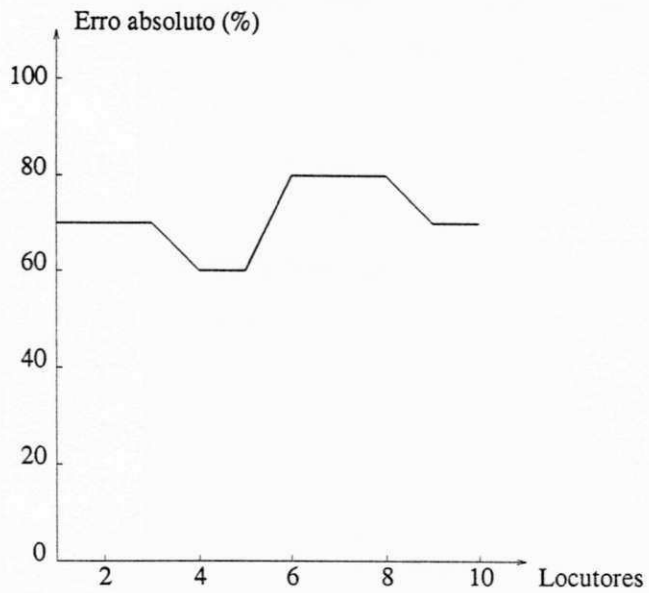


Figura 5.1: Avaliação inicial do reconhecedor HMM para o reconhecimento de dígitos independente do locutor (sem treinamento).

A partir da análise do gráfico dado acima, pode-se concluir sobre a necessidade imperativa do treinamento dos modelos antes da fase de classificação do sistema HMM independente do locutor. É possível notar também, uma pequena queda no desempenho do reconhecedor quando os locutores que emitem os dígitos para a classificação são mulheres (locutores de 6 a 10). Isto indica que o sistema não tem um desempenho uniforme independente do sexo do locutor.

5.3 Reconhecimento independente do locutor utilizando uma única fdp por estado

Constatada a necessidade de treinamento dos modelos HMM's antes da fase de classificação, para o reconhecimento independente do locutor, o segundo experimento avaliou o desempenho do sistema proposto fazendo uso, agora, do algoritmo de treinamento de Baum-Welch.

Para a obtenção dos parâmetros finais dos HMM's foram utilizados, na fase de

treinamento, os 10 conjuntos de dígitos contidos no banco de dados CONJ 1. Considerou-se, ainda apenas uma única função densidade de probabilidade Gaussiana por estado do modelo.

A Figura 5.2, mostrada a seguir, ilustra graficamente o comportamento do reconhecedor HMM de densidade contínua com apenas uma única fdp por estado do modelo (com treinamento). O gráfico mostra o desempenho do reconhecedor para cada um dos locutores do banco de dados CONJ 2.

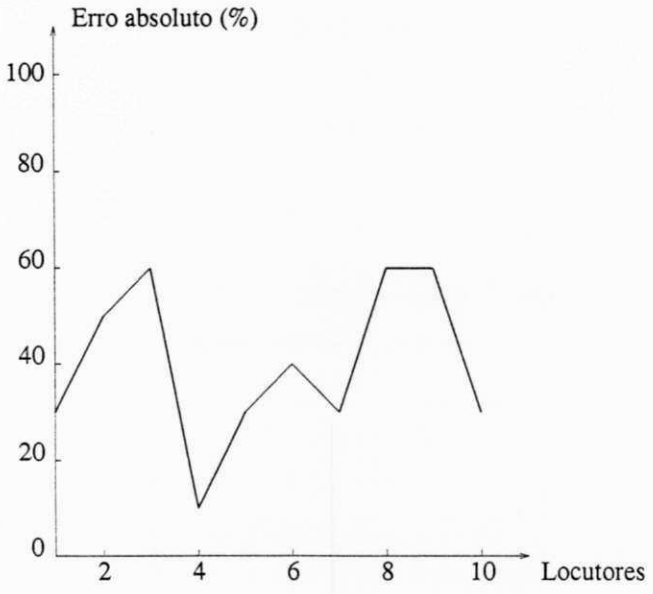


Figura 5.2: Desempenho do reconhecedor HMM com uma única fdp por estado do modelo (com treinamento).

O gráfico da Figura 5.2 mostra uma sensível melhora no desempeho do reconhecedor HMM após o treinamento dos modelos. A taxa média de erros verificada antes do treinamento era de 71%, caindo para 40% após o treinamento. Pode-se notar, entretanto, que o sistema não funciona de maneira uniforme para todos os locutores. Para o locutor 4, por exemplo, a taxa de erros encontrada foi de apenas 10% - um único dígito classificado de maneira errada. Por outro lado, para outros locutores foram obtidas taxas de erro de até 60% (locutores 3, 8 e 9).

Apesar da significativa melhora no desempenho do reconhecedor, em relação ao

experimento anterior, este ainda apresenta uma alta taxa média de erros. Este fato pode ser justificado pela utilização de uma única fdp por estado concomitantemente com uma matriz covariância diagonal (hipótese das variáveis aleatórias serem descorrelacionadas).

Uma segunda avaliação pode ser feita considerando-se a Tabela 5.1 dada a seguir. Nessa tabela, a primeira coluna e primeira linha indicam os dígitos, que correspondem à entrada e saída do reconhecedor HMM. A classificação de um dado dígito de entrada - uma dada linha - como um dos dígitos de saída é representada pelo cruzamento dessa linha com a respectiva coluna. Assim, na diagonal principal estão concentrados os acertos, e os demais valores representam o erro absoluto obtido na fase de classificação para cada um dos dígitos.

SAÍDA / ENTRADA	ZERO	UM	DOIS	TRÊS	QUATRO	CINCO	SEIS	SETE	OITO	NOVE
ZERO	6	1			3					
UM		7			1				2	
DOIS			6		2		2			
TRÊS			1	5	3		1			
QUATRO	2				6			1	1	
CINCO			1		4	5				
SEIS					3		6	1		
SETE					2	1		7		
OITO					4				6	
NOVE	1				1	1			1	6

Tabela 5.1: Avaliação do reconhecedor HMM independente do locutor com uma única fdp por estado.

A Tabela acima apresenta a performance do reconhecedor HMM para cada um dos

dígitos, quando se tem apenas uma única fdp por estado do modelo. Como pode-se observar ao longo da diagonal principal da Tabela 5.1, o sistema apresenta um desempenho quase uniforme para todos os dígitos (variância igual a 0,4). Para os dígitos pares, por exemplo, a taxa de erros média permaneceu constante e igual a 40%. Para os dígitos ímpares, a taxa de erros média também foi de 40%, embora houvesse uma pequena variação de um dígito para o outro. Também pode ser observado que o sistema, apesar do processo de reestimação ter sido utilizado, apresenta ainda uma certa tendência a classificar as sentenças de teste como a palavra que melhor se adaptou ao modelo inicial (todos os dígitos foram classificados, pelo menos uma vez, como sendo o dígito quatro).

5.4 Reconhecimento independente do locutor utilizando duas fdp's por estado

No terceiro experimento realizado com o reconhecedor HMM para o modo de reconhecimento independente do locutor, consideradou-se uma mistura de duas funções densidade de probabilidade para cada estado.

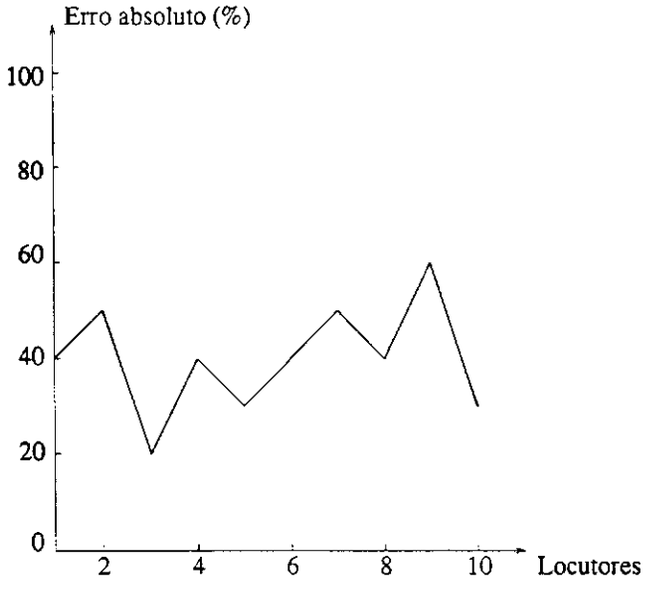


Figura 5.3: Desempenho do reconhecedor HMM com duas fdp's por estado.

O gráfico ilustrado na Figura 5.3, apresenta os resultados obtidos para o desempenho do reconhecedor HMM neste terceiro experimento. A partir da análise deste gráfico, pode-se observar algumas melhoras isoladas (locutores 3 e 8) no desempenho do reconhecedor HMM com duas fdp's, em relação ao modelo com apenas uma fdp por estado. Contudo, para outros locutores (1, 4 e 7) houve uma queda no desempenho do sistema. Além disso, pode-se verificar também que não houve nenhuma melhora na taxa média de erros passando de um modelo para o outro. Ambos apresentam uma taxa de erro média igual a 40%. Houve, porém, uma maior uniformidade no desempenho do reconhecedor, em relação aos locutores, no segundo caso.

Novamente pode-se avaliar o reconhecedor HMM implementado considerando-se o seu desempenho para cada um dos dígitos através da Tabela 5.2.

SAÍDA / ENTRADA	ZERO	UM	DOIS	TRÊS	QUATRO	CINCO	SEIS	SETE	OITO	NOVE
ZERO	3				3				2	2
UM	7	1							2	
DOIS			0	4	3		1	1		1
TRÊS				7			1	1		1
QUATRO	1				4			1	2	2
CINCO						10				
SEIS							8	2		
SETE						2		8		
OITO									10	
NOVE	1				1					8

Tabela 5.2: Avaliação do reconhecedor HMM independente do locutor com uma mistura de duas fdp's contínuas por estado.

O não melhoramento na taxa de erro média pode ser justificado pela quantidade insuficiente de dados de treinamento para a obtenção de modelos mais robustos e

insensíveis as características particulares de alguns locutores usados na fase de treinamento.

À guisa de exemplo, Rabiner [14] utilizou um banco de dados composto por 100 locutores distintos para o treinamento de um sistema de reconhecimento de dígitos falados na língua inglesa. Para o sistema ora em discussão, foram utilizados apenas 10 locutores distintos no procedimento de reestimação. A utilização de um número reduzido de locutores aconteceu, principalmente, devido a dificuldades operacionais encontradas e a indisponibilidade de tempo para coletar dados de um número maior de locutores.

Apesar de apresentarem a mesma taxa média de erros , os experimentos com uma e com duas fdp's por estado apresentam diferentes taxas de erro para cada um dos dígitos. Essa diferença pode ser melhor observada através do gráfico dado pela Figura 5.4.

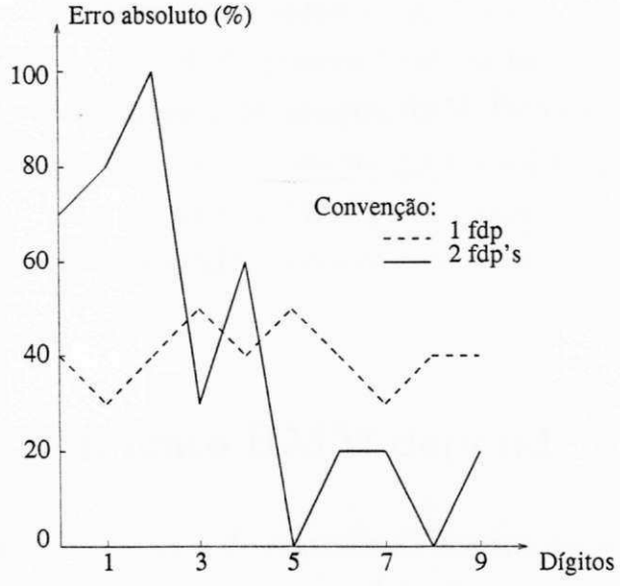


Figura 5.4: Comparação de desempenho do reconhecedor HMM independente do locutor com uma e com duas fdp's por estado.

Pode-se observar do gráfico acima que enquanto o experimento com apenas uma fdp por estado apresenta uma acuracidade bastante uniforme para todos os dígitos (variando de 50% a 70%), o experimento com uma mistura de fdp's apresenta um

desempenho bastante desigual. Para os dígitos cinco e oito, por exemplo, foi obtida uma taxa de acertos de 100% para todos os locutores. Por outro lado, para o dígito dois foi encontrada uma taxa de acertos de 0%. Esta taxa de erro máxima encontrada para o dígito dois pode ser explicada levando-se em consideração a utilização de um número insuficiente de dados de treinamento para a reestimação confiável dos parâmetros do modelo.

Um outro fator que foi observado e que merece ser mencionado é que, para quase todos os dígitos que foram classificados de maneira errônea, a medida de probabilidade apontava quase sempre a palavra correta como a segunda candidata para a classificação. Esse fato indica que utilizando-se um treinamento mais robusto a taxa de erro deverá cair consideravelmente.

Apesar do número insuficiente de seqüências de treinamento para a reestimação confiável dos parâmetros dos modelos, tentou-se ainda o uso de um número maior de fdp's por estado. Contudo, isso não foi possível devido ao fato do algoritmo implementado não utilizar alocação dinâmica de memória RAM, limitando-se, dessa forma, ao uso da memória contínua oferecida pelo sistema operacional DOS. Assim sendo, o uso apenas da memória RAM convencional (640 kbytes) se mostra insuficiente para tratar com o grande volume de dados exigido para o caso do sistema com mais de duas fdp's por estado.

5.5 Reconhecimento HMM dependente do locutor

Com a finalidade de avaliar o desempenho do sistema de reconhecimento HMM implementado, para o caso do reconhecimento dependente do locutor, foram levados a efeito dois experimentos. O primeiro, da mesma forma como foi realizado para a análise do reconhecimento independente do locutor, consistiu apenas da estimativa inicial para os parâmetros dos HMM's (não houve aplicação do algoritmo de reestimação). Para o segundo experimento foi utilizado o banco de dados CONJ 3 na fase de treinamento. Em ambos os casos o banco de dados CONJ 4 foi utilizado na fase de classificação. Também, em ambos casos, foi utilizada apenas uma única fdp para representar as

observações em cada estado.

5.5.1 Reconhecimento dependente do locutor sem treinamento do modelo

As estimativas iniciais dos parâmetros dos HMM's, para o reconhecimento dependente do locutor sem treinamento, ou seja, sem reestimação, foram obtidas através de uma única pronúncia de cada um dos dígitos extraídas do banco de dados CONJ 3.

Os resultados obtidos estão colocados na Tabela 5.3 , dada abaixo:

SAÍDA / ENTRADA	ZERO	UM	DOIS	TRÊS	QUATRO	CINCO	SEIS	SETE	OITO	NOVE
ZERO	7		3							
UM	4	0	1		1	1				3
DOIS			9	1						
TRÊS			2	8						
QUATRO					2					8
CINCO				1		9				
SEIS				6			4			
SETE			1			9		0		
OITO			5			1	1		0	3
NOVE	1		1			2				6

Tabela 5.3: Avaliação do reconhecedor HMM dependente do locutor com uma única fdp por estado (sem reestimação).

Como pode ser observado a partir da Tabela 5.3, mesmo apresentando um bom desempenho para alguns dígitos (2 e 5), o desempenho global do reconhecedor HMM dependente do locutor é bastante critico, apresentando uma taxa de erro média de 55%. E para alguns dígitos (1, 7 e 8) uma taxa de erros de 100%. Isto implica que,

mesmo para o caso do reconhecimento dependente do locutor, se faz necessário um procedimento de treinamento dos modelos antes da fase de classificação.

5.5.2 Reconhecimento dependente do locutor com treinamento do modelo

O último experimento realizado com o sistema de reconhecimento HMM implementado, visa avaliar o seu desempenho para o caso dependente do locutor com a aplicação do algoritmo de reestimação na fase de treinamento dos modelos.

Na fase de treinamento foram utilizados os dez conjuntos de dígitos contidos no banco de dados CONJ 3 para a obtenção dos parâmetros finais dos modelos. Os resultados obtidos estão mostrados no gráfico da Figura 5.5, dada a seguir:

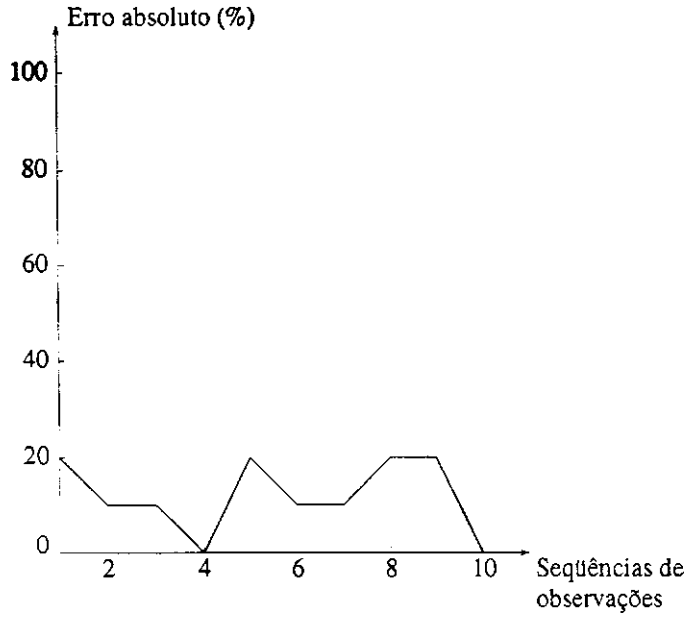


Figura 5.5: Desempenho do reconhecedor HMM dependente do locutor com treinamento.

Como pode ser observado através do gráfico da Figura 5.5, o sistema HMM dependente do locutor apresenta um desempenho ótimo para duas seqüências de teste

utilizadas (4 e 10) e, de uma maneira geral, um desempenho bastante aceitável, considerando-se a utilização de apenas uma única fdp por estado do modelo.

Os resultados deste último experimento podem ser analisados ainda considerando-se a Tabela 5.4 dada a seguir:

SAÍDA / ENTRADA	ZERO	UM	DOIS	TRÊS	QUATRO	CINCO	SEIS	SETE	OITO	NOVE
ZERO	10									
UM		9								1
DOIS			10							
TRÊS			3	7						
QUATRO		1			8				1	
CINCO						9		1		
SEIS							10			
SETE			1			4		5		
OITO									10	
NOVE										10

Tabela 5.4: Avaliação do reconhecedor HMM dependente do locutor com uma única fdp por estado (com reestimação).

Apresentando uma taxa de erro média de 12%, com uma variância de 2,28, e um desempenho ótimo (100% de acertos) para metade das palavras do vocabulário utilizado, o reconhecedor HMM com uma função densidade de probabilidade contínua apresenta, com exceção das taxas de erro encontradas para os dígitos 3 e 7, um desempenho bastante aceitável para o caso do reconhecimento dependente do locutor.

Várias são as possibilidades de mudança no sistema de reconhecimento implementado na busca de uma taxa de erro inferior a encontrada. O primeiro passo seria aumentar a quantidade de dados de treinamento, pois isso possibilitaria, além de uma maior representatividade de cada um dos dígitos, o aumento do número de misturas

de fdp's em cada estado, compensando assim o efeito de um processo correlacionado estar sendo representando por um processo descorrelacionado. Como foi comprovado em um experimento anterior (seção 5.4) o simples aumento do número de componentes de mistura, sem utilizar o número suficiente de dados de treinamento necessário para a reestimação de todos os parâmetros do modelo de modo confiável, não leva a melhora substancial no desempenho do reconhecedor.

Capítulo 6

Conclusão

Este trabalho teve como objetivo o estudo e a aplicação ao reconhecimento automático de fala das funções probabilísticas das cadeias de Markov. Foi implementado, em linguagem de programação C, um sistema de reconhecimento de palavras isoladas para a língua portuguesa, que usa um modelo de densidades de probabilidade contínuas para as fdp's dos vetores característicos em cada estado.

O uso dos modelos de Markov escondidos (*Hidden Markov Models - HMM's*) se apresenta como algo de grande interesse devido à flexibilidade quanto ao modelamento de vários eventos da fala tais como fonemas, difones, sílabas, etc. e por apresentar um custo computacional, na fase de classificação, bastante reduzido quando comparado a outras técnicas mais convencionais.

Por se tratar de um trabalho de pesquisa pioneiro, no âmbito desta instituição de ensino, no estudo e aplicação dos modelos HMM's para a comunicação vocal homem-máquina, muito tempo foi necessário para a aquisição de experiência nessa nova tecnologia.

O reconhecedor HMM implementado teve seu desempenho avaliado para dois modos de reconhecimento: o reconhecimento independente do locutor e o reconhecimento dependente do locutor utilizando, em ambos os casos, um vocabulário de 10 dígitos.

Os resultados obtidos levam às seguintes conclusões:

1. O algoritmo utilizado para a detecção dos pontos limites das palavras, não apresenta um desempenho uniforme para todos os dígitos, considerando-se o ambiente ruídoso no qual foram feitas as gravações utilizadas neste trabalho.

2. Qualquer que seja o modo de reconhecimento considerado, é absolutamente necessária a aplicação de um procedimento de reestimação para a determinação dos valores finais dos parâmetros dos HMM's.

3. A utilização do procedimento de reestimação de Baum-Welch melhora de forma bastante significativa o desempenho do reconhecedor HMM, em relação ao desempenho do sistema sem reestimação.

4. A utilização de uma única fdp contínua, com o uso simultâneo de uma matriz covariância diagonal, apresenta um desempenho bastante uniforme para cada uma das palavras do vocabulário utilizado, embora tenha ainda uma elevada taxa média de erros.

5. O uso de uma mistura de duas fdp's contínuas por estado, apresenta uma relativa melhora no desempenho do sistema para alguns dígitos. Porém, esse desempenho não é uniforme para todas as palavras do vocabulário e a taxa média de erros é mantida invariável.

6. A utilização de uma mistura de fdp's contínuas para representar as observações em cada estado está condicionada a um número suficiente de dados de treinamento para que se possa obter estimativas seguras dos parâmetros do modelo.

7. O número de seqüências de observações utilizado não é suficiente para se obter estimativas melhoradas dos parâmetros dos HMM's, de forma a fornecer uma melhora significativa no desempenho do reconhecimento independente do locutor, através do aumento do número de misturas de fdp's.

8. O desempenho do reconhecedor HMM dependente do locutor é, como era de se esperar, superior ao desempenho apresentado pelo mesmo sistema para o caso do reconhecimento independente do locutor.

9. Os parâmetros da voz utilizados na composição dos vetores característicos (o conjunto de parâmetros cepstrais e o logaritmo da energia) podem ser bem representados pela densidade de probabilidade contínua e oferecem bom desempenho para o reconhecedor HMM, notadamente no modo dependente do locutor.

A seguir são apresentadas algumas sugestões para a continuidade deste trabalho:

- A utilização de um algoritmo de detecção dos pontos de início e fim de palavras capaz de oferecer um desempenho mais uniforme para todas as palavras do vocabulário na presença de ruído. Ou ainda, a realização de uma pré-filtragem do sinal de voz com o objetivo de reduzir o ruído ambiental [45].

- Utilização de um maior número de locutores para a base de dados a ser utilizada no treinamento do sistema para reconhecimento independente do locutor.

- Investigação de novos parâmetros para a composição dos vetores característicos, bem como variações dos parâmetros do modelo utilizado, como por exemplo: número de estados e tipo de matriz covariância (diagonal ou completa).

- Aumento do número de seqüências de observações utilizadas na fase de treinamento, possibilitando assim um aumento no número de componentes de misturas de fdp's por estado.

- Utilização do algoritmo de Viterbi na fase de treinamento dos modelos HMM's, para a segmentação das seqüências de observações em estados.

- Elaboração de um procedimento de treinamento que, junto com o algoritmo de reestimação, possibilitem a estimação "ótima" dos parâmetros do modelo.

- Aumento do número de palavras no vocabulário utilizado.

Ainda, a título de sugestão, são apresentados alguns tópicos para novos trabalhos na área do reconhecimento de fala:

- Simulação, e implementação em *hardware*, de um sistema HMM de palavras conectadas, ou de fala contínua.
- Estudo de um sistema de reconhecimento de fala HMM baseado em modelos de sub-palavras (por exemplo, fonemas).
- Aplicação de novas técnicas para o reconhecimento automático de fala, como por exemplo, redes neuronais e a teoria dos conjuntos nebulosos (*Fuzzy Theory*).
- Utilização de técnicas híbridas. Como por exemplo, a combinação da técnica dos modelos de Markov escondidos com a teoria das redes Neuronais.

Referências

[1] K. Lee, A. G. Hauptmann, and A. Rudnicky. The spoken word. *Byte*, pages 225–232, July 1990.

[2] T. B. Martin. Practical applications of voice input to machines. *Proceedings of the IEEE*, 64(4):487–501, April 1976.

[3] M. Immendörfer. Applications for speech processing in telecommunication and office equipment. *Electrical Communication*, 60(1):71–78, 1986.

[4] L. R. Rabiner. Special issue on man-machine communication by voice. *Proc. of the IEEE*, 64(4):403–404, April 1976.

[5] H. Gu, C. Tseng, and L. Lee. Isolated-utterance speech recognition using hidden Markov models with bounded state durations. *IEEE Trans. on signal processing*, 39(8):1743–1752, August 1991.

[6] K. P. Unnikrishnan and J. J. Hopfield. Connected-digit speaker-dependent speech recognition usin neural network with time-delayed connections. *IEEE Trans. on signal processing*, pages 698–713, March 1991.

[7] M. K. Brown, M. A. Mcgee, L. R. Rabiner, and J. G. Wilpon. Training set design for connected speech recognition. *IEEE Trans. on signal processing*, 39(6):1268–1281, June 1991.

[8] X. D. Huang. Phoneme classification using semicontinuous hidden Markov models. *IEEE Trans. on Signal Processing*, 40(5), May 1992.

[9] M. R. Sambur and L. R. Rabiner. A speaker-independent digit-recognition system. *B. S. T. J.*, 54(1):81–102, January 1975.

[10] Peter Noll N. S. Jayantand. *Digital Coding of Waveforms.* Prentice-Hall, Inc., 1984.

[11] H. Ney, D. Mergel, A. Noll, and A. Paeseler. Data driven search organization for continuous speech recognition. *IEEE Trans. on Signal Recognition*, 40(2), February 1992.

[12] K. Lee, H.-W. Hon, and R. Reddy. An overview of the SPHINX speech recognition system. *IEEE Trans. on ASSP*, 38(1):35–45, January 1990.

[13] S. E. Levinson. Structural Methods in Automatic Speech Recognition. *Proc. of the IEEE*, 73(11):1625–1647, nov 1985.

[14] L. R. Rabiner, B. H. Juang, and S. E. Levinson. Recognition of isolated digits using hidden Markov models with continuous mixture densities. *AT&T Tech. J.*, 64(6):1211–1234, July 1985.

[15] A. Maheswaran and B. R. Davis. Analytical signal processing for pattern recognition. *IEEE Transations on ASSP*, 38(9):1645–1649, September 1990.

[16] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi. On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition. *B. S. T. J.*, 62(4):1075–1105, April 1983.

[17] J. G. Wilpon et al. Speech recognition: From the laboratory to the real world. *AT&T Technical Journal*, 69(5):14–24, September 1990.

[18] J. L. Flanagan and C. J. Del Riesgo. Speech processing: A perspective on the science and its applications. *AT&T Technical Journal*, 69(5):2–13, September 1990.

[19] K. Lee. Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Trans. on ASSP*, 38(4):599–609, April 1990.

[20] A. Ljolje and S. E. Levinson. Development of an acoustic-phonetic hidden Markov model for continuous speech recognition. *IEEE Trans. on Signal Processing*, 39(1):29–39, January 1991.

[21] C. H. Lee, C. H. Lin, and B. H. Juang. A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Trans. on signal processing*, 39(4):806–814, April 1991.

[22] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, February 1989.

[23] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Stat.*, 37:1554–1563, 1966.

[24] L. E. Baum and J. A. Egon. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. *Bull. Amer. Meteorol. Soc.*, 73:360–363, 1967.

[25] L. E. Baum and G. R. Sell. Growth functions for transformations on manifolds. *Pac. J. Math..*, 27(2):211–227, 1968.

[26] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique ocurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41(1):164–171, 1970.

[27] L. E. Baum. *An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov process*, volume 3, pages 1–8. 1972.

[28] J. K. Baker. The Dragon system - an overview. *IEEE Trans. on Acoust. Speech Signal Processing*, 23(1):24–29, February 1975.

[29] F. Jelinek. Continuous speech recognition by statistical methods. *Proc. IEEE*, 64:532–536, April 1976.

[30] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *B. S. T. J.*, 62(4):1035–1074, April 1983.

[31] L. R. Rabiner, B. H. Juang, and S. E. Levinson. Some properties of continuous hidden Markov model representations. *AT&T Tech. J.*, 64(6):1251–1270, July 1985.

[32] L. Deng and P. Kenny at al. Phonemic hidden Markov models with gaussian mixture output densities for large-vocabulary word recognition. *IEEE Trans. on signal processing*, 39(7):1677–1681, July 1991.

[33] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decodind algorithm. *IEEE Trans. Informat. Theory*, IT-13:260–269, April 1967.

[34] G. D. Forney. The Viterbi algorithm. *Proc. IEEE*, 61:268–278, March 1973.

[35] F. Jelinek. Continuous speech recognition by statistical methods. *Proc. IEEE*, 64:532–556, April 1976.

[36] H. Bourlard, C. J. Wellekins, and H. Ney. Connected digit recognition using vector quantization. *Proc. IEEE, ICASSP '84*, pages 26.10.1–4, March 1984.

[37] L. R. Rabiner. An algorithm for determining the endpoints of isolated utterances. *B. S. T. J.*, 54(2):297–315, February 1975.

[38] L. F. Lamel et al. An improved endpoint detector for isolated word recognition. *IEEE Trans. on ASSP*, 29(4), August 1981.

[39] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, New Jersey, 1978.

[40] B. H. Juang and L. R. Rabiner. A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, 64(2):391–408, February 1985.

[41] N. L. Pimenta. Levantamento estatístico da locução e conversação para sinais de voz. Dissertação de mestrado, Universidade Federal da Paraíba-Campus II, julho 1992.

[42] T. Yoh'ichi. A weighted cepstral distance measure for speech recognition. *IEEE Trans. on ASSP*, 35(10), October 1987.

[43] S. Furui. Speaker-independent isolated word recognition based on dynamics-emphasized cepstrum. *The transactions of the IECE of Japan*, E69(12):1310–1317, December 1986.

[44] B. P. Lathi. *An Introduction to Random Signals and Communication Theory.* Intertext Books, London, 1970.

[45] Silvana L. do N. C. Costa and B. G. Aguiar Neto. Adaptive multichannel system to speech enhancement using microphone array. *International Telecommunication Symposium (ITS'94) - Rio de Janeiro - Brazil*, August 1994.