Universidade Federal da Paraíba
Centro de Ciências e Tecnologia
Curso de Pós-Graduação em
Engenharia Elétrica

Verificação de Locutor Utilizando Modelos de Markov Escondidos (HMMs) de Densidades Discretas

Joseana Macêdo Fechine

Campina Grande - Pb

Abril - 1994

Joseana Macêdo Fechine

Verificação de Locutor Utilizando Modelos de Markov Escondidos (HMMs) de Densidades Discretas

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.
Orientador

Campina Grande, Paraíba, Brasil

F291v Fechine, Joseane Macedo
Verificacao de locutor utilizando modelos de Markov escondidos (HMMs) de densidades discretas / Joseane Macedo Fechine. - Campina Grande, 1994.
174 f. : il.

Dissertacao (Mestrado em Engenharia Eletrica) - Universidade Federal da Paraiba, Centro de Ciencias e Tecnologia.

1. Cadeia de Markov 2. Modelos de Markov Escondidos 3. Dissertacao I. Aguiar Neto, Benedito Guimaraes, Dr. II. Título

CDU 519.217.2(043)

Verificação de Locutor Utilizando Modelos de Markov Escondidos (HMMs) de Densidades Discretas

Joseana Macêdo Fechine

Dissertação de Mestrado aprovada em 04/04/1994

Benedito Guimarães Aguiar Neto - Dr. -Ing.
Orientador

Marcos Antônio Gonçalves Brasileiro - Dsc
Componente da Banca

Abraham Alcaim - PhD
Componente da Banca

Campina Grande, Paraíba, Brasil, abril/1994

Dedico este trabalho a Deus em primeiro lugar, aos meus pais, José e Ana Ildaísa e aos meus irmãos, Vicente, Geovane e Guilhermino.

Vale lembrar o adágio chinês a LauTzu (600 A.C.):
"A mais longa jornada não se inicia senão com um simples passo".

Agradecimentos

A realização deste trabalho recebeu o apoio de muitas pessoas, cuja colaboração gostaria de agradecer. Assim, devo os meus sinceros agradecimentos a algumas pessoas e instituições. Dentre elas:

Ao professor Benedito Guimarães Aguiar Neto, pelo esforço, estímulo e dedicação prestados.

À professora Rosângela Maria Vilar França, pela experiência e sugestões.

Aos colegas Cortez, Rinaldo, Gustavo, Mônica, Wallington, Marony, Silvana Porto, Ana Lúcia, Aldenor, Edjander, Eliane, Moreira, Washington, Silvana Cunha, Socorro e demais colegas do LAPS.

Às minhas grandes amigas Kátia, Magna e Kíssia, que tanto me apoiaram nas horas difíceis.

A toda minha família e aos meus amigos, que tanto me incentivaram no decorrer deste trabalho.

A Universidade Federal da Paraíba-Campus II e ao CNPq.

E à cidade de Campina Grande, que me abriga tão carinhosamente.

Resumo

Os Modelos de Markov Escondidos (HMMs) vêm se tornando cada vez mais populares por serem muito ricos em estrutura matemática e, consequentemente formarem uma base teórica muito forte para uso em um largo grupo de aplicações na área de processamento de sinais de voz. Apresentam em geral, uma redução do custo computacional em comparação com métodos mais tradicionais.

O reconhecimento de locutor utilizando HMMs, como toda tarefa de reconhecimento de padrões, se divide em duas fases distintas: treinamento e classificação. Na fase de treinamento, inicialmente é realizada a análise do sinal de voz de forma a se obterem os parâmetros representativos deste locutor. Foram usados, neste trabalho, os coeficientes de Predição Linear (coeficientes LPC), os quais foram representados por um alfabeto discreto obtido através da quantização vetorial. O HMM associado ao locutor é obtido através do algoritmo de reestimação de Baum-Welch, que consiste em uma técnica iterativa que fornece, através do cálculo de uma medida de probabilidade, o modelo que melhor representa o dado locutor.

A fase de classificação, no caso, de verificação de locutor, consiste no cálculo da probabilidade associada ao modelo de referência já armazenado para o locutor a ser verificado. Se o valor de probabilidade calculado é maior que um dado limiar, o locutor é considerado verdadeiro, caso contrário o locutor é considerado impostor.

Abstract

Hidden Markov Models (HMMs) are becoming popular in pattern recognition because they present a strong mathematical structure solid and so they provide a theorectical basis for very many applications in voice processing systems. They can also provide a reduction in complexity when compared to other methods.

Speaker recognition using HMMs, like other pattern recognition techniques, can be performed in two phases: training and classification.

For the training phase each speaker uses a individual HMM. The model is built after the speech has beeen analised and the Linear Predictive Coding (LPC) parameters representing that particular speaker have been obtained. The LPC coeficients are then discretizes by a vector quantizer. The discretized parameters are used for running an iteractive algorithm (Baum-Welch algorithm) calculating a probability wich best represents that speaker.

Classification, in this application, speaker verification, consists in using the HMM obtained in the training phase to calculate and check whether that particular speaker provides an acceptable probability to be considered a valid user (customer).

Sumário

Lista de Tabelas

Lista de Figuras

Lista de símbolos

1. F_0 - freqüência fundamental;
2. P - período fundamental ;
3. t - tempo(s);
4. $\Delta(t)$ - trem de pulsos glotais;
5. $A_s(t)$ - controle da amplitude dos pulsos glotais;
6. $A_f(t)$ - controle da amplitude do ruído;
7. $s(t)$ - sinal de voz;
8. $g(t)$ - resposta ao impulso do filtro conformador (formato do pulso glotal);
9. $n(t)$ - fonte de ruído branco;
10. $u(t)$ - excitação;
11. $f(t)$ - resposta ao impulso do trato vocal;
12. $z(t)$ = resposta ao impulso da radiação;
13. f - freqüência (Hz);
14. $\Omega - 2\pi f$, freqüência angular (rad/s);
15. $U(\Omega)$ - espectro da excitação;

16. $F(\Omega)$ - resposta em freqüência do trato vocal (função de transferência do aparelho fonador);

17. $Z(\Omega)$ - impedância de carga dos lábios e/ou narinas;

18. $\delta_P(t)$ - seqüência periódica de impulsos com período P;

19. $D(\Omega)$ e $N(\Omega)$ - são polinômios cujas raízes correspondem respectivamente aos pólos e zeros do aparelho fonador;

20. F_1, F_2 e F_3 - Três primeiras freqüências formantes;

21. R - resistência;

22. L - impedância;

23. a - raio da circunferência, cuja área é igual a área de abertura dos lábios ou narinas;

24. c - a velocidade do som;

25. f_s - freqüência de amostragem;

26. $\Delta(n)$ - representação discreta de $\Delta(t)$;

27. $A_s(n)$ - representação discreta de $A_s(t)$;

28. $A_f(n)$ - representação discreta de $A_f(t)$;

29. $s(n)$ - representação discreta de $s(t)$;

30. $g(n)$ - representação discreta de $g(t)$;

31. $n(n)$ - representação discreta de $n(t)$;

32. $f(n)$ - representação discreta de $f(t)$;

33. $z(n)$ - representação discreta de $z(t)$;

34. $u(n)$ - representação discreta de $u(t)$;

35. $\delta_p(n)$ - representação discreta de $\delta_p(t)$;

36. T - período de amostragem (s) ou número de vetores em uma seqüência de observação;

37. s_k, s_k^* - freqüência ressonante complexa do trato vocal;

38. σ_k e F_k - partes real e imaginária da freqüência de ressonância do trato vocal, respectivamente;

39. Θ_k - ângulo no plano-z;

40. $H(z)$ - transformada-z da função de transferência do modelo que representa a geração do sinal de voz;

41. $G(z)$ - transformada-z de $g(n)$;

42. $F(z)$ - transformada-z de $f(n)$;

43. $Z(z)$ - transformada-z de $z(n)$;

44. Z_0 - valor inicial da impedância de carga dos lábios e/ou narinas;

45. $U(z)$ - transformada-z de $u(n)$;

46. G - parâmetro de ganho de $H(z)$;

47. c_k, $1 \leq k \leq p$ - coeficientes de $H(z)$ (coeficientes de predição linear);

48. $e(n)$ - erro de predição;

49. $\tilde{s}(n)$ - sinal $s(n)$ após a predição;

50. N_A - comprimento da janela que contém um segmento de voz;

51. $v(n)$ - $s(n)$ ponderado pela janela;

52. p - número de amostras passadas do sinal utilizadas na combinação linear, ordem do preditor;

53. $\tilde{v}(n)$ - sinal $v(n)$ após a predição;

54. $E(n)$ - erro quadrático de predição;

55. $R_r(k)$ - função de autocorrelação a curto prazo;

56. $C(z)$ - polinômio de grau M_A.

57. $m = w(n)$ - função de alinhamento;

58. K- dimensão do quantizador vetorial;

59. M - número de níveis do quantizador vetorial (número de símbolos do alfabeto discreto, número de símbolos distintos observados no estado);

60. x - vetor de entrada do quantizador vetorial;

61. $\tilde{x}$ - vetor de reprodução do quantizador vetorial;

62. Y - alfabeto de reprodução do quantizador vetorial;

63. y_i - vetor de reprodução (vetor código);

64. $q(x)$ - quantizador vetorial de x;

65. S - partição do espaço vetorial;

66. C_i - número de células do espaço K-dimensional;

67. $d(x, \hat{x})$ - medida de distorção;

68. q_t - estado do HMM no instante t;

69. $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$ - parâmetros que caracterizam um HMM, para um l-ésimo locutor;

70. N - número de estados do HMM;

71. $\mathcal{A} = [a_{ij}]$, $1 \leq i, j \leq N$ - matriz transição de estados do HMM;

72. $\mathcal{B} = [b_j(k)]$, $1 \leq j \leq N$ e $1 \leq k \leq M$ - função de probabilidade das observações;

73. $\pi = \{\pi_i\}$, $1 \leq i \leq N$ - vetor de probabilidade do estado inicial;

74. $\mathbf{O}^l = \{O_1, O_2, ...O_T\}$ - seqüência de vetores de observação no intervalo $[1, T]$ para um l-ésimo locutor;

75. x - variável aleatória;

76. $\mathrm{F}(x_n, x_{n-1}, ..., x_1)$ - Função Distribuição de probabilidade da seqüência de variáveis aleatórias;

77. x_N - uma variável aleatória do tipo discreto;

78. $P\{x_n = k_n / x_{n-1}\}$ - expressão da probabilidade condicional de x_n;

79. k_n - seqüência de eventos;

80. $\{V_1, ..., V_K, ..., V_M\}$ - conjunto de M resultados esperados do modelo (também chamado de conjunto de símbolos ou alfabeto);

81. $Q = \{q_1, ..., q_T\}$ - seqüência de estados percorrida pelo modelo no intervalo de tempo $[1, T]$.

82. $c_t(m)$, $1 \leq m \leq p$ - coeficiente LPC, para um dado instante t;

83. $\hat{c}_t(m)$, $1 \leq m \leq p$ - componetes do vetor de um dicionário do quantizador vetorial, para um dado instante t;

84. $d(\hat{c}_t, c_t)$ - a distância entre dois vetores LPC, c_t e $\hat{c}_t$, para um dado instante t;

85. $||D_M||$ - medida de distorção (distância) do quantizador vetorial;

86. $P_l(\mathbf{O}^l / \lambda_l)$ - cálculo da probabilidade associada ao HMM referente ao l-ésimo locutor;

87. $\alpha_t(i)$ - probabilidade de avanço ("forward probability");

88. $\beta_t(i)$ - probabilidade de retrocesso ("backward probability");

89. $\overline{a_{ij}}$ - valor reestimado de a_{ij};

90. $\overline{b_j(k)}$ - valor reestimado de $b_j(k)$;

91. $c[i]$ - classe i ocupada por um vetor quantizado;

92. ϵ - limiar de comparação e ajuste;

93. R_{bj} - número de $b_j(k)s$ modificados para um dado j;

94. esc_t - coeficiente de escalonamento;

95. Kl - número de repetições da sentença de um dado locutor;

96. / / - As duas barras indicam a pronúncia de fonemas ou palavras. A palavra "som", por exemplo, pode ser representada por /som/.

Lista de abreviaturas

1. LPC - Linear Predictive Coding (Codificação por Predição Linear);
2. DTW - Dinamic Time Warping (Alinhamento Dinâmico no Tempo);
3. QV - Quantização Vetorial;
4. CELP - Code-Excited Linear Predictive;
5. HMM - Hidden Markov Model (Modelo de Markov Escondido);
6. RAL - Reconhecimento Automático de Locutor;
7. L1, L2, L3, L4, L5 e L6 - locutores 1, 2, 3, 4, 5 e 6, respectivamente;
8. **a** - aceita o locutor;
9. **r** - repete a sentença;
10. **na** - não aceita o locutor;

Capítulo 1

Introdução

1.1 Comunicação Vocal Homem-Máquina

A voz é o meio mais natural de comunicação do homem. Quando duas pessoas estão conversando, descobrimos com facilidade a idade, sexo e se a língua que está sendo falada é de nosso conhecimento.

A partir, unicamente da voz, somos capazes de identificar uma série de características de uma pessoa, tais como, seu grupo sócio-cultural, seu estado emocional, seu estado de saúde, a região onde mora (através do sotaque) e uma larga quantidade de outras características.

Torna-se claro portanto, que a partir do sinal de voz é possível distinguir as características de cada pessoa. Partindo-se desse princípio, o homem procurou desenvolver equipamentos que permitissem, através da voz, a sua comunicação com as máquinas.

Com o desenvolvimento tecnológico foram surgindo uma série de equipamentos eletrônicos de uso doméstico, com o objetivo de melhorar a qualidade de vida do homem moderno. Tais equipamentos, embora sofisticados, enfrentam ainda dificuldades quanto a sua utilização, devido a forma artificial com que o usuário deve interagir com os mesmos. Assim, parece claro que o desenvolvimento de uma interface vocal, tornaria mais fácil e produtiva a relação Homem-Máquina [1, 2].

Os primeiros trabalhos descrevendo máquinas que podiam, de alguma forma, reconhecer com certo sucesso a pronúncia de determinadas palavras datam de 1952 [3]. Uma grande quantidade de trabalhos no assunto surgiram nos anos 60, a nível de laboratório, graças as descobertas de algumas propriedades da voz através do uso de espectógrafos [4] e das novas facilidades que os computadores digitais vieram oferecer.

Em seguida, verificou-se a necessidade de desenvolver máquinas capazes não só de entender o que estava sendo dito, mas de responder ao que lhe era perguntado. Os esforços iniciais para construção de máquinas falantes datam do final do séc. XVIII, quando foram elaborados curiosos engenhos acústicos que produziam sons semelhantes à voz e eram "tocados" à maneira de um instrumento musical [5].

A comunicação vocal entre pessoas e máquinas inclui síntese de voz para texto, reconhecimento automático de voz (conversão voz-texto), e o reconhecimento de locutores a partir de suas vozes. Portanto, a comunicação vocal Homem-Máquina se divide nas seguintes sub-áreas principais [6]:

1. Sistema de Resposta Vocal
2. Sistemas de Reconhecimento de Fala
3. Sistemas de Reconhecimento de Locutor

Sistemas de resposta vocal são projetados para responder a um pedido de informação utilizando mensagens faladas. Assim, a comunicação de voz em sistemas de resposta vocal se faz em uma única direção, isto é, da máquina para o homem [6].

As áreas 2 e 3 realizam a comunicação vocal do homem para a máquina. O reconhecimento de fala, pode ser subdividido em um largo número de sub-áreas dependendo de alguns fatores, tais como, tamanho do vocabulário, população de locutores, etc. A tarefa básica no reconhecimento de fala é reconhecer uma determinada elocução de uma sentença ou "entender" um texto falado (i.é., responder de forma correta ao que está sendo falado) [6].

Dado um sinal de voz de entrada, o objetivo do reconhecimento de locutor é identificar a pessoa mais provável de ser o locutor (dentre uma população conhecida) -

Identificação de Locutor, ou verificar se o locutor é quem ele alega ser - **Verificação de Locutor** [6].

A Figura 1.1 mostra a representação geral de um problema de reconhecimento de locutor [7].

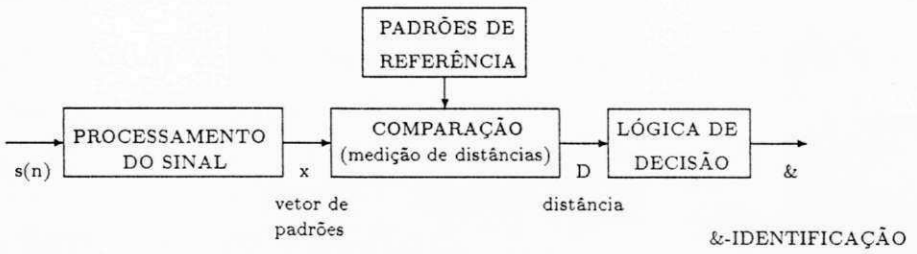


Figura 1.1: Representação geral do problema de reconhecimento de locutor

Esses sistemas desempenham as seguintes funções:

1. Verificação de locutor - Comparação com um único padrão pré-estabelecido.
2. Identificação de locutor - Comparação com todos os padrões pré-estabelecidos.

1.2 Reconhecimento de Locutor

A identificação de pessoas a partir de suas vozes é um problema científico de grande importância e possui aplicações em diversas áreas. Uma das aplicações imediatas é o seu uso para o controle de acesso a algum ambiente restrito pelo uso da senha verbal. Outra aplicação está ligada à criminalística, com o mesmo propósito que hoje é dado às impressões digitais.

A identificação da voz tem a conveniência da facilidade de coleção de dados. Outra vantagem dessa técnica quando comparada com exame de fundo de olho, impressões

digitais e assinaturas, se refere à sua facilidade de utilização em sistemas onde se exige o reconhecimento à distância; por exemplo transações bancárias por telefone. Além disso, a voz não pode ser perdida nem tão pouco esquecida [8].

O processo de reconhecimento de identidade vocal consiste na extração de parâmetros de voz de um dado locutor de forma a definir um modelo que preserve as suas características vocais que o diferenciam de outros indivíduos.

Este trabalho trata da **Verificação de Locutor**, a qual consta de duas fases: fase de treinamento e fase de verificação.

FASE DE TREINAMENTO (Figura 1.2)

1. Define uma identificação prévia para o usuário: Ex. Senha numérica dada através de um teclado.
2. Extrai parâmetros da fala a partir da sentença de teste.
3. Define e armazena padrões de referência para a identidade vocal do usuário.

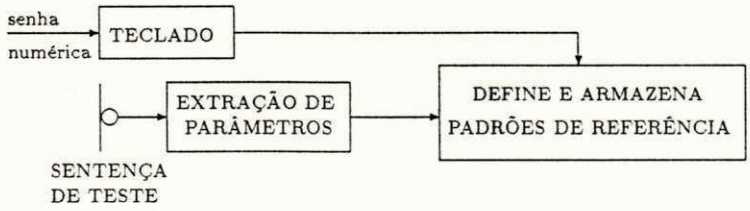


Figura 1.2: Fase de treinamento

FASE DE VERIFICAÇÃO (Figura 1.3)

1. Indica identificação pré-estabelecida.

2. Pronuncia a sentença de teste.

3. Extrai parâmetros da fala e calcula padrão de identidade vocal.

4. Compara se a identidade pertence ao locutor a partir da comparação do padrão calculado com o padrão de referência.

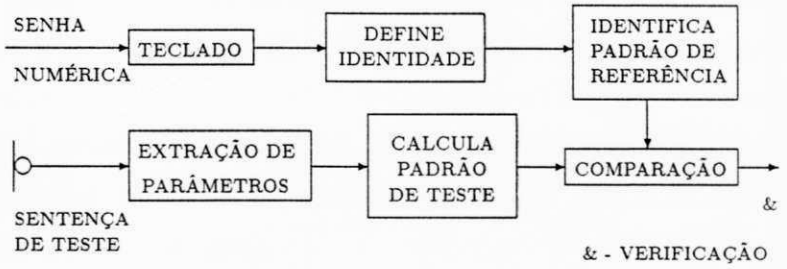


Figura 1.3: Fase de verificação

Várias são as técnicas utilizadas para reconhecimento de identidade vocal (reconhecimento de locutor) [7]. Dentre elas podem ser citadas: Alinhamento Dinâmico no Tempo (DTW) [7], Coeficientes de Predição Linear (LPC) [9], Quantização Vetorial (QV) [10], Modelos de Markov Escondidos (HMMs) [11], dentre outras.

Apesar do sucesso obtido com a maioria dessas técnicas, o uso de Modelos de Markov Escondidos vêm se tornando cada vez mais popular em sistemas de reconhecimento de voz e locutor devido a algumas vantagens. Em primeiro lugar, os HMMs são muito ricos em estrutura matemática e, consequentemente podem formar uma base teórica muito forte para uso em um largo grupo de aplicações, p.ex., modelagem do sinal de voz e capacidade de solucionar problemas mais difíceis, como por exemplo, o reconhecimento de locutor em sistemas independentes do texto. Segundo, quando aplicados apropriadamente, trabalham muito bem para várias aplicações práticas. Além disso, apresentam uma redução do custo computacional em comparação com outros métodos (p.ex. DTW) [11, 12].

Para a construção do HMM referente a cada locutor, inicialmente são calculados os coeficientes LPC, que serão os parâmetros representativos do sinal de voz. Em seguida, a quantização vetorial é requerida para representar os vetores de observação (coeficientes LPC) pelos símbolos correspondentes (obtendo-se assim, uma compressão de dados). Por fim, cada locutor será representado por um HMM, obtido a partir desse conjunto de símbolos associados a cada vetor de observação [11, 13].

O objetivo do presente trabalho foi desenvolver uma técnica de verificação de locutor dependente do texto, ou seja, a sentença usada para treinamento é a mesma usada para verificação, utilizando Modelos de Markov Escondidos (HMMs). Os resultados obtidos utilizando HMM foram, de forma geral, muito bons, apresentando baixos índices de falsa rejeição (locutor verdadeiro é considerado impostor) e de falsa aceitação (impostor é considerado locutor verdadeiro).

Os testes de desempenho do sistema foram levados a efeito utilizando 6 locutores. Os erros de verificação observados no sistema são provocados em parte pelas limitações do modelamento da identidade vocal por HMM e em parte pela presença de ruídos ambientais.

1.3 Organização da dissertação

No Capítulo 2 são descritos os parâmetros necessários para representação e modelagem dos sinais de voz.

No Capítulo 3 são apresentadas as técnicas mais usuais de reconhecimento de locutor e quais os motivos que levaram à escolha de Modelos de Markov Escondidos (HMMs) como técnica utilizada nesse trabalho.

O Capítulo 4 é dedicado a descrição do Modelo de Markov Escondido de densidades discretas especificando os parâmetros necessários à modelagem dos sinais de voz correspondentes a cada locutor, para sua posterior verificação.

No Capítulo 5 são apresentados e avaliados os resultados experimentais, obtidos a partir da implementação do algoritmo do HMM.

No 6º e último capítulo é feito um comentário geral a respeito do método proposto e avaliação das perspectivas que estão surgindo no contexto de Reconhecimento Automático de Locutor.

Capítulo 2

Extração de parâmetros para reconhecimento de locutor

2.1 Introdução

De forma semelhante à sistemas de reconhecimento de fala, existem vários parâmetros que podem caracterizar um dado locutor. Esses parâmetros representam as características da fala, as quais são extraídas a partir dos sinais de voz. Para o estudo e avaliação das características da fala, de forma a utilizá-las no modelamento de uma identidade vocal para um dado locutor, se faz necessário um conhecimento do processo de produção da fala e o seu modelo correspondente (analógico e digital), bem como um estudo dos aspectos acústicos que possam diferenciar indivíduos.

2.2 O mecanismo de produção da voz

Sinais de voz são compostos de uma seqüência de sons. Esses sons e a transição entre eles serve então como uma representação simbólica da informação. A combinação desses sons (símbolos) é governada por regras de linguagem. O estudo dessas regras e suas aplicações na comunicação humana é do domínio da lingüística, e o estudo e

classificação dos sons de voz é chamado fonética [6].

Para gerar o som desejado, o locutor exerce uma série de controles sobre o aparelho fonador, representado nas Figuras 2.1 e 2.2, produzindo a configuração articulatória e a excitação apropriadas [5]. A Figura 2.1 evidencia as características importantes do sistema vocal humano [5]. O trato vocal começa na abertura entre as cordas vocais, ou glote e termina nos lábios. O trato vocal assim, consiste da faringe (a conexão entre o esôfago e a boca) e termina na boca ou cavidade oral. Para homens adultos, o trato vocal possui, em média, 17cm. A área da seção transversal do trato vocal, determinada pelas posições da língua, dos lábios, maxilar e úvula varia de 0 (completamente fechado) a até aproximadamente $20cm^2$. O trato nasal começa na úvula e termina nas narinas. Quando a úvula é abaixada, o trato nasal é acusticamente acoplado ao trato vocal para produzir os sons nasais da voz. Verifica-se que a forma do trato nasal, não pode ser modificada voluntariamente pelo locutor. Após a filtragem, determinada pela conformação do aparelho fonador, o fluxo de ar injetado pelos pulmões é acoplado ao ambiente externo através dos orifícios dos lábios e/ou narinas [6].

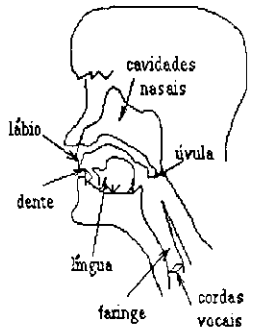


Figura 2.1: Anatomia do aparelho fonador

No estudo dos processos de produção da voz, é útil abstrair as características importantes do sistema físico que conduzem ao modelamento matemático. A Figura 2.3 mostra o diagrama esquemático do trato vocal. O diagrama completo inclui o sistema

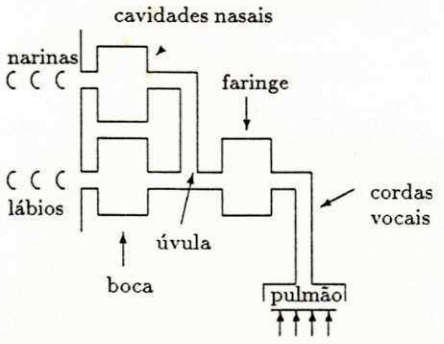


Figura 2.2: Modelo acústico do aparelho fonador

sub-glotal composto dos pulmões, brônquios e traquéia. O sistema sub-glotal funciona como uma fonte de energia para produção da voz. A voz é simplesmente a onda acústica radiada do sistema quando o ar é expelido dos pulmões [6].

O trato vocal e o trato nasal são mostrados na Figura 2.3 como tubos de seção transversal não uniforme. O som se propaga através desses tubos, o espectro de freqüência é modelado pela seletividade de freqüência do tubo. Este efeito é muito similar aos efeitos de ressonância observados em instrumentos de sopro. No contexto da produção da voz, as freqüências de ressonância do tubo do trato vocal são chamadas freqüências formantes ou simplesmente formantes. As freqüências formantes dependem sobretudo da forma e dimensões do trato vocal, cada forma é caracterizada por um conjunto de freqüências formantes. Sons diferentes são formados variando a forma do trato vocal. Assim, as propriedades espectrais do sinal de voz variam com o tempo e com a forma do trato vocal [6, 14].

Os sons da voz podem ser classificados dentro de 3 classes distintas de acordo com o modo de excitação. As classes são as seguintes [6]: sons sonoros, sons surdos e sons explosivos.

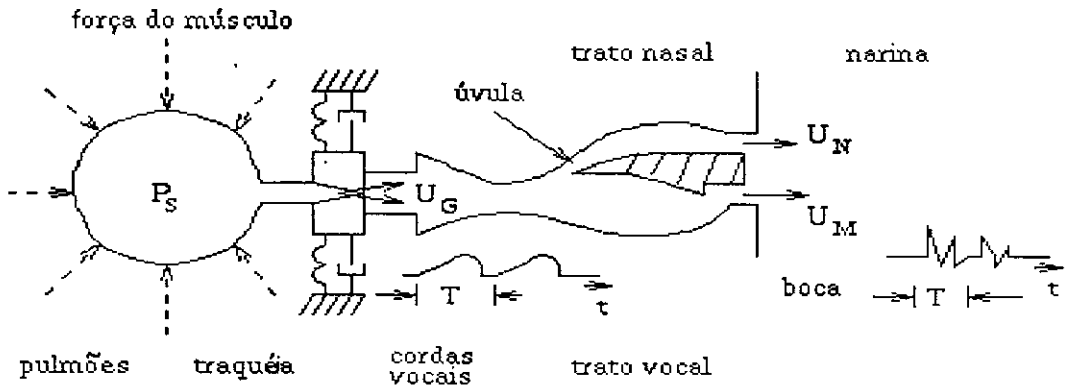


Figura 2.3: Diagrama esquemático do aparelho do trato vocal

2.2.1 Sons Sonoros

O fluxo de ar vindo dos pulmões é controlado pela abertura e fechamento das cordas vocais, ou melhor dizendo, dobras vocais que são ligamentos semelhantes a dois lábios que podem ser tensionados e aproximados sob o controle do locutor. A abertura entre as dobras é denominada glote. Estando a glote completamente fechada, o fluxo de ar vindo dos pulmões é interrompido e a pressão sub-glótica aumenta até que as dobras vocais sejam separadas, liberando o ar pressionado, gerando um pulso de ar de curta duração. Com o escoamento do ar, a pressão glótica é reduzida, possibilitando uma nova aproximação das cordas vocais. O processo se repete de forma quase periódica. Desta forma, são obtidas ondas de pressão, quase periódicas, excitando o trato vocal, que atuando como um ressonador modifica o sinal de excitação, produzindo freqüências de ressonância denominadas de formantes que caracterizarão os diferentes sons sonoros [6].

As vogais /a/ e /i/ (Figuras 2.4 e 2.5), cujo grau de nasalização é determinado pelo abaixamento da úvula, são exemplos típicos de sons sonoros.

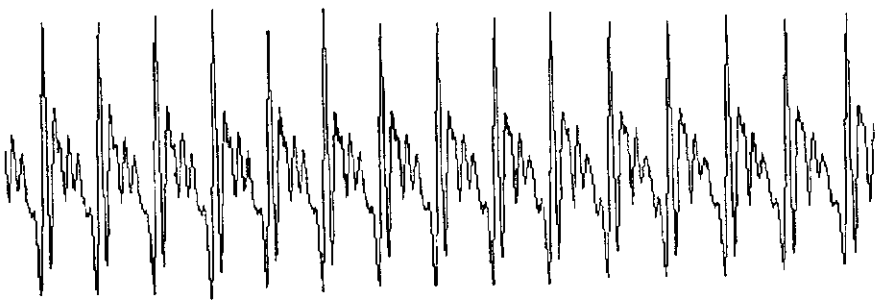


Figura 2.4: Forma de onda da vogal não nasalizada /a/.

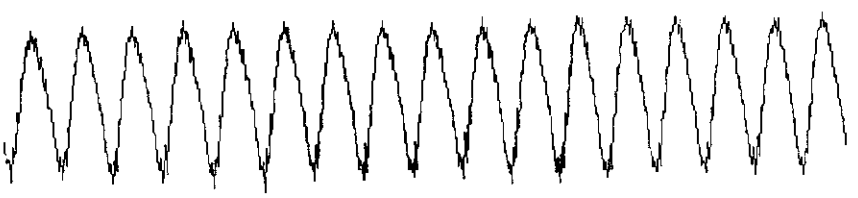


Figura 2.5: Forma de onda da vogal não nasalizada /i/.

Algumas consoantes, como /l/ (Figura 2.6) e /m/, também são produzidas com a excitação glotal.

A freqüência média dos pulsos é denominada freqüência fundamental de excitação, F_0 e o período fundamental, P, é dado por

$$P = \frac{1}{F_0} \tag{2.1}$$

A frequência fundamental dos sons sonoros fica entre 80 Hz (para homens) e 350 Hz (para crianças), sendo 240 Hz um valor típico para mulheres [15].

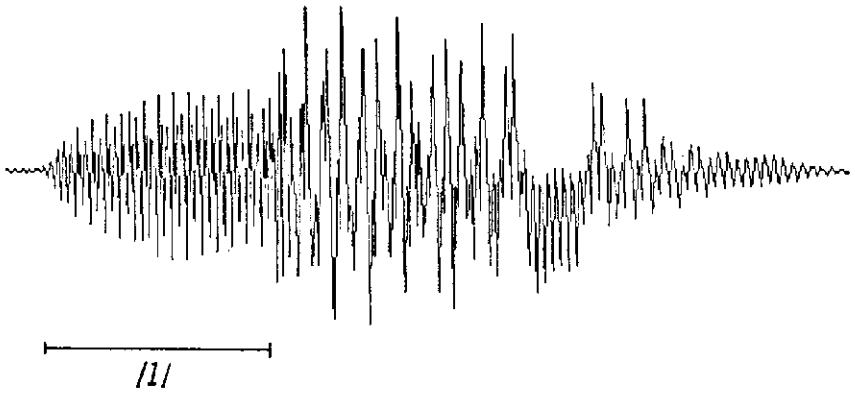


Figura 2.6: Forma de onda da sílaba /lá/.

Tendo em vista a pequena abertura da glote em relação às cavidades superiores do aparelho fonador, considera-se que a vazão glótica não é influenciada pelos movimentos dos articuladores. Ou seja, o sistema glotal pode ser visto como uma fonte de corrente de alta impedância acoplada ao trato vocal. Fazendo uma analogia com a eletricidade, a pressão corresponde à tensão e a vazão à corrente [5].

2.2.2 Sons Surdos

Os sons surdos são gerados pela produção de uma constrição em algum ponto do trato vocal (usualmente próximo ao final da boca), assim o ar adquire velocidade suficientemente alta para produzir turbulência gerando uma fonte de ruído de espectro largo (semelhante ao ruído branco) para excitar o trato vocal.

Na produção desses sons a glote permanece aberta, não havendo vibração das cordas vocais. Por exemplo, na produção do /f/ (Figura 2.7), lábios e dentes são ligeiramente pressionados, deixando assim uma passagem estreita para o ar, produzindo um fluxo de ar turbulento nas imediações da constrição, o qual excita as cavidades do trato vocal. O som produzido desta forma tem características ruidosas com concentração relativa de energia nas mais altas componentes de freqüência do espectro de sinais de voz [6, 14].

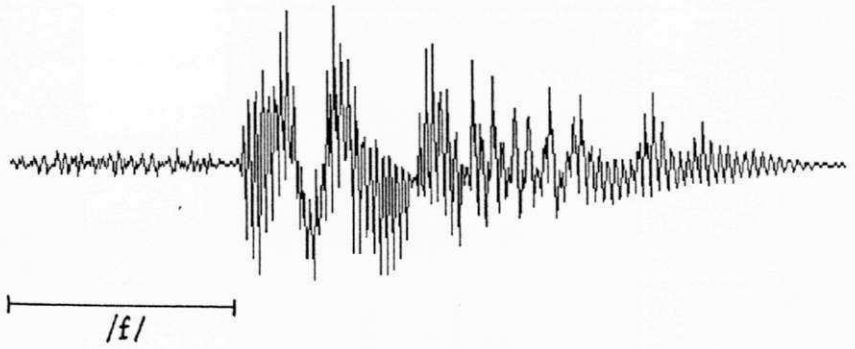


Figura 2.7: Forma de onda da sílaba /fá/.

2.2.3 Sons Explosivos

Na geração dos sons explosivos, o ar é totalmente dirigido à boca, estando esta completamente fechada. Com o aumento da pressão, a oclusão é rompida bruscamente, gerando um pulso que excita o aparelho fonador. Com a excitação ocorre um movimento rápido dos articuladores em direção à configuração do próximo som. Exemplos de sons explosivos são os fonemas /p/ (Figura 2.8), /t/, /k/, dentre outros [6].

Figura 2.8: Forma de onda da sílaba /pé/.

2.2.4 Sons com excitação mista

Os sons fricativos sonoros, como /j/, /v/ (Figura 2.9) e /z/, são produzidos combinando-se a vibração das cordas vocais e a excitação turbulenta. Nos períodos em que

a pressão glótica atinge um valor máximo, o escoamento através da obstrução torna-se turbulento, gerando o caráter fricativo do som; quando a pressão glótica cai abaixo de um dado valor, termina o escoamento turbulento de ar e as ondas de pressão apresentam um comportamento mais suave.

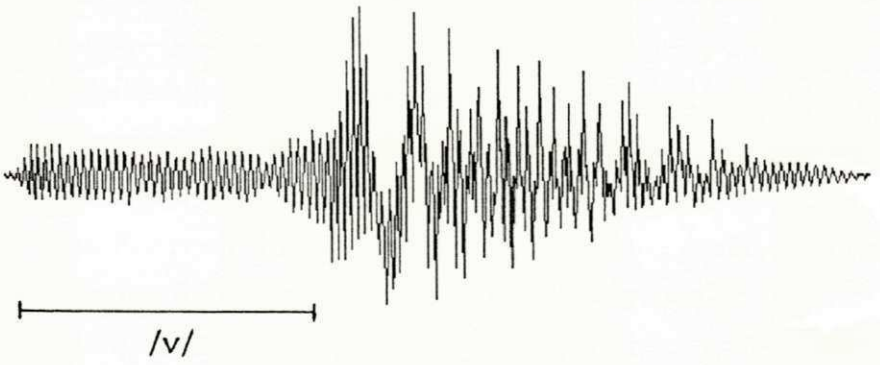


Figura 2.9: Forma de onda da sílaba /vá/.

Os sons oclusivos (ou explosivos) sonoros, como /d/ e /b/ (Figura 2.10), são produzidos de forma semelhante aos correspondentes não sonoros, /p/ e /t/, porém há vibração das cordas vocais durante a fase de fechamento da cavidade oral [14].

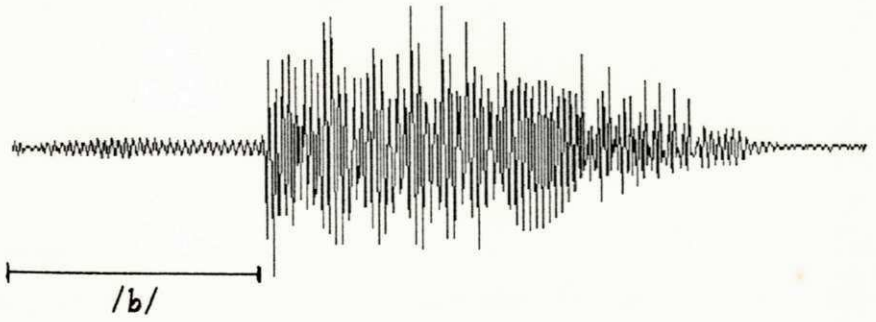


Figura 2.10: Forma de onda da sílaba /bé/.

2.3 Modelos para produção da voz

Ondas sonoras são criadas pela vibração e se propagam no ar ou em outro meio pela vibração das partículas do meio. Assim, os processos físicos são a base para a descrição da geração e propagação do som no sistema vocal. Em particular, a lei fundamental da conservação da massa, conservação do momento, e conservação da energia juntamente com as leis da termodinâmica e mecânica dos fluídos, todas se adaptam a compressibilidade, baixa viscosidade do fluído (ar) que é o meio de propagação do som pela voz. Usando esses princípios físicos, um conjunto de equações diferenciais parciais pode ser obtido para descrever o movimento do ar no sistema vocal. A formulação e solução dessas equações é extremamente difícil exceto sobre suposições simples em torno da configuração do trato vocal e perdas de energia no sistema vocal. Uma teoria acústica detalhada deve considerar os efeitos das seguintes características [5, 6, 14]:

1. Variação da configuração do trato vocal com o tempo;
2. Perdas próprias por condução de calor e fricção nas paredes do trato vocal;
3. A maciez das paredes do trato vocal;
4. Radiação do som pelos lábios;
5. Junção nasal;
6. Excitação do som no trato vocal, etc.

Um modelo detalhado para geração de sinais de voz, que leva em conta os efeitos da propagação e da radiação conjuntamente pode, em princípio, ser obtido através de valores adequados para excitação e parâmetros do trato vocal. A teoria acústica sugere uma técnica simplificada para modelar sinais de voz, a qual é bastante utilizada como base em inúmeros modelos para síntese de voz, mostrada em diagrama de blocos na Figura 2.11. Essa técnica apresenta a excitação separada do trato vocal e da radiação. Os efeitos da radiação e o trato vocal são representados por um sistema linear variante com o tempo. O gerador de excitação gera um sinal similar a um trem de pulsos

(glotal), ou sinal aleatório (ruído). Os parâmetros da fonte e sistema são escolhidos de forma a obter na saída um sinal de voz desejado [6]. Este modelo será discutido com mais detalhes em seguida.

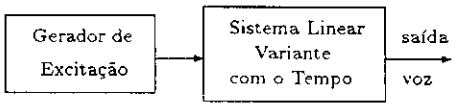


Figura 2.11: Modelo para produção da voz

2.3.1 Modelo analógico para produção da voz

Como visto anteriormente, é possível obter representações matemáticas para produção da voz. É importante conhecer as características básicas do sinal de voz e verificar como essas características são relacionadas com grandezas físicas para produção da voz. Foi visto que os sons de voz são gerados de 3 formas, e que cada uma gera uma saída diferente. Verificou-se também que o trato vocal gera freqüências de ressonância sobre a excitação de forma a produzir os diferentes sons da voz.

Um modelo analógico para produção da voz, pode ser representado pela Figura 2.12. Onde $\delta_P(t)$ é uma seqüência periódica de impulsos com período P, $g(t)$ é a resposta ao impulso do filtro conformador (formato do pulso glotal), $\Delta(t)$ = trem de pulsos glotais, $n(t)$ é uma fonte de ruído branco, $A_s(t)$ = controle da amplitude dos pulsos glotais, $A_f(t)$ = controle da amplitude do ruído, $u(t)$ representa a excitação, $f(t)$ = resposta ao impulso do trato vocal, $z(t)$ = resposta ao impulso da radiação e $s(t)$ é o sinal de voz.

De acordo com as considerações anteriores, o espectro do sinal de voz, $S(\Omega)$, $\Omega = 2\pi f$[rad/s], pode ser dado por:

$$S(\Omega) = U(\Omega).F(\Omega).Z(\Omega) \tag{2.2}$$

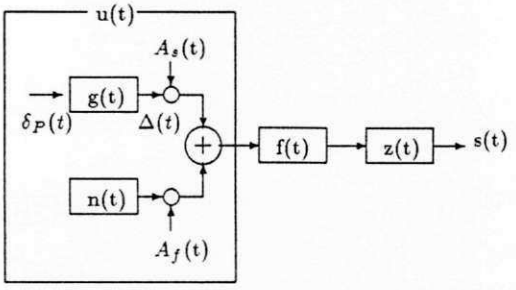


Figura 2.12: Modelo analógico para a produção da voz

onde,

$U(\Omega)$ = espectro da excitação;

$F(\Omega)$ = resposta em freqüência do trato vocal;

$Z(\Omega)$ = impedância de carga dos lábios e/ou narinas.

2.3.1.1 Excitação

O bloco referente à excitação, u(t) (Figura 2.12), está dividido em duas partes, correspondentes ao modelamento da vibração das cordas vocais e ao fluxo turbulento.

Os fonemas fricativos sonoros são produzidos com a combinação das fontes de excitação sonora e turbulenta, cujas intensidades são controladas por $A_s(t)$ e $A_f(t)$, respectivamente. A explosão dos fonemas oclusivos não é incorporada explicitamente no modelo pois, com boa aproximação, ela pode ser representada por um ruído aleatório de curta duração [5].

O trem de pulsos glotais, $\Delta(t)$, é modelado pela convolução

$$\Delta(t) = \delta_P(t) * g(t) \tag{2.3}$$

onde

$$\delta_P(t) = \sum_{k=0}^{\infty} \delta(t - kP), \tag{2.4}$$

é uma seqüência periódica de impulsos com período P; e g(t) é a resposta ao impulso de um filtro conformador, que responde pelo formato do pulso glotal. O pulso glotal tem a forma semelhante a uma onda dente de serra, cujo tempo de subida é maior que o tempo de descida. Por simplificação, assume-se que o filtro possui resposta ao impulso triangular, como mostra a Figura 2.13.

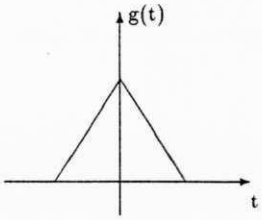


Figura 2.13: Resposta ao impulso do filtro conformador

$$g(t) = \begin{cases} 1 + t/\tau, & -\tau \leq t \leq 0 \\ 1 - t/\tau, & 0 \leq t \leq \tau \end{cases} \tag{2.5}$$

A excitação turbulenta, n(t), é simulada por uma fonte de ruído branco, com densidade de probabilidade uniforme, média nula, variância unitária e descorrelacionada com as outras variáveis do modelo [5, 6].

2.3.1.2 Trato Vocal

De forma geral, a função de transferência do aparelho fonador, $F(\Omega)$, é dada por

$$F(\Omega) = \frac{N(\Omega)}{D(\Omega)} \tag{2.6}$$

onde $D(\Omega)$ e $N(\Omega)$ são polinômios cujas raízes correspondem respectivamente aos pólos e zeros do aparelho fonador. O grau desses polinômios é determinado pelo som a ser modelado.

Para sons vocálicos não nasalizados, $F(\Omega)$ apresenta apenas pólos, simplificando a análise. Entretanto, para sons surdos, as cavidades anteriores à constrição são representadas por um pólo, assim como a própria constrição. As cavidades posteriores são representadas por um zero.

Resultados simples e eficazes podem ser obtidos a partir do estudo de modelos contendo apenas pólos.

No estudo dos vocálicos, os picos de freqüência do trato vocal, ou seja, a envoltória do espectro, recebem a denominação de formantes sendo os três primeiros formantes F_1, F_2 e F_3 , utilizados no reconhecimento das vogais [6].

2.3.1.3 Radiação

Na realidade, o tubo do trato vocal termina com a abertura entre os lábios (ou as narinas no caso de sons nasais). Assim, um modelo razoável está descrito na Figura 2.14a, que mostra o lábio abrindo como um orifício em uma esfera. Neste modelo, para baixas freqüências, a abertura pode ser considerada uma superfície radiante, com as ondas de som radiadas sofrendo difração por uma esfera que representa a cabeça [6].

Os resultados dos efeitos da difração são complicados e difíceis de representar; entretanto, para determinar a condição de fronteira dos lábios, tudo que é necessário é uma estreita relação entre pressão e vazão da superfície de radiação. Nivelar isto é muito complicado para a configuração da Figura 2.14a. Entretanto, se a área de radiação (abertura dos lábios) é pequena comparada ao tamanho da esfera, uma aproximação razoável assume que a área de radiação está contida em um plano de extensão infinita como descrito na Figura 2.14b.

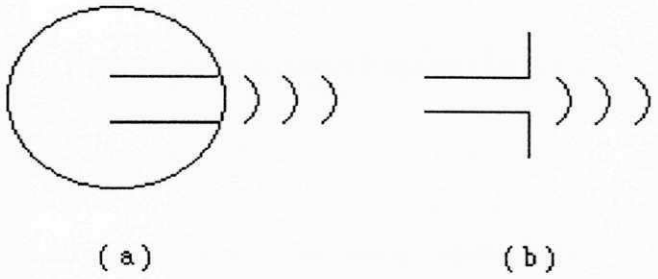


Figura 2.14: a) radiação de uma esfera; b) radiação de um plano infinito.

O mecanismo de radiação é modelado por uma impedância que transforma as ondas de vazão em ondas de pressão. Considerando-se as aberturas dos lábios e narinas desprezíveis em relação à superfície total da face, que é tratada então como um refletor plano de área infinita, obtem-se a "impedância de carga" das aberturas [14], dada por:

$$Z(\Omega) = \frac{j\Omega LR}{R + j\Omega L} \tag{2.7}$$

que é equivalente à ligação de uma resistência R em paralelo com uma impedância L.

Valores de R e L que fornecem uma boa aproximação para o plano de radiação infinito são [6]:

$$R = \frac{128}{9\pi^2} \quad e \quad L = \frac{8a}{3\pi c} \tag{2.8}$$

onde a é o raio da circunferência, cuja área é igual a área de abertura dos lábios ou narinas (varia entre 0,5-1,5 cm) e c é a velocidade do som (aproximadamente 35000 cm/s) [6].

Para freqüências abaixo de 4 KHz, a Equação 2.7 pode ser simplificada como:

$$Z(\Omega) = j\Omega L \tag{2.9}$$

2.3.2 Modelo digital para produção da voz

As discussões anteriores procuraram ressaltar os aspectos físicos envolvidos na produção da voz e desenvolver um modelo que servirá de base para um modelo discreto, apropriado para a análise prática realizada em computadores.

Para produzir sinais "tipo-voz" o modo de excitação e as propriedades de ressonância do sistema linear que representa o trato vocal devem variar com o tempo.

Para alguns sons de voz é razoável assumir que as propriedades gerais da excitação e do trato vocal não se alteram para períodos de 10-20 mseg [6].

Nas próximas discussões, será admitido que o espectro do sinal de voz, $S(\Omega)$, é limitado às freqüências abaixo de $\Omega_{máx} = 2\pi f_{máx}$ [rad/s] e amostrado à freqüência $\geq 2.f_{máx}$ (Hz), de acordo com o Teorema da Amostragem. A freqüência angular do sinal discreto, $W = 2\pi f/fs$ [rad], está normalizada em relação a freqüência de amostragem. Com estas considerações, a versão digital da Figura 2.12 (Modelo analógico para a produção da voz) pode ser dada pela Figura 2.15, onde $z = e^{jw}$.

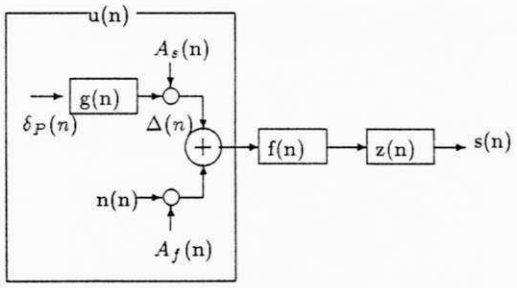


Figura 2.15: Modelo digital para a produção da voz

Na Figura 2.15, $\Delta(n)$ = trem de pulsos glotais, $\delta_P(n)$ é uma seqüência periódica de impulsos com período P, $A_s(n)$ = controle da amplitude dos pulsos glotais, $A_f(n)$ = controle da amplitude do ruído, g(n) é a resposta ao impulso do filtro conformador (formato do pulso glotal), n(n) é uma fonte de ruído branco, $u(n)$ representa a excitação, $f(n)$ = resposta ao impulso do trato vocal, $z(n)$ = resposta ao impulso da radiação e $s(n)$ é o sinal de voz.

2.3.2.1 Excitação

Recordando que a maioria dos sons de voz podem ser classificados como sons surdos ou sonoros, sabe-se que, em termos gerais, o que é requerido é uma fonte que possa produzir tanto formas de ondas com pulsos quase-periódicos como formas de onda de ruído aleatório.

O trem de pulsos glotais, $\Delta(n)$, da Figura 2.15 é modelado pela convolução

$$\Delta(n) = \delta_p(n) * g(n) \tag{2.10}$$

onde

$$\delta_p(n) = \sum_{k=0}^{\epsilon} \delta(n - kP) \tag{2.11}$$

é uma seqüência periódica de impulsos espaçados pelo intervalo de P instantes de amostragem.

Da mesma forma do modelo analógico, g(n) (Figura 2.16) corresponde à resposta ao impulso do pulso glotal.

$$g(n) = \begin{cases} n+1, & 0 \leq n \leq N \\ 2N - n - 1, & N \leq n \leq 2N \end{cases} \tag{2.12}$$

O efeito do pulso glotal no domínio da freqüência corresponde a introdução do efeito da filtragem passa-baixa.

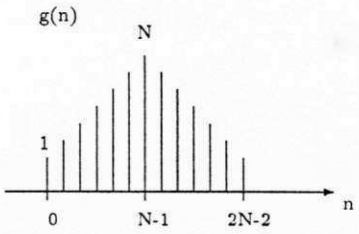


Figura 2.16: Resposta ao impulso do filtro conformador

A excitação turbulenta é simulada por uma seqüência de números aleatórios, n(n), correspondendo a um ruído branco, com densidade de probabilidade uniforme, média nula, variância unitária e descorrelacionado com as outras variáveis do modelo.

2.3.2.2 Trato vocal

As freqüências de ressonância (formantes) do trato vocal correspondem aos pólos da função de transferência $F(z)$, onde $F(z)$ é a representação de $f(n)$ no domínio z. Um modelo só de pólos constitui uma representação bastante satisfatória dos efeitos do trato vocal para a maioria dos sons da voz; entretanto, a teoria acústica nos diz que nasais e fricativos requerem tanto ressonâncias quanto anti-ressonâncias (pólos e zeros). Nesses casos, é necessário incluir zeros na função de transferência ou então seguir a teoria de Atal [16] dizendo que o efeito de um zero na função de transferência pode ser alcançado incluindo mais pólos. Na maioria dos casos esta técnica é utilizada.

Desde que os coeficientes do denominador de $F(z)$ são reais, as raízes do denominador polinomial irão ser também reais ou ocorrerá pares de complexos conjugados. Uma freqüência ressonante complexa do trato vocal é [6]

$$s_k, s_k^* = -\sigma_k + -j2\pi F_k \tag{2.13}$$

A Figura 2.17 descreve freqüências ressonantes complexas tanto no plano-s quanto

no plano-z [6].

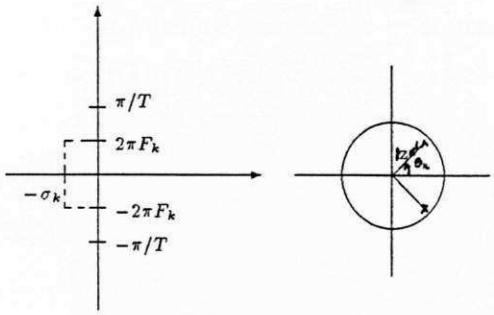


Figura 2.17: Representações da ressonância do trato vocal no (a) plano-s; e (b) plano-z.

Os pólos conjungados complexos correspondentes na representação discreta no tempo irão ser

$$z_k, z_k^* = e^{-\sigma_k T} e^{+-j2\pi F_k T} = e^{-\sigma_k T} cos(2\pi F_k T) + -j e^{-\sigma_k T} sen(2\pi F_k T) \tag{2.14}$$

A largura da banda da ressonância no trato vocal é aproximadamente $2\sigma_k$ e a freqüência central é $2\pi F_k$ [5]. No plano-z, o raio da origem até o pólo determina a largura da banda, isto é,

$$|z_k| = e^{-\sigma_k T} \tag{2.15}$$

e o ângulo no plano-z é

$$\Theta_k = 2\pi F_k T \tag{2.16}$$

Assim se o denominador de $F(z)$ é fatorado, as freqüências formantes análogas correspondentes e as larguras da bandas podem ser encontradas usando as equações 2.15 e 2.16. Como mostrado na Figura 2.17 as freqüências complexas naturais do trato vocal humano estão todas na metade esquerda do plano-s desde que o sistema seja estável. Assim, $\sigma_k > 0$, implica em $|z_k| < 1$; isto é, todos os pólos correspondentes ao modelo discreto no tempo precisam estar inseridos no círculo unitário como requerido para estabilidade [6, 14].

2.3.2.3 Radiação

Até agora considerou-se a função de transferência $Z(z)$, onde $Z(z)$ é a representação de $z(n)$ no domínio z, que relata a vazão da fonte para a vazão dos lábios. Pretendendo-se obter um modelo da pressão dos lábios (como é usualmente o caso), então os efeitos da radiação precisam ser incluídos. Como visto anteriormente no modelo analógico, a vazão e a pressão são relacionados pela Equação 2.7. Deseja-se agora, uma relação similar da transformada-z, da seguinte forma:

$$Z(z) = Z(s)|_{s=\frac{2}{T}\{\frac{1-z^{-1}}{1+z^{-1}}\}} \tag{2.17}$$

Para baixas freqüências pode ser demonstrado que a pressão é aproximadamente a derivada da vazão. Assim, para obter uma representação discreta no tempo desta estreita relação é necessário usar uma técnica de digitalização que evita interpenetração de espectro - "aliasing". Por exemplo, usando o método da transformação bilinear no projeto do filtro digital [6], pode ser mostrado que uma aproximação razoável para os efeitos da radiação é dada por:

$$Z(z) = Z_0(1 - z^{-1}) \tag{2.18}$$

Pode ser visto na Figura 2.18 que a pressão está relacionada com a vazão por uma operação de filtragem passa-alta.

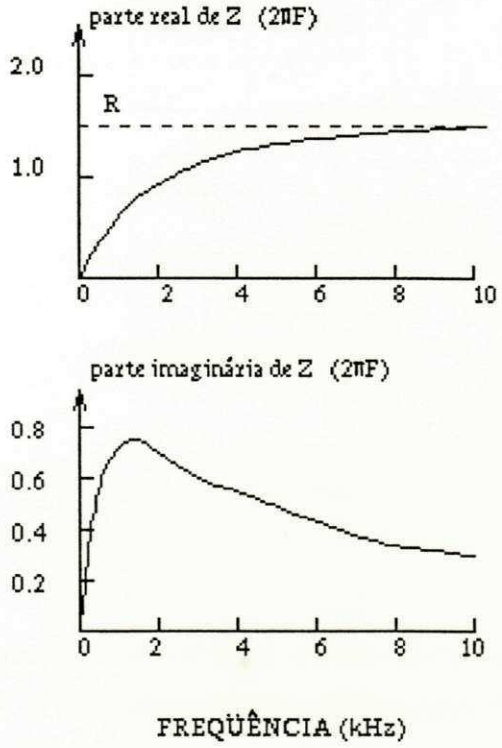


Figura 2.18: Partes real e imaginária da impedância de radiação

$Z(z)$ pode ser implementada de forma conveniente e os parâmetros requeridos irão, é claro, ser apropriados para a configuração escolhida.

2.3.3 O Modelo completo

Colocando todos os componentes necessários obtem-se o modelo da Figura 2.19 [6]. Onde $A_s(n)$ e $A_f(n)$ controlam a intensidade da excitação do sinal de voz e do ruído, respectivamente.

Chaveando entre geradores de excitação sonora e não sonora alterna-se o modo de excitação. O trato vocal pode ser modelado em uma larga variedade de formas.

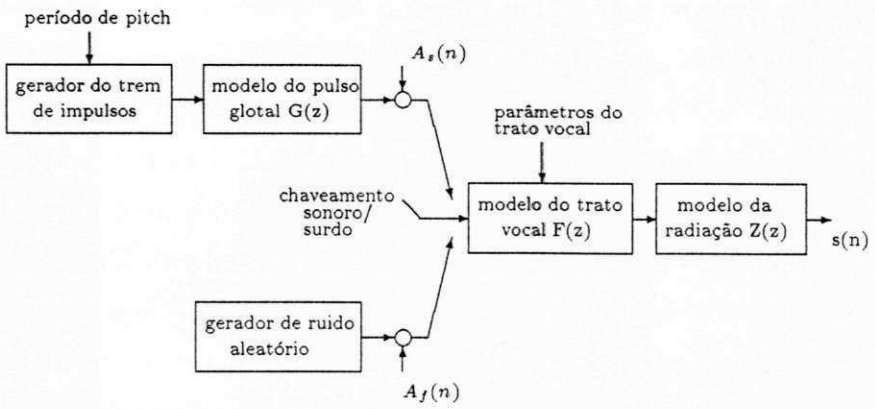


Figura 2.19: Modelo discreto para produção da voz

Em alguns casos é conveniente combinar o pulso glotal e modelos de radiação em um sistema simples. De fato, poder-se-á ver que no caso de análise de predição linear (seção 2.4.2) é conveniente combinar o pulso glotal, radiação e componentes do trato vocal todos juntos e então representá-los com uma simples função de transferência

$$H(z) = G(z)F(z)Z(z) \tag{2.19}$$

só de pólos.

Uma questão natural para este ponto concentra-se nas limitações deste modelo. Certamente o modelo precisa de mais equações parciais. Fortunamente, nenhuma das deficiências deste limita seriamente a sua aplicabilidade. Primeiro, existe a questão da variação dos parâmetros com o tempo. Em sons contínuos como as vogais, os parâmetros variam muito pouco e o modelo trabalha muito bem. Com sons transientes tais como paradas, não apresenta um desempenho muito bom. Poderia ser enfatizado que nosso uso de funções de transferência e funções de resposta em freqüência implicitamente assume que é possível representar o sinal de voz em pequenos intervalos de tempo. Isto é, os parâmetros do modelo são considerados constantes ao longo dos

intervalos de tempo, tipicamente 10-20 mseg. A função de transferência $H(z)$, então, serve para definir a estrutura do modelo cujos parâmetros variam muito pouco com o tempo. Uma segunda limitação é a necessidade de fornecer zeros como requerido teoricamente para nasais e fricativos. Isto é definitivamente uma limitação para sons nasais, mas não tão severa para sons fricativos. Zeros podem ser incluídos no modelo se desejado. Terceiro, uma simples dicotomia da excitação sonoro-não sonoro é inadequada para fricativos sonoros. Adicionar simplesmente as excitações sonoro e não sonoro é inadequada visto que a fricção é correlacionada com os picos do escoamento glotal. Um modelo mais sofisticado para fricativos sonoros tem sido desenvolvido [17] e pode ser aplicado quando necessário. Finalmente, uma consideração relativamente menos importante é que o modelo da Figura 2.19 requer que o pulso glotal seja espaçado por um múltiplo inteiro do período de amostragem, T. Witham e Steiglitz [18] têm considerado formas de eliminação desta limitação em situações requerendo controle preciso de pitch [6].

Nesta última seção foi mostrado que o sinal de voz pode ser razoavelmente modelado pela resposta de um filtro $H(z)$ composto apenas por pólos. A excitação do filtro é uma seqüência de números aleatórios na produção de sons não sonoros, ou um trem de impulsos, na produção de sons sonoros. Este resultado, unido a hipótese de que o sinal de voz é ergódico e portanto estacionário no sentido amplo, possibilitam a utilização de técnicas de Predição Linear no seu estudo. O grande mérito da Predição Linear, quando aplicada à análise do sinal de voz, se encontra na possibilidade de estimar os parâmetros do filtro $H(z)$ de forma simples e precisa, a partir do próprio sinal [19].

2.4 Características que diferenciam os locutores

Quando a tarefa é identificar a pessoa que está falando em vez de reconhecer o que está sendo dito, o sinal de voz deve ser processado visando extrair as características do locutor [7].

É difícil separar no sinal de voz as características que refletem a identidade dos sons produzidos, ou seja, identificar a quem pertence o sinal produzido, pois estas

dependem de aspectos próprios de cada locutor. Em geral, há duas fontes de variação entre locutores [7]:

1. Diferenças na forma das cordas vocais e do trato vocal;
2. Diferenças no estilo de locução;

Não há referências acústicas que tratem especificamente ou exclusivamente da identificação do locutor. A maioria dos parâmetros e características usadas na análise de voz contêm informação útil tanto para identificação do locutor quanto da mensagem falada. Os dois tipos de informação, entretanto, são codificadas muito diferentemente.

Como não há um conjunto de referências acústicas simples que distingam confiavelmente os locutores, reconhecedores de locutor utilizam, tipicamente, médias estatísticas a curtos intervalos de tempo, determinadas ao longo de várias elocuções ou exploram a análise de sons específicos.

O último método é comum em aplicações dependente do texto, onde elocuções do mesmo texto são usadas para treinamento e teste; o método das médias estatísticas é muitas vezes utilizado em casos independente do texto, onde treinamento e teste envolvem elocuções de diferentes textos.

Por simplicidade, a maioria dos sistemas de Reconhecimento Automático de Locutor utilizam parâmetros de padrões de voz tais como: 8-12 coeficientes LPC (Coeficientes de Predição Linear) ou 17-20 bancos de filtros passa-faixas de energia. Entretanto, vendo o reconhecimento de locutor como um problema de separação de densidades de probabilidades no espaço N-dimensional, melhores resultados, com baixa computação, podem ser obtidos pela seleção mais cuidadosa dos parâmetros ou características que contém o espaço. Idealmente, o espaço deve usar características um pouco independentes que apresentem variações intralocutor pouco similares e grandes variações interlocutor.

Uma forma de selecionar as características acústicas para Reconhecimento de Locutor é examinar que características se correlacionam com a percepção humana de similaridade de voz. Quando a análise do escalonamento multidimensional é aplicada para julgamentos semelhantes, as seguintes características são calculadas para a maior

parte das variações entre locutores: F_0 (freqüência fundamental), as três primeiras freqüências formantes F_1, F_2 e F_3, duração da palavra, sexo e idade do locutor. Embora sexo e idade do locutor não sejam características acústicas, entretanto a freqüência F_0 pode contribuir para uma estimação dessas características. Por outro lado, características temporais e espectrais se constituem em fortes candidatos ao reconhecimento de locutor.

As fontes de variação do locutor podem ser classificadas em função das características fisiológicas ou de comportamento, que conduzem a dois tipos de características úteis. As características inerentes ao locutor e as características instruídas. As características inerentes ao locutor são relativamente fixas e dependem sobretudo da anatomia do seu trato vocal. Sabendo que estas podem ser afetadas pelas condições de saúde (p.ex., gripes que congestionam as passagens nasais). Essas características são menos susceptíveis a imitação de impostores que as características instruídas. Estas últimas se referem ao movimento dinâmico do trato vocal, ou seja, a forma como o locutor fala. Sabendo que as características instruídas podem ser usadas para distinguir pessoas com trato vocal semelhante, entretanto são bastante dependentes do estado emocional do indivíduo. Impostores geralmente encontram facilidade para enganar reconhecedores baseados em características instruídas. Características estatísticas baseadas em médias estatísticas a longo intervalo de tempo refletem mais as características inerentes do que as instruídas e são adequadas para reconhecimento de locutor independente do texto [7].

2.4.1 Uso da Freqüência Fundamental

O cálculo da freqüência fundamental (F_0) sobre todos os dados de teste para um locutor, frequentemente funciona como uma simples característica para classificar locutores de forma grosseira dentre grupos gerais (quanto ao sexo e idade: homens, mulheres e crianças).

Em [7] foi utilizada a freqüência fundamental F_0 para reconhecimento de locutor, a qual foi estimada em 40 intervalos iguais em uma elocução totalmente sonora durante

2 segundos. Utilizando-se os 4 primeiros momentos de F_0 produziu-se 78% de reconhecimento. Para colocar a utilidade de F_0 em perspectiva, entretanto, um conjunto de 12 parâmetros cepstrais precisaram somente de 0.5 segundos dos 2 segundos de elocução para alcançar 98% de precisão. Assim, um conjunto de características espectrais é mais poderoso do que o uso de F_0 isoladamente, para reconhecimento de locutor. Entretanto, verifica-se que o uso de F_0 combinado a outros parâmetros pode levar a bons resultados [7, 8].

2.4.2 Uso dos Coeficientes de Predição Linear

Uma das mais importantes técnicas para análise de voz é o método da análise linear preditiva. Este método tem sido a técnica predominante para estimar os parâmetros básicos da voz, ou seja, pitch, formantes, espectro, funções área do trato vocal e para representação da voz em transmissão a baixa taxa de bits ou armazenagem. A importância desse método reside tanto na habilidade de fornecer estimativas extremamente corretas dos parâmetros da voz, quanto na relativa velocidade de computação [5].

A idéia básica da predição linear reside no fato de que a voz amostrada pode ser aproximada como uma combinação linear das amostras de voz passadas.

A filosofia da predição linear está intimamente relacionada com o modelo de voz discutido anteriormente, que mostrou como o sinal de voz pode ser modelado como saída de um sistema linear variante no tempo excitado por pulsos quase periódicos (para os sons sonoros), ou ruído aleatório (para sons não sonoros). Os métodos de predição fornecem um método robusto, realizável e correto para estimação dos parâmetros que caracterizam o sistema linear variante com o tempo.

As técnicas de predição linear poderiam ser aplicadas em um esquema de quantização para reduzir a taxa de bits na representação digital do sinal de voz. Essas técnicas são referidas como Codificação por Predição Linear (LPC) [6].

As técnicas e métodos de predição linear estão disponíveis na literatura de engenharia há um longo tempo e têm sido empregados vastamente, principalmente em sistemas

de controle, automação, telecomunicações e teoria da informação e codificação.

A predição linear pode ser aplicada utilizando os seguintes métodos:

1. o método da covariância [16];
2. o método da autocorrelação [20];
3. a formulação do filtro inverso [6];
4. a formulação da estimação espectral [6];
5. a formulação da máxima verossimilhança [6];
6. a formulação do produto interno [6];

dentre outros.

Um estudo dos vários métodos e a comparação entre eles podem ser encontrados em [6]. Neste trabalho será utilizado o Método da Autocorrelação, discutido em seguida.

A forma particular do modelo digital de produção de voz que é apropriada para a utilização da predição linear está descrita na Figura 2.20.

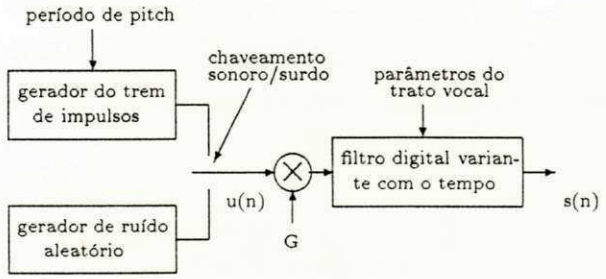


Figura 2.20: Diagrama de blocos para o modelo simplificado de produção de voz.

Neste caso, os efeitos da radiação, trato vocal, e excitação glotal são representados por um filtro digital variante no tempo cuja função de transferência tem a seguinte forma [6]:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} c_k z^{-k}}, \; S(z) = U(z)H(z) \tag{2.20}$$

Onde:

$S(z)$ - Transformada-z da seqüência de voz $s(n)$;

$U(z)$ - Transformada-z do sinal de excitação u(n).

Este sistema é excitado por um trem de impulsos para sons sonoros ou por uma seqüência de ruído aleatório para sons não sonoros. Assim, os parâmetros do modelo são: classificação sonoro/não sonoro, período fundamental, parâmetro de ganho G, e os coeficientes c_k do filtro digital. Esses parâmetros, é claro, variam muito pouco em curtos intervalos de tempo [6].

A maior vantagem do modelo é que o ganho, G, e os coeficientes do filtro c_k podem ser estimados de forma computacionalmente eficiente pelo método de predição linear.

Para o sistema da Figura 2.20, as amostras de voz $s(n)$ são relacionadas com a excitação u(n) pela Equação diferença [6]

$$s(n) = \sum_{k=1}^{p} c_k s(n-k) + Gu(n) \tag{2.21}$$

Uma predição linear com coeficientes de predição, $c(k)$ é definida como um sistema cuja saída é

$$\tilde{s}(n) = \sum_{k=1}^{p} c_k s(n-k) \tag{2.22}$$

O erro de predição, e(n), é definido como

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} c_k s(n-k) \tag{2.23}$$

Para formular o problema, inicialmente é selecionado um segmento do sinal de voz através de uma janela de comprimento finito e igual a N_A (Figura 2.21). A melhor

escolha do valor de N_A permite uma boa aproximação às hipóteses de ergodicidade e estacionariedade no sentido amplo, já citadas anteriormente. Em virtude da inércia dos articuladores, é intuitivo que o sinal de voz possa ser considerado estacionário em intervalos apropriados, de curta duração.

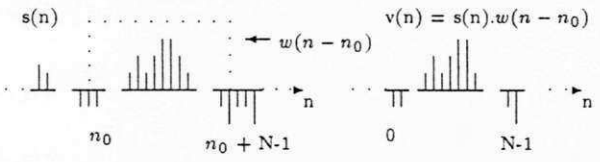


Figura 2.21: Exemplo de um segmento de voz selecionado a partir da seqüência $s(n)$ através de uma janela retangular.

Para simplificar as notações, a origem do eixo n é redefinida a cada segmento selecionado.

Nas próximas equações, $v(n)$ corresponderá ao segmento selecionado e ponderado pela janela sendo, assim, nulo no intervalo $n < 0$ e $n > N_A$. A origem do eixo "n" será estabelecida no início de cada segmento, para simplificar as notações.

A Equação 2.20 será escrita no domínio do tempo como [5]

$$v(n) = Gu(n) + \sum_{i=1}^{p} c_i v(n-i) \tag{2.24}$$

onde $v(n)$, $0 \leq n < N_A$, é o segmento do sinal de voz.

Como dito anteriormente, a idéia principal da Predição Linear consiste em aproximar cada amostra do sinal de voz pela combinação linear de amostras passadas do sinal. Sendo p o número de amostras passadas utilizadas na combinação linear, pode-se formalizar a aproximação da amostra genérica $v(n)$ pela relação [5]:

$$\tilde{v}(n) = \sum_{i=1}^{p} c_i v(n-i) \tag{2.25}$$

Capítulo 3

Técnicas para reconhecimento de locutor

3.1 Introdução

O ouvido humano é um orgão maravilhoso. Além da habilidade para receber e decodificar o que é falado, o ouvido é capaz de desempenhar diversas outras tarefas. Dentre elas, podem-se citar, por exemplo, localização de objetos, músicas, e a identificação de pessoas a partir de suas vozes. Muitos são os esforços para desenvolver máquinas que, tal qual o ser humano, possam conhecer mensagens faladas, bem como identificar quem as está falando.

O reconhecimento de locutor é um exemplo de uma identificação pessoal biométrica. Este termo é usado para diferenciar técnicas que se baseiam na identificação de certas características intrínsicas da pessoa (como a voz, impressão digital, ou estruturas genéticas) daquelas que usam artefatos para identificação (como chaves, emblemas, cartões magnéticos, dentre outros). Esta distinção faz com que as técnicas biométricas sejam, provavelmente, mais confiáveis. Assim, a motivação principal para o estudo do reconhecimento de locutor é tornar a identificação da voz o mais realizável possível. O que é bastante útil para aplicações de segurança, tais como controle de acesso a ambientes restritos (a voz atuando para abrir e fechar uma porta), controle de acesso de

onde $\tilde{v}(n)$ é a aproximação de $v(n)$ e c_i é o i-ésimo coeficiente da combinação linear; $\tilde{v}(n)$ é normalmente denominada a estimativa ou predição de ordem p da amostra $v(n)$.

O erro de predição da cada amostra, $e(n)$, é definido por

$$e(n) = v(n) - \tilde{v}(n) = v(n) - \sum_{i=1}^{p} c_i v(n-i) \tag{2.26}$$

e o erro quadrático, $E(n)$, acumulado em todo o segmento é dado por

$$E(n) = \sum_{n=-\infty}^{\infty} e(n)^2 \tag{2.27}$$

Como o segmento de voz é nulo para $n < 0$ e para $n > N_A$, o erro de predição (eq. 2.27) é, portanto, nulo para $n < 0$ e $n > N_A + p - 1$. A partir desta consideração, e substituindo a Equação 2.26 na Equação 2.27, obtém-se:

$$E(n) = \sum_{n=0}^{N_A+p-1} [v(n) - \sum_{i=1}^{p} c_i v(n-i)]^2 \tag{2.28}$$

O conjunto de coeficientes c_i que minimiza $E(n)$ é obtido a partir de

$$\frac{\partial[E(n)]}{\partial[c_i]} = 0, \ 1 \leq i \leq p \tag{2.29}$$

Com a substituição da Equação 2.28 em 2.29 e a realização das p derivadas parciais, chega-se ao seguinte sistema de equações lineares:

$$\sum_{k=1}^{p} c_k R_r(|i-k|) = R_r(i), \ 1 \leq i \leq p \tag{2.30}$$

onde

$$R_r(k) = \sum_{n=0}^{N_A-k-1} v(n)v(n+k) \tag{2.31}$$

é a função de autocorrelação a curto prazo. As equações 2.30 e 2.31, conhecidas como Equação de Wiener-Hopf, pode ser vista mais facilmente se colocada da seguinte forma (forma matricial) [5]:

$$\begin{vmatrix} R_r(0) & R_r(1) & \ldots & R_r(p-1) \\ R_r(1) & R_r(0) & \ldots & R_r(p-2) \\ R_r(2) & R_r(1) & \ldots & R_r(p-3) \\ \ldots & \ldots & \ldots & \ldots \\ R_r(p-1) & R_r(p-2) & \ldots & R_r(0) \end{vmatrix} \begin{vmatrix} c_1 \\ c_2 \\ c_3 \\ \ldots \\ c_p \end{vmatrix} = \begin{vmatrix} R_{r1} \\ R_{r2} \\ R_{r3} \\ .. \\ R_{rp} \end{vmatrix} \tag{2.32}$$

Os coeficientes c_i do preditor são determinados a partir da solução das eqs. 2.30 e 2.31 (ou 2.32) e são os coeficientes c_i do filtro $H(z)$ da Figura 2.20 (ou a partir do modelo simplificado, Figura 2.22).

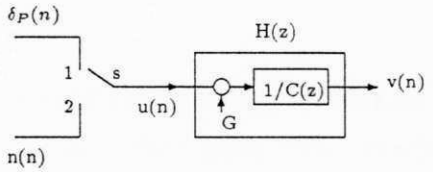


Figura 2.22: Modelo Digital Simplificado para a Produção da Fala

Na Figura 2.22, a chave "s" seleciona ou a excitação periódica (posição 1), ou a excitação turbulenta (posição 2). O filtro $H(z)$ possui apenas pólos. G determina a amplitude do sinal de voz e $C(z)$ é um polinômio de grau p.

Resumindo, os coeficientes de $H(z)$ para um segmento de voz de comprimento N_A são estimados da seguinte forma:

1. Cálculo das autocorrelações a curto prazo através da Equação 2.31;
2. Solução do sistema de equações 2.32.

Utilizando a simetria da matriz de autocorrelação, pode-se elaborar algoritmos recursivos bastante eficientes para solução do sistema, a exemplo do algoritmo de Levinson-Durbin [5, 19] largamente utilizado.

Após a estimação dos coeficientes do polinômio C(z), falta determinar o ganho, G, expresso por [6] como,

$$G = [R_r(0) - \sum_{k=1}^{p} c_k R_r(k)]^{1/2} \tag{2.33}$$

onde $R_r(k)$ é a função de autocorrelação calculada com atraso k. Esta relação é válida tanto para excitação periódica (sons sonoros) quanto para excitação turbulenta (sons surdos) do modelo.

2.5 Discussão

Ondas sonoras são criadas pela vibração e se propagam no ar ou em outro meio pela vibração das partículas do meio. Para gerar o som desejado, o locutor exerce uma série de controles sobre o aparelho fonador, produzindo a configuração articulatória e a excitação apropriadas, gerando os diversos sons da fala (sons sonoros, sons surdos e sons explosivos). A compreenssão dos fenômenos físicos associados à produção da fala é de fundamental importância para a determinação de um modelo apropriado para representação dos sons da voz.

Uma das mais importantes técnicas para análise de voz é o método da análise linear preditiva. Este método tem sido a técnica predominante para estimar os parâmetros básicos da voz e bastante utilizado para a transmissão a baixa taxa de bits, utilizando-se, por exemplo, a quantização vetorial. A Predição Linear é bastante utilizada em reconhecimento de voz e locutor devido a sua capacidade de modelar, de forma bastante satisfatória, o sinal de voz além de apresentar uma relativa velocidade de computação.

dados em computador, ou controle automático de transações telefônicas (p.ex. reservas de vôo ou banco por telefone). Outro benefício relacionado ao sistema biométrico, é que os atributos não podem ser perdidos nem tão pouco precisam ser relembrados.

O Reconhecimento Automático de Locutor (RAL) é um exemplo de uma tarefa de reconhecimento de padrões. Em essência RAL requer um mapeamento entre identificação de voz e de locutor, tal que cada possível forma de onda de entrada é identificada com seu locutor correspondente.

O ponto principal do processo de reconhecimento é uma comparação entre padrões obtidos a partir da representação de parâmetros/características de um sinal de voz desconhecido ou de teste com padrões de referência previamente armazenados, obtidos das características dos possíveis locutores a serem testados. Em identificação automática de locutor, o vetor de padrões de teste é, usualmente, comparado com todos os padrões de referência armazenados em uma memória de dados, podendo a memória, muitas vezes, ser parcionada visando obter-se um procedimento mais eficiente. A comparação envolve uma medida de quão similar o teste e a referência são. O padrão de referência mais estreitamente "casado" com o teste é usualmente escolhido, produzindo uma saída correspondente à aquela referência. Contudo, se o casamento é relativamente pobre ou se outras referências fornecem casamento similar, uma decisão pendente pode ser adiada e ao locutor é solicitado que repita seu padrão [8].

A decisão de aceitar ou rejeitar depende, usualmente, de um limiar: se a distância entre um vetor de padrão de teste e um vetor de padrão de referência excede um limiar, o sistema rejeita o par.

O padrão de representação de uma pequena parcela ou bloco de voz usando K características ou parâmetros, pode ser visto como um vetor K-dimensional. Uma memória de vetor de padrões é estabelecida durante o treinamento quando cada locutor pronuncia um vocabulário, e os segmentos acústicos são convertidos dentro de características identificadas com cada locutor. Para representar elocuções de palavras ou sentenças, o vetor deve incluir as variações das características ao longo do tempo.

A aplicação de métodos de reconhecimento de padrões para RAL envolve vários passos: normalização, parametrização, extração de características, uma comparação

de similaridade e uma decisão, como mostra a Figura 3.1 [8].

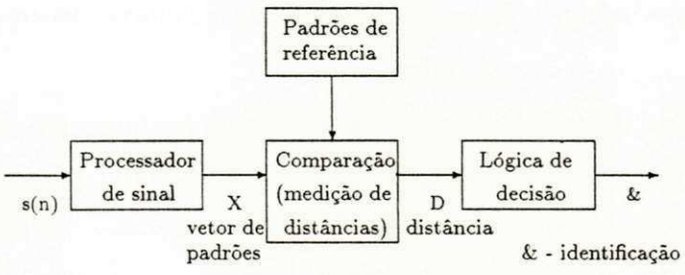


Figura 3.1: Modelo tradicional de reconhecimento de padrões para reconhecimento de locutor.

O passo inicial de normalização tenta eliminar a variação do sinal de voz de entrada devido ao ambiente (i.é., ruído de fundo, nível de gravação, etc.). A forma mais simples de normalização ajusta a amplitude máxima do sinal para um nível pré-estabelecido visando avaliar as variações no nível de gravação, distância para o microfone, intensidade do sinal de voz original e perda na transmissão. Tais variações são assumidas como constantes ou pouco variáveis, o que permite a atualização da amplitude por um fator de escala (pelo qual o sinal recebido é multiplicado) em longos intervalos, correspondendo tipicamente a elocuções limitadas por pausas facilmente identificáveis [8].

A maior redução de dados ocorre na conversão do sinal dentro de certos parâmetros e características. Os parâmetros acústicos derivam-se diretamente dos métodos padrões de análise e codificação de voz (coeficientes de predição linear (LPC), energia do sinal, taxa de cruzamento por zero, etc.). Para parametrizar eficientemente o sinal de voz, um modelo padrão de voz é usado, o qual separa excitação e resposta do trato vocal. A excitação é tipicamente representada em termos de uma decisão vocal, amplitude total, e uma estimativa da freqüência fundamental (durante a identificação do sinal de voz como sonoro). Há uma pequena concordância em relação a quais parâmetros espectrais utilizar; a maioria dos reconhecedores representam a envoltória espectral

com aproximadamente 8-14 coeficientes. Parâmetros comuns são: coeficientes LPC, coeficientes "cepstrais" [9] e energia do sinal, dentre outros. Todos eles tentam capturar dentro de alguns poucos parâmetros, informação espectral suficiente para identificar os locutores [8].

Todas as tarefas de reconhecimento de padrões, incluindo RAL, utilizam duas fases: TREINAMENTO e RECONHECIMENTO. Realizado "off-line" e muitas vezes combinando métodos manuais e automáticos, a fase de treinamento estabelece uma memória de referência ou dicionário de (voz) padrões de referência, aos quais são atribuídos rótulos. Na fase do reconhecimento automático (usualmente em tempo real) são obtidos padrões de teste que são comparados com os padrões de referência e então, utilizando-se uma regra de decisão, é identificado aquele mais semelhante ao padrão de entrada desconhecido. De uma forma geral os métodos conhecidos para reconhecimento de locutor diferenciam-se na forma como os parâmetros extraídos são utilizados na construção dos padrões. Dessa forma, podem ser divididos em dois grupos: MÉTODOS PARAMÉTRICOS e MÉTODOS ESTATÍSTICOS [21].

Nos métodos paramétricos, após a detecção de fim de palavra é levado a efeito uma redução de dados explícita, após a qual é obtido um padrão de referência que continua ainda na forma paramétrica. A regra de decisão no processo de comparação de padrões baseia-se em medidas de distância.

Nos métodos estatísticos a construção dos padrões é obtida através de modelos estatísticos, tais como Modelos de Markov Escondidos (HMMs) [11, 13]. Os parâmetros extraídos são portanto, com o auxílio da teoria das probabilidades, representados por modelos estocásticos nos quais está presente uma redução implícita de dados. Nesses métodos não é feita uma comparação direta de padrões e a decisão é feita através do cálculo de probabilidades associadas aos modelos.

Os métodos paramétricos têm sido bastante estudados, a exemplo daqueles que utilizam programação dinâmica como método para comparação de padrões [7]. Este método tem possibilitado bons resultados. Apesar do sucesso, métodos alternativos de reconhecimento têm sido estudados devido principalmente aos seguintes fatores [22]:

1. O alto custo computacional do método usando programação dinâmica;

2. As dificuldades de estender o método para problemas mais difíceis, como por exemplo, o reconhecimento de locutor para sistemas independentes do texto;

Devido a uma ou mais das razões acima, vários métodos paramétricos têm sido propostos, tais como o uso da quantização vetorial no cálculo da programação dinâmica [22] ou o uso da quantização vetorial para eliminar o processamento da própria programação dinâmica [22]. Embora os reconhecedores baseados em quantização vetorial tenham obtido um desempenho muito bom no reconhecimento de locutor, e tenham contribuído para a redução dos custos computacionais, esses têm feito muito pouco para reduzir as dificuldades computacionais encontradas nos métodos paramétricos. Dessa forma, o reconhecedor HMM tem sido de grande interesse devido ao seu baixo custo computacional durante a fase de reconhecimento, e por basear-se em modelos estocásticos do sinal de voz sendo capaz de modelar vários eventos, tais como fonemas, sílabas, etc. [22], o que o torna bastante flexível.

Em um sistema de reconhecimento de padrões convencional, a palavra de teste desconhecida é alinhada no tempo para cada um dos padrões de referência através de alguma forma de distorção na escala do tempo, geralmente, uma distorção dinâmica na escala do tempo ("Dynamic Time Warping" - DTW). Ao contrário, nenhum alinhamento direto é realizado no sistema HMM, apenas um alinhamento indireto na escala do tempo é resultante da medição do valor de probabilidade. Dessa forma, torna-se interessante apresentar algumas características do Alinhamento Dinâmico no Tempo (DTW), Quantização Vetorial (QV) e Modelos de Markov Escondidos (HMMs) [13].

3.2 Reconhecimento de locutor utilizando Alinhamento Dinâmico no Tempo

3.2.1 Introdução

A maior parte dos reconhecedores de locutor (e voz) de alto desempenho recorre ao uso do alinhamento não linear para solucionar os problemas de alinhamento do

vetor de padrões, na tentativa de alinhar segmentos acústicos similares dos vetores de padrões de referência e teste. O método chamado Alinhamento Dinâmico no Tempo (Dynamic Time Warping - DTW), combina alinhamento e computação da distância através de um método de programação dinâmica [8]. Desvios de uma comparação linear bloco por bloco são permitidos se a distância para o bloco em processamento é pequena comparada com outras comparações locais. DTW alinha vetores de padrões encontrando um alinhamento no tempo que minimiza a medida de distância total, que soma as distâncias dos blocos na comparação de vetores de padrões.

Considerando dois padrões R e T de R e T blocos cada, correspondendo ao vetor de padrões de referência e teste, respectivamente. DTW encontra uma função de alinhamento $m = w(n)$, que mapeia o eixo do tempo n dos vetores padrões de teste dentro do eixo de tempo m do vetor de padrões de referência (Figura 3.2) [8].

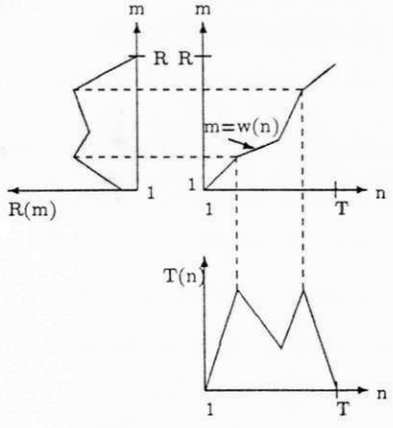


Figura 3.2: Exemplo de um alinhamento não linear no tempo de um padrão de teste T(n) e um padrão de referência R(m).

A maior parte dos sistemas de Reconhecimento Automático de Locutor utiliza vocabulários que envolvem seqüências de diferentes eventos acústicos. As elocuções de teste e referência são subdivididas no tempo, produzindo seqüências de vetores de

parâmetros. Mais comumente, cada sinal de voz é dividido em blocos de igual duração (parcialmente sobrepostos) em torno de 10-30 ms, cada um, produzindo um vetor.

Visto que a segmentação automática de elocuções dentro de unidades linguísticas expressivas (p.ex. fones, sílabas, etc.) é difícil, vetores de padrões são usualmente comparados bloco por bloco, o que leva a problemas de alinhamento. Elocuções são geralmente faladas a diferentes taxas, até mesmo para um simples locutor repetindo a mesma palavra. Logo, elocuções de teste e referência normalmente têm diferentes durações. Uma forma de permitir a comparação linear bloco por bloco é normalizar o intervalo entre blocos de forma que um número comum de blocos é usado para todos os vetores de padrões. Por exemplo, se a duração típica de uma palavra é 400ms, e uma resolução no tempo de 20 blocos/palavra é desejada, o intervalo do bloco iria exceder 20ms para palavras mais longas do que 400ms e é proporcionalmente menor para palavras mais curtas. Tal normalização linear no tempo, ou alinhamento, pode também ser realizado através do ajuste de intervalo do bloco antes da parametrização ou através da decimação/interpolação da seqüência de características [8].

Alinhamento preciso no tempo é crucial para uma bom desempenho do RAL. Casamento de vetores de padrões correspondendo ao mesmo locutor resulta em uma pequena distância quando segmentos acústicos paralelos de dois vetores de padrões são comparados. Alinhamento linear é em geral insuficiente para alinhar eventos de voz, porque os efeitos da variação da taxa de locução são não lineares: vogais e sílabas tônicas tendem a expandir/contrair mais do que as consoantes e sílabas não tônicas. Assim, alinhamento dinâmico linear de duas elocuções da mesma palavra freqüentemente alinha segmentos acústicos de diferentes fones. Se blocos são suficientemente desalinhados, a distância total para a palavra pode ser grande o suficiente para rejeitar uma decisão positiva de casamento, mesmo se eles representam a mesma palavra falada por uma pessoa.

3.2.2 Reconhecimento de locutor DTW convencional baseado na análise LPC

O reconhecedor LPC/DTW tem como base o alinhamento de sinais distorcidos por pertubações lineares ou não lineares dos 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 de voz [13].

A Figura 3.3, dada abaixo, mostra um diagrama de blocos do reconhecedor de palavras isoladas LPC/DTW convencional [13].

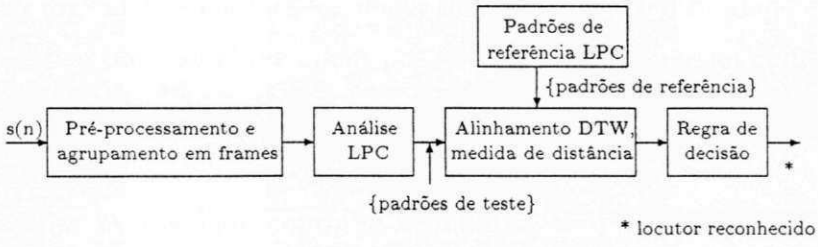


Figura 3.3: Diagrama de blocos do reconhecedor LPC/DTW

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 levadas a uma regra de decisão, que fornece uma classificação da palavra falada, e possivelmente um conjunto ordenado (pela distância) das n melhores candidatas.

Os padrões de referência da palavra, para o reconhecedor da Figura 3.3 são gerados por um algoritmo de treinamento. Tipicamente, em torno de 12 padrões de referência por palavras são suficientes para o reconhecimento do locutor [12, 13]. No reconhecedor

LPC/DTW o procedimento de treinamento trata-se de um processo de armazenagem e coleção de dados computacionalmente simples.

3.3 Reconhecimento de locutor utilizando Quantização Vetorial (QV)

3.3.1 Introdução

As projeções atuais para as comunicações no mundo inteiro apontam para a transmissão digital como um meio dominante da comunicação para voz e dados. Espera-se da transmissão digital que ela forneça maior flexibilidade, credibilidade e custos mais baixos. Além disso, pode-se obter maior privacidade e segurança na comunicação.

Os custos do meio de transmissão como também de armazenamento digital são proporcionais à quantidade de dados digitais a serem transmitidos ou armazenados. Portanto, há uma necessidade contínua de minimizar o número de bits necessário para transmitir sinais, de forma a manter a inteligibilidade e a qualidade em valores aceitáveis. Na engenharia elétrica, o campo que trata deste problema é chamado compressão ou codificação de dados, que aplicado a voz, é conhecido por codificação ou compressão de voz.

A conversão de uma fonte analógica numa fonte digital, consiste de duas etapas: amostragem e quantização. Na amostragem, o sinal analógico é convertido num sinal discreto no tempo pela medição de valores do sinal em intervalos regulares de tempo. Na quantização, o sinal contínuo em amplitude é convertido num sinal de amplitudes discretas, que é diferente do sinal contínuo em amplitude pelo erro ou ruído de quantização [23].

A quantização de cada amostra ou parâmetro do sinal separadamente é chamada quantização escalar. A quantização conjunta de um bloco de amostras ou de parâmetros do sinal é chamada quantização de bloco ou quantização vetorial.

O propósito da Quantização Vetorial (QV) em codificação de voz é bastante amplo.

Várias são as aplicações. Por exemplo, reduzir a taxa de transmissão dos "vocoders" (voice coders - codificadores de voz) de 2400 bits/s para operar em taxa muito mais baixas, mantendo ainda qualidade e inteligibilidade aceitáveis e mais recentemente, codificação dos parâmetros do filtro em codificadores CELP. A codificação de voz em taxas na faixa de 200-800 bits/s tem atraído substancial interesse [24, 25] para uso tanto em operações comerciais quanto governamentais. A QV tem sido usada efetiva e regularmente em sistemas de reconhecimento de voz e de locutor. Além disso, o problema da QV é parte do problema geral de reconhecimento de padrões dentro de um número discreto de categorias que otimizam algum critério de fidelidade. A teoria básica da QV vem da teoria da informação e tem largas aplicações para a transmissão da informação.

O algoritmo de Linde, Buzo e Gray (LBG) é um algoritmo eficiente e intuitivo para o projeto de bons quantizadores vetoriais com medidas de distorção muito gerais, desenvolvido para usar ou em descrições de fonte probabilísticas conhecidas ou numa longa seqüência de dados de treinamento [23]. O algoritmo LBG é baseado no método de Lloyd [19, 24, 25], não é uma técnica variacional, e não envolve diferenciação. Portanto, ele pode trabalhar bem mesmo quando a distribuição tem componentes discretas, como no caso em que se tem uma distribuição das amostras da seqüência de treinamento.

3.3.2 Quantização Vetorial

Um quantizador vetorial K-dimensional de M-níveis, é um mapeamento, q, que assume para cada vetor de entrada, $x = \{x_0, ..., x_{k-1}\}$, um vetor de reprodução, $\hat{x} = q(x)$, extraído de um alfabeto de reprodução finito $Y = \{y_i; i = 1, ..., M\}$. O quantizador q é completamente descrito pelo alfabeto de reprodução (ou dicionário) Y junto com a partição, $S = \{S_i; i = 1, ..., M\}$, do espaço vetorial de entrada nos conjuntos $S_i = \{x : q(x) = y_i\}$ do mapeamento dos vetores de entrada no i-ésimo vetor de reprodução. y_i é um vetor de dimensão K denominado vetor de reprodução ou vetor de saída.

O conjunto Y é referido como dicionário de reconstrução, M é o tamanho do dicionário, e y_i são os vetores códigos de dimensão K. O tamanho M do dicionário é

também conhecido como o número de níveis, um termo emprestado da terminologia da quantização escalar. Assim, diz-se um quantizador de M níveis ou um dicionário de M níveis [25].

A seqüência de vetores de reprodução y_i é mapeada em uma seqüência digital adequada para transmissão ou armazenamento com dimensão $log_2 M$. A taxa de bits/amostra é dada portanto por:

$$\frac{log_2 M}{K} \tag{3.1}$$

A Quantização Vetorial é uma técnica de codificação usada tipicamente para transmissão a baixa taxa de bits. Aplicada a Reconhecimento de Locutor, a QV fornece uma alternativa para DTW ("Dynamic Time Warping"). A eficiente taxa de redução de dados da QV dentro da parametrização de voz é útil em Reconhecimento de Locutor para minimizar a memória utilizada. A principal vantagem da QV em reconhecimento de locutor está no método de reprodução do dicionário para determinação da similaridade entre elocuções.

Em codificação de voz via QV, blocos de voz são tipicamente representados por K parâmetros, os quais são codificados juntos como um bloco ou vetor. Se os elementos do vetor são correlacionados de alguma forma, tal codificação será mais eficiente do que tratando os K parâmetros individualmente. Um conjunto apropriadamente escolhido de 1024 (210 para um QV de 10 bits) deveria ser capaz de representar adequadamente a envoltória espectral visando cobrir todos os possíveis sons da fala [8]. A desvantagem da QV está no aumento da complexidade de análise do codificador. Depois que a análise normal é completada (produzindo K parâmetros escalares para um dado bloco da análise), o codificador deve então determinar qual o vetor de dimensão K, dentre um conjunto de M possibilidades armazenados em um dicionário, corresponde mais estreitamente ao conjunto de parâmetros escalares. Uma medida de distância (p.ex. Medida de Distorção do Erro Quadrático) é usada como um critério de decisão para o projeto e operação do dicionário.

O ponto em questão na implementação do QV consiste no projeto e busca do dicionário. A criação do dicionário necessita da análise de uma longa seqüência de

treinamento de voz, tipicamente uns poucos minutos são suficientes para que possa conter exemplos de fonemas em diferentes contextos. Um procedimento de projeto iterativo é usado afim de convergir sobre um dicionário local ótimo (ótimo no sentido de que a medida de distorção média é minimizada através do conjunto de treinamento).

Comparada com a codificação escalar, a maior complexidade da QV está no tempo necessário para busca do dicionário para que a palavra código apropriada melhor represente um dado vetor de voz. Para busca completa do dicionário, o vetor de todo bloco é comparado com cada uma das M palavras código requerendo cálculos de M distâncias (cada uma contendo K operações quadradas e $2k-1$ adições, no caso de uma Distância Euclidiana Simples). O projeto do dicionário é um problema de tempo, para o qual uma computação elevada é necessária, se o dicionário é usado durante um longo período de tempo. Entretanto, para aplicações de codificação em tempo real, o custo de uma busca completa do dicionário deve ser balanceado com um sistema de melhor desempenho com um K maior.

Um dicionário é usualmente projetado para cada combinação de locutor e palavra do vocabulário, baseado em uma ou mais elocuções da palavra. Cada padrão de teste é avaliado por todos os dicionários, e o locutor correspondente ao dicionário que apresenta a menor medida de distância é selecionado como a saída do Sistema de Identicação de Locutor (para Verificação de Locutor, a distorção do dicionário é comparada com um limiar).

Na sua forma simples, um dicionário não tem informação explícita do tempo, ou em termos de ordem temporal ou durações relativas, visto que as entradas dos dicionários não estão ordenadas e podem derivar de qualquer parte das palavras de treinamento. Entretanto, referências da duração são parcialmente preservadas porque as entradas são escolhidas de forma a minimizar a distância média através de todos os blocos do treinamento, e os blocos correspondentes a segmentos acústicos mais longos (p.ex. vogais) são mais freqüentes nos dados de treinamento. Tais segmentos são mais prováveis para especificar as posições da palavra código do que os blocos de consoantes menos freqüentes, especialmente em pequenos dicionários.

Aumentando o tamanho do dicionário cresce o tempo de computação mas decresce

a probabilidade de erro pela redução dos desvios padrões das distorções. O aumento do desempenho com a duração depende do grau de correlação entre palavras da elocução de teste.

Em resumo, o Reconhecimento de Locutor via QV pode produzir alta precisão tanto nos casos dependente quanto independente do texto, com elocuções de teste relativamente curtas. Em reconhecimento de voz a QV tem muitas vezes a vantagem de requerer menor memória de referência em comparação com a de vetores de padrões de palavras no método DTW [8].

3.3.3 Projeto do dicionário

Para o projeto do dicionário, o espaço K-dimensional do vetor aleatório x é particionado em M regiões ou células $\{C_i,\ 1 \leq i \leq M\}$ e associa a cada célula C_i um vetor y_i. O quantizador então assume o vetor código Y_i se x está em C_i.

$$q(x) = y_i, \quad se \quad x \in C_i \tag{3.2}$$

A Figura 3.4 mostra um exemplo de um particionamento do espaço bi-dimensional ($K = 2$) para o propósito da quantização vetorial. A região limitada pelas linhas mais fortes é a célula C_i. As posições dos vetores códigos correspondentes às outras células são mostradas pelos pontos. O número de vetores códigos, para este caso, é $M = 8$ [24, 25].

Para $K = 1$ (uma dimensão), a quantização vetorial se reduz a quantização escalar. A Figura 3.5 mostra um exemplo de um particionamento da linha real para quantização escalar. Os valores códigos (saída ou níveis de reconstrução) são mostrados por pontos negros dentro dos intervalos. O número de níveis na Figura 3.5 é $M = 10$.

Na quantização escalar, as células podem ter tamanhos diferentes, mas têm a mesma forma. Por outro lado, na quantização vetorial, as células têm formas diferentes. Esta liberdade de ter vários formatos de células no espaço multidimensional dá à quantização vetorial uma vantagem sobre a quantização escalar [25].

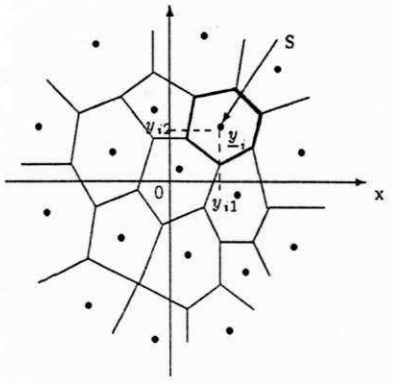


Figura 3.4: Partição do espaço bi-dimensional ($K = 2$) em $M = 8$ células.

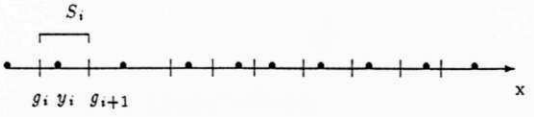


Figura 3.5: Particionamento da linha real em 10 células ou intervalos para quantização escalar ($K = 1$).

Quando x é quantizado como y (ou $\hat{x}$), resulta um erro de quantização e uma medida de distorção pode ser definida entre x e $\hat{x}$.

Nos projetos dos sistemas de compressão de dados (ou voz), tenta-se projetar o quantizador de forma que a distorção na saída seja minimizada para uma determinada taxa de transmissão. Assim, uma das decisões mais importantes no projeto de um quantizador é qual a medida de distorção a ser utilizada.

3.3.4 Medidas de Distorção

Muitos pesquisadores têm descoberto que uns poucos decibéis de diminuição na distorção é muito perceptível pelo ouvido humano em uma situação, mas não em outra. Idealmente, uma medida de distorção deve ser tratável para permitir análise, de forma a ser analisada em tempo real e usada em sistemas de distorção mínima, e subjetivamente relevante tal que medidas de distorção quantitativamente grandes ou pequenas se correlacionem com qualidade subjetiva ruim e boa.

Assume-se que a distorção causada pela reprodução de um vetor de entrada x por um vetor de reprodução $\hat{x}$ é dada por uma medida de distorção não-negativa $d(x, \hat{x})$. Muitas medidas são propostas pela literatura tais como: Medida de Distorção do Erro Quadrático [23], o Erro Médio Quadrático Ponderado, Medida de Distorção de Itakura-Saito [24], dentre outras. Neste trabalho foi utilizada a Medida de Distorção do Erro Quadrático descrita a seguir.

Medida de Distorção do Erro Quadrático

É a medida mais simples e mais comum, por sua simplicidade e tratamento matemático. Os espaços de reprodução e de entrada são espaços Euclidianos k-dimensionais onde [23]

$$d(x, \hat{x}) = \sum_{i=0}^{k-1} |x_i - \hat{x}|^2 \tag{3.3}$$

Tem-se, então, o quadrado do espaço Euclidiano entre os vetores para a distorção do erro quadrático.

3.3.5 Escolha do alfabeto de reprodução inicial

Existem várias formas de se escolher o alfabeto de reprodução inicial $\hat{A}_0$ exigido pelo algoritmo do quantizador vetorial (ver apêndice A). Um dos métodos, o escolhido neste trabalho, para usar nas distribuições amostrais é o método "k-means", pela escolha dos primeiros M vetores da seqüência de treinamento [23, 24].

3.4 Reconhecimento de locutor utilizando Modelos de Markov Escondidos

3.4.1 Introdução

Embora inicialmente estudado entre os anos 60 e 70, os métodos estatísticos de Markov ou Modelos de Markov Escondidos (Hidden Markov Models - HMMs) têm se tornado cada vez mais populares nos últimos anos. Há duas fortes razões para que isto tenha ocorrido. Primeiro, os modelos são muito ricos em estrutura matemática e consequentemente podem formar uma base teórica para uso em um largo grupo de aplicações; segundo, os modelos quando aplicados apropriadamente, trabalham muito bem para várias aplicações práticas.

O uso de HMMs foi proposto por Baker e, independentemente, por um grupo da IBM e mais recentemente pela Phillips [11].

Uma função probabilística de um canal de Markov (escondido) é um processo estocástico gerado por dois mecanismos interrelacionados. Um canal de Markov básico tem um número finito de estados, e um conjunto de funções aleatórias, onde cada uma está associada a cada um dos estados. Para instantes de tempo discretos, assume-se que o processo está em algum estado e uma seqüência de observação é gerada por uma função aleatória correspondendo ao estado corrente. O canal de Markov básico escolhe o estado de acordo com uma matriz de probabilidade de transição associada. O observador vê somente a saída da função aleatória associada com cada estado e não pode observar diretamente os estados do canal de Markov básico; daí o termo Modelo de Markov Escondido [12].

Em princípio, o canal de Markov básico pode ser de uma ordem e as saídas dos estados podem ser processos aleatórios de multivariáveis possuindo algumas funções densidade de probabilidade associadas. Neste trabalho, restringiu-se as considerações para canais de Markov de ordem 1, i.é., aqueles nos quais a probabilidade de transição para algum estado depende somente deste estado e do estado predecessor [11, 12, 13].

No domínio de voz, os Modelos de Markov Escondidos (HMMs) têm sido de grande

interesse devido ao seu baixo custo computacional durante a fase de reconhecimento (para tanto é necessário apenas o cálculo de uma medida de probabilidade, diferentemente dos métodos paramétricos que envolvem, durante a fase de reconhecimento, o cálculo de medidas de distância, acarretando num maior tempo de cálculo) e por basear-se em modelos estocásticos do sinal de voz capazes de modelar vários eventos, tais como fonemas, sílabas, etc. [22], o que os tornam bastante flexíveis. Algumas das vantagens do uso de HMM são [26]:

1. A habilidade para treinar vários exemplos. Os parâmetros do modelo são automaticamente agrupados para representar as entradas.

2. As características temporais do sinal de entrada (modelo "esquerda-direita") são modeladas inerentemente.

3. Considera as variações estatísticas do sinal de entrada por estarem implícitas na própria formulação probabilística.

4. Não é necessário uma distribuição estatística a priori de entradas para estimação dos parâmetros, que não é o caso, usualmente, em outras técnicas estatísticas.

É completamente natural pensar no sinal de voz como sendo gerado por tal processo. Pode-se imaginar o trato vocal como sendo constituído de um número finito de configurações articulatórias ou estados. A cada estado é associado um sinal com características espectrais que caracterizam o estado. Assim, a potência espectral de curtos intervalos do sinal de voz é determinada somente pelo estado corrente do modelo, enquanto a variação da composição espectral do sinal com o tempo é governada predominantemente pela lei probabilística de transição de estados do canal de Markov básico [12].

Para sinais de voz derivados de um pequeno vocabulário de palavras isoladas, o modelo é bastante fiel. O precedente é, claramente, uma simplificação pretendida somente para o propósito de motivação da discussão teórica seguinte.

3.4.2 Modelos de Markov Escondidos (HMMs)

Existem dois casos gerais do modelo que são de interesse, ou seja, o caso "ergódico" no qual a cadeia de Markov é ergódica (i.é., todos os estados são aperiódicos e recorrentes não nulos, Figura 3.6) e o caso "esquerda-direita" (Figura 3.7) no qual uma transição do estado q_i para o estado q_j é possível se $j \geq i$ (i.é., existe uma progressão seqüencial através dos estados do modelo). Ambos os casos são de interesse para aplicações reais [27].

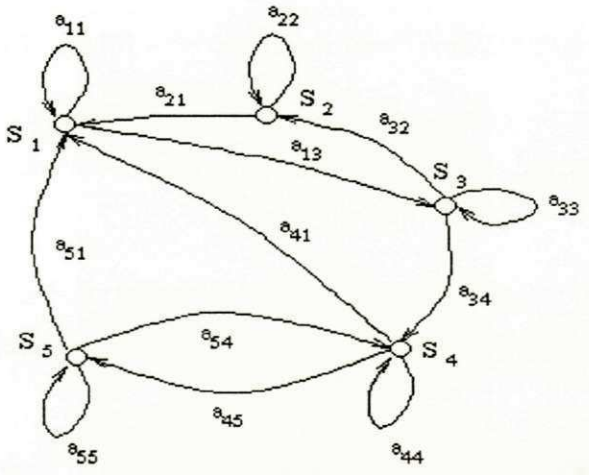


Figura 3.6: HMM - "ergódico" com 5 estados

O tipo de HMM considerado neste trabalho é o tipo "esquerda-direita" mostrado na Figura 3.7. O sinal é assumido como sendo uma função estocástica da seqüência de estados da cadeia de Markov. O objetivo é escolher-se os parâmetros do HMM que correspondam de maneira ótima às características observadas de um dado sinal [22].

Não existem muitos trabalhos que utilizam HMMs para verificação de locutor, é mais comum o uso para reconhecimento de fala. Porém, dentre as poucas referências [28, 29] no assunto, consideram-se modelos do tipo "esquerda-direita". Estes modelos têm as seguintes propriedades [12]:

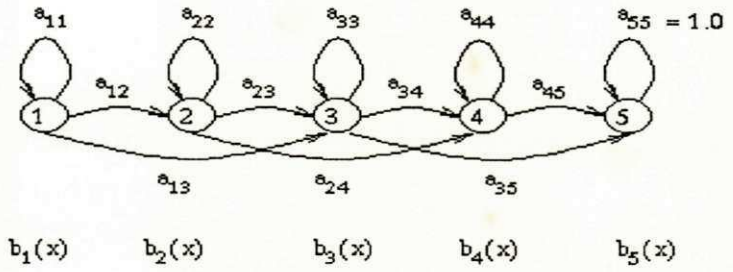


Figura 3.7: HMM - "esquerda-direita" com 5 estados

1. A primeira observação é produzida quando a cadeia de Markov encontra-se em um estado determinado, chamado estado inicial, designado por q_1.

2. A última observação é gerada enquanto a cadeia de Markov está em um outro estado determinado, chamado estado final ou estado de absorção, designado por q_N.

3. Uma vez que a cadeia de Markov deixa um estado, aquele estado não pode ser mais visitado num tempo posterior.

Os parâmetros que caracterizam o HMM, $\lambda = (\mathcal{A}, \mathcal{B}, \pi)$, da Figura 3.7 são:

1. N, número de estados do modelo. Estados individuais são denotados como $(q_1, q_2, ..., q_N)$. Embora os estados sejam escondidos, para algumas aplicações práticas há algumas significações físicas relacionadas aos estados ou aos conjuntos de estados do modelo. Por exemplo, em lançamento de moedas, cada estado corresponde a uma moeda distinta. Geralmente os estados são interconectados de tal forma que um estado pode ser relacionado de alguma forma com outro estado (p.ex. um modelo ergódico).

2. $\mathcal{A} = [a_{ij}]$, $1 \leq i, j \leq N$, a matriz transição de estados, onde a_{ij} é a probabilidade de ocorrer uma transição do estado q_i para o estado q_j.

a_{ij} = prob(q_j em t+1/q_i em t). Em modelos "esquerda-direita" usou-se a restrição $a_{ij} = 0, j < i, j > i + 2$.

3. $\mathcal{B}$= $[b_j(k)]$, $1 \leq j \leq N$ e $1 \leq k \leq M$, é uma matriz de função de probabilidade das observações. Indica a probabilidade de observar, em uma dado estado q_j, a saída do modelo através de um vetor aleatório com uma função densidade de probabilidade (f.d.p) b_j [27].

4. $\pi = \pi_i = P\{q_i/t = 1\}$, $1 \leq i \leq N$, vetor de probabilidade do estado inicial, indica a probabilidade de iniciar o processo no estado q_i para t=1.

Assume-se que o sinal a ser representado pelo HMM consiste de uma seqüência de vetores de observação $\mathbf{O} = \{O_1, O_2, ...O_T\}$, onde cada vetor O_t é formado pelos coeficientes LPC obtidos para cada bloco de amostras do sinal de voz analisado, que caracteriza o sinal no t-ésimo intervalo de tempo. Assim, cada bloco de amostras do sinal de voz, corresponderá a um determinado intervalo de tempo. Desta forma, pode-se considerar dois tipos de funções de probabilidades das observações, ou seja, contínua e discreta.

Em alguns estudos, assume-se que todos os parâmetros de interesse possuem distribuições Gaussianas, tem-se então, o HMM de densidades contínuas [22]. Uma forma alternativa para o uso de HMMs é a combinação com a quantização vetorial [13], onde os parâmetros de interesse (vetores LPC - Coeficientes de Predição Linear) são transformados em um conjunto de observações discretas. Tem-se então os chamados HMMs de densidades discretas [11].

Neste trabalho foi utilizado HMM com função de probabilidade das observações discreta, denominado HMM de densidades discretas.

Como na maioria dos sistemas de verificação, assume-se um conjunto de dados de treinamento, a partir dos quais é construída uma série de Modelos de Markov, um para cada locutor. Então, quando deseja-se verificar um locutor, calcula-se a medida de probabilidade associada ao HMM de referência já armazenado, correspondente a este locutor. O locutor é aceito se seu valor de probabilidade é maior ou igual ao limiar estabelecido, caso contrário o locutor é rejeitado.

3.5 Discussão

De uma forma geral, os métodos conhecidos para reconhecimento de locutor distinguem-se pela forma como os parâmetros extraídos são utilizados na construção dos padrões. Portanto, podem ser divididos em métodos paramétricos e métodos estatísticos. Os métodos paramétricos mais usuais são: Análise por Predição Linear, Quantização Vetorial, Alinhamento Dinâmico no Tempo, dentre outros. O método estatístico mais usual é a representação dos sinais de voz por Modelos de Marvov Escondidos.

Apesar do sucesso obtido com a maioria dos métodos paramétricos, o uso de HMM vêm se tornando cada vez mais popular devido ao seu baixo custo computacional durante a fase de reconhecimento, e por basear-se em modelos estocásticos do sinal de voz, sendo capaz de modelar vários eventos, tais como fonemas, sílabas, etc., o que o torna bastante flexível.

O próximo capítulo irá tratar das várias características e parâmetros necessários para representar um dado locutor utilizando os Modelos de Markov Escondidos.

Capítulo 4

Verificação de Locutor utilizando HMMs de densidades discretas

4.1 Processos Discretos de Markov

Os modelos de Markov se constituem em modelos estatísticos, podendo ser utilizados para construir uma seqüência de padrões, que podem representar um dado locutor. Como dito no capítulo anterior, os HMMs podem ser do tipo discreto ou contínuo de acordo com as funções de probabilidades das observações, ou seja, contínuas ou discretas.

Para o HMM do tipo discreto, representa-se cada vetor O_t de uma seqüência de vetores de observação do l-ésimo locutor, $\mathbf{O}^l = \{O_1, ..., O_T\}$, $1 \leq l \leq L$ e $1 \leq t \leq T$, por um dos M possíveis símbolos $v_k \in V$, $1 \leq k \leq M$, onde V representa um alfabeto discreto obtido através da quantização vetorial dos vetores de observação. Neste caso, a matriz $\mathcal{B} = [b_j(k)]$ indica a probabilidade de observar-se um símbolo v_k dado o estado corrente q_j, $1 \leq j \leq N$. Assim, $b_j(k)$ é a probabilidade de que se obtenha o resultado v_k, no instante de tempo t (t-ésimo vetor, tempo discreto), no estado q_j.

4.1.1 Parâmetros do Modelo

Em resumo, o modelo para cada l-ésimo locutor é denotado por $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$. Inicia-se no estado particular q_i para t=1, depende da distribuição do estado inicial e produz um símbolo de saída $O_t = v_k$, de acordo com $b_i(k)$. Em seguida, este se move para o estado q_j ou permanece no estado q_i de acordo com a_{ij}. Este processo de saída do símbolo e a transição para o próximo estado se repete até que o objetivo seja atingido (p.ex. quando o número de iterações estabelecido é alcançado). Em modelos "esquerda-direita", o processo se inicia no estado q_1 com $t = 1$ e termina quando é atingido T passos, ou seja, $t = T$. Assim, a partir da seqüência de observação $\mathbf{O^l} = \{O_1, ..., O_T\}$ e dos parâmetros necessários, obtêm-se o HMM referente a cada l-ésimo locutor [11, 26].

Seja $V = \{v_1, v_2, ..., v_M\}$ o alfabeto discreto utilizado para representar a seqüência de observação $\mathbf{O^l} = \{O_1, ..., O_T\}$. Define-se:

a) A probabilidade inicial como,

$$\pi_i = P\{q_i/t = 1\} \tag{4.1}$$

b) A matriz de probabilidade condicional ou ainda matriz de probabilidade de transição de estados como,

$$\mathcal{A} = [a_{ij}] = P\{q = j \; para \; t + 1/q = i \; para \; t\} \tag{4.2}$$

cujos elementos a_{ij} indicam a probabilidade de ocorrer a transição do estado q_i no instante de tempo t para o estado q_j no instante $t + 1$. A transição pode ser de tal forma que o processo permaneça no estado q_i em $t + 1$ ou se mova para o estado q_j.

c) A matriz de elementos $b_j(k)$ é definida como a matriz de função densidade de probabilidade das observações.

$$\mathcal{B} = [b_j(k)] = P\{v_k \; para \; t/q = j \; para \; t) \tag{4.3}$$

A probabilidade de ocorrência de uma dada seqüência será:

$$P\{O_1, O_2, .., O_T\} = \pi_i.\mathcal{A}.\mathcal{B} \tag{4.4}$$

Ou seja, o produto entre a probabilidade associada ao instante de tempo inicial (o início do processo), a probabilidade de transição entre os vários estados do processo e a função densidade de probabilidades das observações, para cada instante de tempo t. Obtendo-se assim, a probabilidade de ocorrência de uma dada seqüência de observação $\{O_1, O_2, .., O_T\}$.

Para melhor esclarecimento dos conceitos básicos acerca dos HMMs, em seguida será apresentado um exemplo bastante simples.

4.1.2 Exemplo de um processo discreto de Markov

LANÇAMENTO DE DUAS MOEDAS

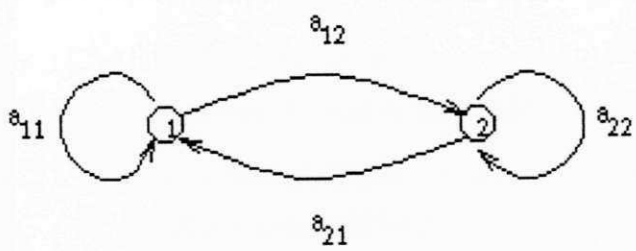


Figura 4.1: Modelo HMM para o lançamento de duas moedas

Para calcular a ocorrência da seqüência CARA, CARA, COROA:

CARA = 1, COROA = 2

ONDE;

1. $\pi_i = P\{q_i/t = 1\}$, a probabilidade inicial de, no instante $t = 1$, o modelo encontrar-se no estado i (iniciar o experimento pela moeda 1 ou 2).

2. $a_{ij} = P\{q_{t+1}=\mathrm{j}/q_t = i\}$, a probabilidade de fazer a transição do estado $q = i$ no instante t para o estado $q = j$ no instante $t + 1$.

3. $b_j(k) = P_j\{x = k\}$, a probabilidade de, estando no estado j, obter o resultado k (probabilidade de, estando na moeda 1 ou 2 obter o resultado CARA = 1 ou COROA = 2).

4. q_t, o estado do modelo no instante de tempo t (moeda selecionada no instante t).

Considerando o lançamento de duas moedas, para obter-se CARA, CARA e COROA, a probabilidade associada será:

$P\{x = 1/t = 1, x = 1/t = 2, x = 2/t = 3\}$

$= \pi_1 P_1\{x = 1\}P\{q_2 = 2/q_1 = 1\}P_2\{x = 1\}P\{q_3 = 2/q_2 = 2\}P_2\{x = 2\}$

Onde:

$\pi_1 = P\{q_1/t = 1\}$, a probabilidade inicial de, no instante $t = 1$, o modelo encontrar-se no estado 1 (iniciar o experimento pela moeda 1).

$b_1(1) = P_1\{x = 1\}$, a probabilidade de, estando no estado 1, obter o resultado 1 (probabilidade de, estando na moeda 1 obter o resultado $x = 1$ (CARA)).

$a_{12} = \mathrm{P}\{\mathrm{q}_2 = 2/q_1 = 1\}$, a probabilidade de fazer a transição do estado 1, no instante $t = 1$, para o estado 2, no instante $t = 2$.

$b_2(1) = P_2\{x = 1\}$, a probabilidade de, estando no estado 2, obter o resultado 1 (probabilidade de, estando na moeda 2 obter o resultado $x = 1$ (CARA)).

$a_{12} = \mathrm{P}\{\mathrm{q}_3 = 2/q_2 = 1\}$, a probabilidade de fazer a transição do estado 1, no instante $t = 2$, para o estado 2, no tempo $t = 3$.

$b_2(2) = P_2\{x = 2\}$, a probabilidade de, estando no estado 2, obter o resultado 2 (probabilidade de, estando na moeda 2 obter o resultado $x = 2$ (CAROA)).

OU

$= \pi_2 P_2\{x = 1\} P\{q_2 = 1/q_1 = 2\} P_1\{x = 1\} P\{q_3 = 2/q_2 = 1\} \mathrm{P}_2\{x = 2\}$

Onde:

$\pi_2 = P\{q_2/t = 1\}$, a probabilidade inicial de, no instante $t = 1$, o modelo encontrar-se no estado 2 (iniciar o experimento pela moeda 2).

$b_2(1) = P_2\{x = 1\}$, a probabilidade de, estando no estado 2, obter o resultado 1 (probabilidade de, estando na moeda 2 obter o resultado $x = 1$ (CARA)).

$a_{21} = \mathrm{P}\{\mathrm{q}_2 = 1/q_1 = 2\}$, a probabilidade de fazer a transição do estado 2, no instante $t = 1$, para o estado 1, no instante $t = 2$.

$b_1(1) = P_1\{x = 1\}$, a probabilidade de, estando no estado 1, obter o resultado 1 (probabilidade de, estando na moeda 1 obter o resultado $x = 1$ (CARA)).

$a_{12} = \mathrm{P}\{\mathrm{q}_3 = 2/q_2 = 1\}$, a probabilidade de fazer a transição do estado 1, no instante $t = 2$, para o estado 2, no instante $t = 3$.

$b_2(2) = P_2\{x = 2\}$, a probabilidade de, estando no estado 2, obter o resultado 2 (probabilidade de, estando na moeda 2 obter o resultado $x = 2$ (CAROA)).

OU AINDA

$= \pi_2 P_2\{x = 1\} P\{q_2 = 2/q_1 = 2\} P_2\{x = 1\} P\{q_3 = 2/q_2 = 2\} P_2\{x = 2\}$

Onde:

$\pi_2 = P\{q_2/t = 1\}$, a probabilidade inicial de, no instante $t = 1$, o modelo encontrar-se no estado 2 (iniciar o experimento pela moeda 2).

$b_2(1) = P_2\{x = 1\}$, a probabilidade de, estando no estado 2, obter o resultado 1 (probabilidade de, estando na moeda 2 obter o resultado $x = 1$ (CARA)).

$a_{22} = \mathrm{P}\{\mathrm{q}_2 = 2/q_1 = 2\}$, a probabilidade de permanecer no estado 2, no instante $t = 2$.

$b_2(1) = P_2\{x = 1\}$, a probabilidade de, estando no estado 2, obter o resultado 1 (probabilidade de, estando na moeda 2 obter o resultado $x = 1$ (CARA)).

$a_{22} = \mathrm{P}\{\mathrm{q}_3 = 2/q_2 = 2\}$, probabilidade de permanecer no estado 2, no instante $t = 3$.

$b_2(2) = P_2\{x = 2\}$, a probabilidade de, estando no estado 2, obter o resultado 2 (probabilidade de, estando na moeda 2 obter o resultado $x = 2$ (CAROA)).

4.2 Aplicação da Quantização Vetorial no HMM

Quando deseja-se utilizar um HMM com uma densidade discreta, a QV é requerida para mapear cada vetor de observação contínuo em um índice de dicionário discreto. Uma vez que o dicionário dos vetores tenha sido obtido, o mapeamento entre vetores contínuos e índices do dicionário é feito através da computação do mais próximo, i.é., o vetor contínuo é substituído pelo índice do vetor mais próximo do dicionário (em relação a alguma medida de distância) [11].

A idéia principal do projeto é a obtenção de uma técnica iterativa ótima para projeto do dicionário baseada em uma seqüência representativa de vetores de treinamento. O procedimento, basicamente divide os vetores de treinamento em M diferentes grupos (onde M é o tamanho do dicionário), cada conjunto é representado por um simples vetor (v_m, $1 < m < M$), que é geralmente o centróide dos vetores em um conjunto de treinamento localizado na m-ézima região, e então iterativamente otimiza a partição e o dicionário (i.é., o centróide de cada partição).

Está associada à quantização vetorial uma penalidade de distorção, visto que está sendo representada uma região completa do espaço vetorial por um simples vetor. Portanto, é vantajoso tomar uma penalidade de distorção tão menor quanto possível. Entretanto, isto implica em um dicionário de tamanho elevado, o que leva a problemas de implementação de HMMs com um grande número de parâmetros [11].

Embora a distorção decresça quando M cresce, não se pode utilizar um valor muito grande de M, pois assim têm-se um número elevado de bits/amostra, o que não é desejado. No capítulo seguinte serão mostrados os valores de M utilizados para teste e qual deles proporcionou um melhor desempenho do sistema.

Para implementação de HMMs de densidades discretas em reconhecimento de locutor, assume-se que as entradas do modelo são seqüências de símbolos discretos escolhidas de um alfabeto finito. Estes símbolos discretos são obtidos usando o método da

quantização vetorial [30, 31] de vetores LPC cujas características serão analisadas em seguida.

4.2.1 Análise das características LPC

A análise espectral é uma das formas mais usuais de obter-se os vetores O_t de uma seqüência de observação $\mathbf{O}^l = \{O_1, ..., O_T\}$ para as amostras de voz de um l-ésimo locutor. O tipo de análise espectral aqui usada é chamada Codificação por Predição Linear (LPC) cujo diagrama de blocos é mostrado na Figura 4.2 [11].

O sinal $\tilde{s}(n)$, para cada l-ésimo locutor, é segmentado em T blocos, de N_A amostras, os quais são obtidos a partir de um janelamento pela função da Hamming. A cada t-ésimo bloco ($1 \leq t \leq T$) é determinado um vetor de autocorrelação $R_{rt}(m)$ com o qual se obtêm os respectivos coeficientes LPC, $c_t(m)$. Estes coeficientes constituem cada vetor O_t da seqüência de observação $\mathbf{O}^l = \{O_1, ..., O_T\}$ [11].

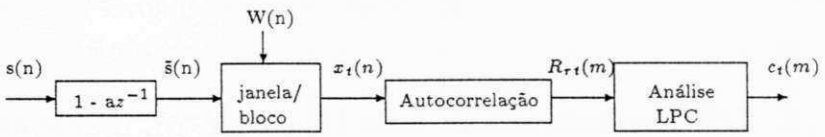


Figura 4.2: Diagrama de blocos para análise das características do locutor para um reconhecedor HMM.

Análise:

$\tilde{s}(n) = s(n) - as(n-1)$ $(a = 0.95)$,

$x_t(n) = \tilde{s}(n).W(n)$, $0 \leq n \leq N_A$, $1 \leq t \leq T$

$R_{rt}(m) = x_t(n)x_t(n+m)$, $0 \leq m \leq p$

$c_t(\mathrm{m})$ = coeficientes LPC, $0 \leq m \leq p$

Os passos do processamento são os seguintes:

1. Pré-ênfase: As amostras são pré-enfatizadas por um filtro de 1ª ordem cuja função de transferência é $1 - 0.95z^{-1}$. Tal filtro produz um nivelamento do espectro, procurando compensar os efeitos da radiação do som pelos lábios [6].

2. Agrupamento em blocos: São usados blocos de N_A amostras consecutivas de voz ($N_A = 240$ correspondendo a 30ms de sinal) os quais contêm 50% de superposição com relação ao bloco anteriormente analisado.

3. Janelamento do bloco: Cada bloco de N_A amostras é multiplicado por uma janela de Hamming - W(n) para minimizar os efeitos adversos da separação entre os blocos [11].

4. Análise de Autocorrelação: Para cada bloco de amostras de voz é calculada a função de autocorrelacão fornecendo um conjunto de (p+1) coeficientes, onde p é a ordem da análise LPC desejada (neste trabalho foi utilizado $p = 12$ [28]).

5. Análise LPC: Para cada bloco, um vetor de coeficientes LPC é obtido, utilizando o método recursivo de Levinson-Durbin [6].

4.2.2 Quantização Vetorial dos coeficientes LPC

Obtendo-se o conjunto de vetores LPC, c_t, $t = 1, 2, ..., T$. A idéia principal da quantização vetorial é determinar um conjunto ótimo de vetores de um dicionário que represente os vetores LPC, $\hat{c}_t(m)$, $m = 1, 2, ..., p$, tal que para um dado p, a medida de distorção obtida pela troca do conjunto de vetores de treinamento, c_t, em relação ao dicionário, seja mínima [13].

Dito de uma maneira mais formal, define-se $d(\hat{c}_t(m), c_t(m))$ como a distância entre dois vetores LPC, $\hat{c}_t$ e c_t. Assim o objetivo da quantização vetorial é encontrar o conjunto, $\hat{c}_t$, tal que [13]

$$||D_M|| = \min_{\hat{c}_t}\{\frac{1}{T}\sum_{i=1}^{T}\min_{1\leq m\leq p}[d(\hat{c}_t, c_t)]\} \tag{4.5}$$

seja satisfeita. O valor $||D_M||$ é a medida de distorção do quantizador vetorial.

Neste trabalho foi utilizada como medida de distorção, a medida de distância do Erro Quadrático (citada no capítulo anterior) [11].

O procedimento da quantização vetorial é descrito a seguir [10, 13]:

1. Dado um vetor de parâmetros de entrada, calcula-se a distância deste vetor com cada centróide do dicionário;
2. Compara-se as distâncias, determinando a menor;
3. Seleciona-se o centróide correspondente, como vetor representante do vetor de parâmetros de entrada;
4. Utiliza-se o código do centróide como referência do vetor de entrada.

4.3 Aplicação de HMMs em Verificação de Locutor

O processo de reconhecimento de padrões, mais especificamente locutor, utilizando Modelos de Markov Escondidos - HMMs, se resume no seguinte: classificar as elocuções de um conjunto de L locutores, utilizando suas respectivas seqüências de observação $\mathbf{O}^l = \{O_1, ..., O_T\}$ de acordo com seu respectivo modelo λ_l. Determina-se a medida de similaridade, P_l, entre uma seqüência de observação $\mathbf{O}^l$ e o modelo associado λ_l, através do cálculo da probabilidade de que o modelo λ_l represente as ocorrências associadas à seqüência de observação $\mathbf{O}^l$. O locutor será então classificado se P_l for maior que um dado limiar. O processo de verificação de locutor envolve duas fases distintas [12]: treinamento e verificação, descritas em seguida.

4.3.1 Fase de Treinamento

Nesta fase é feita a estimação dos parâmetros dos modelos $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$, um modelo para cada l-ésimo locutor (Figura 4.3) [26].

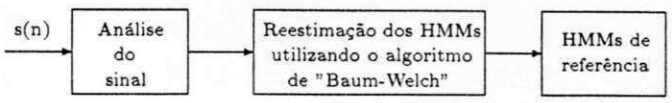


Figura 4.3: Diagrama de blocos que representa a fase de treinamento da verificação de locutor utilizando HMM.

Desde que exista um procedimento de reestimação convergente para o modelo de densidades discretas, teoricamente é possível escolher-se aleatoriamente valores iniciais para cada um dos parâmetros do modelo (sujeitos às restrições iniciais) e deixar o procedimento de reestimação determinar os valores ótimos (máxima verossimilhança). No caso específico de modelos de Markov, a estimação é realizada usando o processo iterativo de Baum-Welch [11, 12].

O processo de reestimação de "Baum-Welch" pode ser descrito através dos seguintes passos:

1. Atribuição inicial dos valores para os parâmetros do modelo $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$ e para a probabilidade P_l;

2. Reestimação dos parâmetros do modelo através do algoritmo de reestimação de "Baum-Welch", cujas equações serão descritas a seguir, obtendo-se $\overline{\lambda_l}$;

3. Cálculo da probabilidade $\overline{P_l}$ associada ao modelo $\overline{\lambda_l}$ reestimado e comparação com a probabilidade anteriormente calculada P_l;

4. Se $\overline{P_l} - P_l \leq \delta$ (limiar), o processo de reestimação é finalizado. Caso contrário, retorna-se ao passo 2.

As atribuições iniciais dos parâmetros do modelo, devem obedecer à regras simples, de forma a satisfazer as restrições do modelo "esquerda-direita". O vetor de probabilidade inicial $\pi_i = \{1, ..., 0\}$, visto que o modelo é "esquerda-direita" e portanto, sempre

é inicializado no estado 1, não sendo necessário reestimá-lo. A matriz $\mathcal{A} = [a_{ij}]$ inicial é gerada obedecendo a seguinte restrição: $a_{ij} = 0$, $j < i, j > i + 2$, já que para modelos "esquerda-direita" um estado visitado no instante de tempo t não poderá ser retornado num instante de tempo posterior. Esta restrição deverá se manter até o final do processo de reestimação. Para matriz $\mathcal{B} = [b_j(k)]$, assume-se que todos os símbolos nos estados são "igualmente prováveis" e $b_j(k)$ é inicializado com $1/M$ para todo j, k, para simplificar.

As equações do método de reestimação de "Baum-Welch" são [11, 12]:

1. $\overline{a_{ij}}$ = (número esperado de transições do estado q_i para o estado q_j)/(número esperado de transições no estado q_i)
2. $\overline{b_j(k)}$ = (número esperado de vezes no estado j observando o símbolo v_k)/(número esperado de vezes no estado j)

ou seja,

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_t(i)}, \quad 1 \leq i \leq N,\ 1 \leq j \leq N \tag{4.6}$$

$$\overline{b_j(k)} = \frac{\sum_{t=1, O_t = v_k}^{T} \alpha_t(j) \beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j) \beta_t(j)}, \quad 1 \leq j \leq N,\ 1 \leq k \leq M \tag{4.7}$$

De forma que:

$$\sum_{j=1}^{N} a_{ij} = 1;\ \sum_{k=1}^{M} b_j(k) = 1;\ \sum_{i=1}^{N} \pi_i = 1,\ a_{ij} \geq 0;\ b_j(k) \geq 0;\ \pi_i \geq 0 \tag{4.8}$$

Cada parâmetro $b_j(O_t)$, $1 \leq j \leq N$ e $1 \leq t \leq T$, é obtido a partir da comparação, em relação a um dado estado j e variando t, com os valores da matriz $[b_j(k)]$ referentes ao índice k do símbolo associado ao vetor O_t no mesmo estado j. Atribui-se a $b_j(O_t)$ o valor de $b_j(k)$ correspondente ao referido símbolo v_k, no estado j.

A probabilidade $\alpha_t(i)$ é denominada probabilidade de avanço ("forward probability"), pois está associada à ocorrência de uma dada seqüência de observação $\mathbf{O}^1 =$

$\{O_1, ..., O_T\}$, segundo o tempo crescente (iniciando em $t = 1$ indo até $t = T$), sendo formulada em [11] como:

1. Inicialização:

$$\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N \tag{4.9}$$

2. Indução:

$$\alpha_{t+1}(j) = \left\{ \sum_{i=1}^{N} \alpha_t(i) a_{ij} \right\} b_j(O_{t+1}), \quad 1 \leq t \leq T-1, \ 1 \leq j \leq N \tag{4.10}$$

A probabilidade P_l associada ao modelo $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$, referente ao l-ésimo locutor, é determinada a partir da seguinte expressão [11]:

Para algum t, $1 \leq t \leq T$,

$$P_l = Prob(\mathbf{O}^1/\lambda_l) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{4.11}$$

Fazendo-se $t = T - 1$, obtem-se:

$$P_l(\mathbf{O}^1/\lambda_l) = \sum_{i=1}^{N} \alpha_T(i) \tag{4.12}$$

Onde:

$$\alpha_T(i) = P_l(O_1, ..., O_T/\lambda_l) \tag{4.13}$$

O cálculo das probabilidades de avanço ("forward"), inicia-se atribuindo-se ao estado q_i o vetor inicial O_1. O passo de indução é o ponto principal do cálculo da probabilidade de avanço, como ilustrado na Figura 4.4.

Esta Figura mostra como o estado q_j pode ser alcançado no instante de tempo $t+1$ a partir dos N possíveis estados, q_i, $1 \leq i \leq N$, no instante de tempo t associado à seqüência de vetores $\mathbf{O}^1$. Assim $\alpha_t(i)$ é a probabilidade de que os vetores $\{O_1, ..., O_T\}$

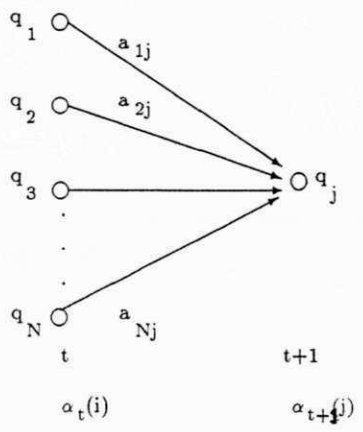


Figura 4.4: Ilustração da seqüência de operações necessárias para computação da variável "forward" $\alpha_{t+1}(j)$.

tenham ocorrido estando no estado q_i no instante t. O produto $\alpha_t(i)a_{ij}$ é então a probabilidade de que o evento $\{O_1, ..., O_T\}$ seja observado a partir de q_i no instante t, tal que o estado q_j seja alcançado no instante $t+1$. Somando esse produto ao longo dos N estados possíveis q_i, $1 \leq i \leq N$ no instante t obtem-se a probabilidade associada ao estado q_j no instante $t+1$. Desde que isto seja feito e q_j seja conhecido, é fácil ver que $\alpha_{t+1}(j)$ é obtido de acordo com o vetor O_{t+1}, no estado j, ou seja, multiplicando as quantidades somadas pela probabilidade $b_j(O_{t+1})$. A computação da Equação (4.7) é realizada para todos os j-ésimos estados, $1 \leq j \leq N$, para um dado t; a computação então é iterativa para $t = 1, 2, ..., T-1$.

O cálculo da probabilidade de avanço é baseado na estrutura de treliça mostrada na Figura 4.5. Desde que há somente N estados (nós para cada instante de tempo na treliça), todas as possíveis seqüências de estado serão agrupadas dentro desses N nós, não importando o tamanho da seqüência de observação. No instante $t = 1$, o primeiro instante de tempo na treliça, referente ao primeiro vetor O_1 de uma dada seqüência de observação $\mathbf{O}^l$, é necessário calcular valores de $\alpha_1(i)$, $1 \leq i \leq N$. Para os instantes

$t = 2, 3, ..., T$, é necessário calcular os valores de $\alpha_t(j)$, $1 \leq j \leq N$, onde cada um dos cálculos envolve somente N valores anteriores de $\alpha_{t-1}(j)$, uma vez que cada um dos N pontos da grade é obtido a apartir dos mesmos N pontos da grade no instante de tempo anterior [11].

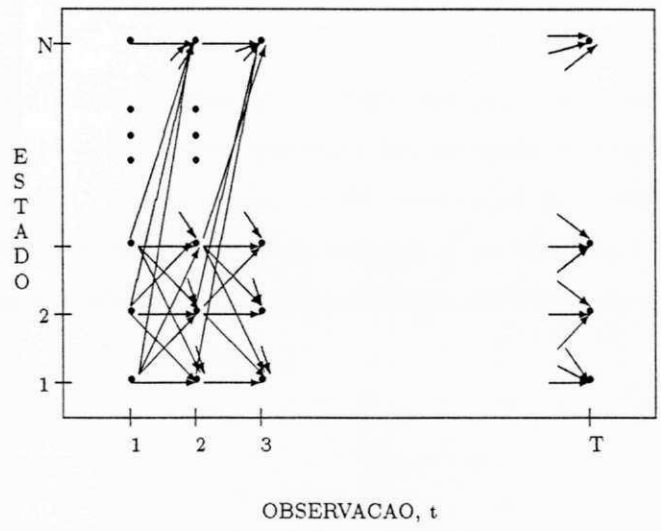


Figura 4.5: Implementação da computação de $\alpha_t(i)$ em termos de uma treliça de observações t e estados i.

De forma similar, $\beta_t(i)$ é denominada probabilidade de retrocesso ("backward probability"), pois está associada à ocorrência da seqüência de observação $\mathbf{O}^1 = \{O_1, ..., O_T\}$ segundo o tempo decrescente (iniciando em $t = T$ indo até $t = 1$) sendo definida como [11]:

$$\beta_t(i) = P_l(O_T, O_{T-1}, ..., O_t/\lambda_l) \tag{4.14}$$

ou, seja, a probabilidade da seqüência de observação parcial do instante de tempo $t+1$ até o fim, dado o estado q_i no instante de tempo t e o modelo λ_l. Assim pode-se obter $\beta_t(i)$ indutivamente da seguinte forma [11]:

1. Inicialização:

$$\beta_T(i) = 1, \quad 1 \leq i \leq N \tag{4.15}$$

2. Indução:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, ..., 1, \ 1 \leq i \leq N \tag{4.16}$$

O passo 1, inicialização, define arbitrariamente $\beta_T(i) = 1$ para todo i. O passo 2, ilustrado na Figura 4.6, mostra que para ter ocorrido o estado q_i no instante de tempo t, levando-se em conta a seqüência de observação no instante de tempo $t+1$, é necessário considerar todos os possíveis estados q_j no instante $t+1$, considerando a transição de q_i para q_j (o termo a_{ij}), como também a observação O_{t+1} no estado j (O termo $b_j(O_{t+1})$).

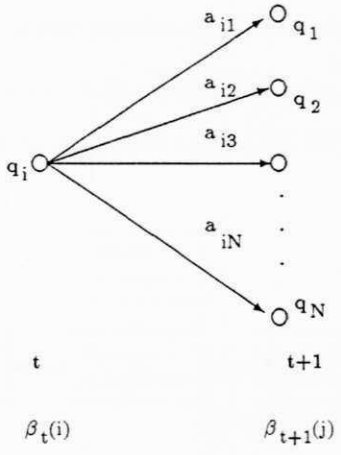


Figura 4.6: Ilustração da seqüência de operações necessárias para computação da variável "backward" $\beta_t(i)$.

Não há nenhuma técnica iterativa ótima para reestimar os parâmetros do modelo $\mathcal{A}$, $\mathcal{B}$ e π, os quais maximizam $P_l(\mathbf{O}^l/\lambda_l)$, dada uma seqüência de observação finita como dado de treinamento. Entretanto, um método iterativo proposto por Baum e Welch

em [32] é utilizado para escolher λ_l tal que $P_l(\mathbf{O}^1/\lambda_l)$ seja localmente máxima. Eles mostraram que o modelo reestimado $\overline{\lambda_l} = (\overline{\mathcal{A}}, \overline{\mathcal{B}}, \pi)$ (em modelos "esquerda-direita" π não precisa ser reestimado) é melhor ou igual ao modelo estimado anteriormente λ_l, desde que $P_l(\mathbf{O}^1/\overline{\lambda_l}) \geq P_l(\mathbf{O}^1/\lambda_l)$. Assim, utiliza-se $\overline{\lambda_l}$ no lugar de λ_l repetindo o processo de reestimação para uma dada seqüência observada, $\mathbf{O}^1$, até que algum ponto limite é atingido, ou seja, é atingido um número de iterações desejado ou o valor de probabilidade escolhido. O resultado final ou estimado é denominado estimação de máxima verossimilhança do HMM [11].

4.3.2 Fase de Verificação

Nesta fase, é realizada a estimação da probabilidade de ocorrência de uma dada seqüência de observação $\mathbf{O}^1 = \{O_1, ..., O_T\}$, associada ao modelo $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$, obtido durante a fase de treinamento, correspondente ao l-ésimo locutor.

Uma vez que os HMMs tenham sido treinados para cada locutor, a estratégia de verificação é direta, como descrito na Figura 4.7 [33, 29].

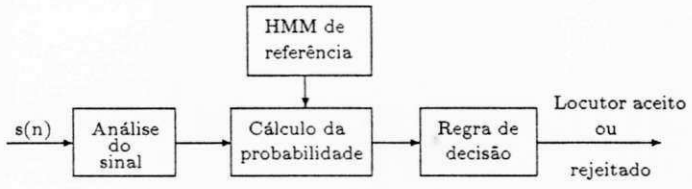


Figura 4.7: Diagrama de blocos que representa a fase de verificação do locutor utilizando HMM.

As etapas necessárias à verificação de um l-ésimo locutor são as seguintes:

1. Análise do sinal, realizada através dos seguintes passos: aquisição, pré-ênfase e janelamento em T blocos, contendo cada bloco, N_A amostras.

2. Análise LPC de uma elocução de voz, como sinal de entrada, obtendo-se a seqüência de observação $\mathbf{O}^{\mathbf{l}} = \{O_1, ..., O_T\}$, associada ao l-ésimo locutor.

3. Geração de uma tabela de códigos associadas à seqüência de observação $\mathbf{O}^{\mathbf{l}} = \{O_1, ..., O_T\}$, através da quantização vetorial.

4. Cálculo da probabilidade associada ao modelo, a partir dos parâmetros do modelo de referência $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$ já obtidos durante a fase de treinamento deste locutor.

O procedimento para o cálculo da probabilidade $P(\mathbf{O}^{\mathbf{l}}/\lambda_1)$ é o mesmo já mostrado anteriormente na Equação (4.11), descrito a seguir:

Fazendo-se $t = T - 1$, obtem-se:

$$P_l(\mathbf{O}^{\mathbf{l}}/\lambda_l) = \sum_{i=1}^{N} \alpha_T(i) \tag{4.17}$$

Onde:

$$\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N \tag{4.18}$$

$$\alpha_{t+1}(j) = \left\{ \sum_{i=1}^{N} \alpha_t(i) a_{ij} \right\} b_j(O_{t+1}), \quad 1 \leq t \leq T-1, \ 1 \leq j \leq N \tag{4.19}$$

Os coeficientes a_{ij} e π_i correspondem, exatamente, aos valores de referência da matriz $\mathcal{A}$ e vetor π, respectivamente.

Os coeficientes $b_j(O_t)$ são obtidos a partir da matriz $\mathcal{B} = [b_j(k)]$, da seguinte forma: a cada vetor O_t de um l-ésimo locutor corresponde, após a quantização vetorial, um determinado índice do quantizador vetorial(símbolo v_k). Cada coeficiente $b_j(k)$ representa a probabilidade de ocorrência de um dado símbolo v_k, no estado j. Assim, cada coeficiente $b_j(O_t)$ corresponde ao valor da probabilidade do símbolo associado àquele estado j.

O exemplo hipotético abaixo ilustra melhor o cálculo de $b_j(O_t)$.

Para um dado locutor tem-se a seqüência de observação $\mathbf{O}^l = \{O_1, ..., O_T\}$, onde $O_t = \{c_1, c_2, ..., c_p\}$ é uma seqüência de vetores LPC.

Após a quantização vetorial com 256 níveis tem-se, supostamente, os símbolos com índices = $\{12, 25, 256, ..., 32\}$ associados à seqüência de observação.

Assim, ao vetor O_1 está associado o símbolo de índice = 12, ao vetor O_2 o símbolo de índice = 25 e assim sucessivamente até O_T.

A matriz $\mathcal{B}$ fornece o valor de probabilidade associado a cada um desses símbolos para cada estado. Portanto o valor de $b_1(O_1)$, será igual a $b_1(12)$, $b_1(O_2)$, será igual a $b_1(25)$ e assim sucessivamente até $b_1(O_T)$.

O mesmo procedimento é tomado para o cálculo de $b_j(O_t)$ nos demais estados do HMM.

O locutor é aceito se seu valor de probabilidade $P_l(\mathbf{O}^l/\lambda_l)$ é maior que um dado limiar, caso contrário, o locutor é rejeitado.

4.4 Discussão

Este capítulo procurou descrever o Modelo de Markov Escondido de densidades discretas e sua utilização em verificação de locutor. Para um melhor entendimento da teoria associada a este modelo, foi utilizado inicialmente um exemplo bastante simples (lançamento de duas moedas), que pode caracterizar de forma clara os parâmetros associados ao modelo. Em seguida, foram descritas as duas tarefas necessárias para a verificação de locutor utilizando os Modelos de Markov Escondidos (HMMs). Estas tarefas são: o treinamento e a verificação. A primeira tarefa é realizada através do algoritmo de reestimação de "Baum-Welch", de forma a obter-se um modelo referente a cada locutor. A segunda tarefa é mais simples: a partir dos parâmetros de voz de um dado locutor a ser verificado, é calculado o valor de probabilidade associado a este locutor, utilizando os parâmetros do modelo já armazenados para este locutor durante a fase de treinamento. Se o valor de probabilidade obtido for maior que um dado limiar, o locutor é aceito, caso contrário é rejeitado.

Capítulo 5

Descrição do Sistema HMM-QV Proposto e Resultados Experimentais

5.1 Descrição do Sistema

O sistema de verificação de locutor pode ser descrito pelo diagrama de blocos apresentado na Figura 5.1 [33]:

De acordo com a Figura 5.1, observa-se que a verificação de locutor utilizando HMM, como dito no capítulo anterior, consta de duas fases, treinamento (modo 1) e verificação (modo 2). Na fase de treinamento, o conjunto de observações de treinamento $\mathbf{O}^l = \{O_1, ..., O_T\}$ de cada locutor é primeiro, representado por símbolos através da quantização vetorial. Após a quantização vetorial, uma elocução é representada por uma seqüência de códigos, onde cada bloco de amostras do sinal, está associado a um código da tabela de vetores-código (dicionário) da quantização vetorial. Em seguida, os parâmetros do modelo são reestimados através do algoritmo de "Baum-Welch", obtendo-se um modelo de referência $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$ para cada locutor.

Na fase de verificação, os conjuntos de parâmetros (coeficientes LPC) do locutor

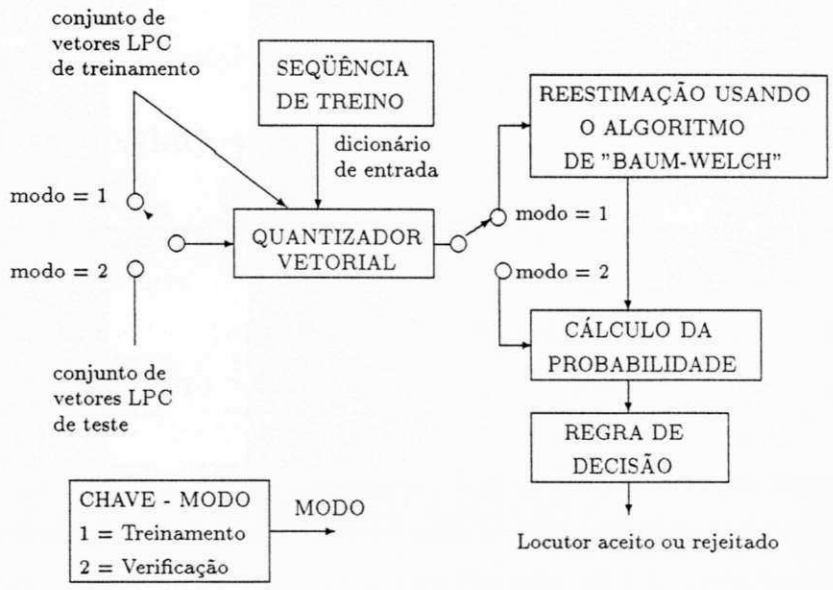


Figura 5.1: Diagrama de blocos para verificação de locutor utilizando HMM-QV.

a ser verificado, são, novamente, representados por um conjunto símbolos através da quantização vetorial. Após a quantização vetorial, uma elocução é representada por uma seqüência de códigos, onde cada bloco de amostras do sinal, está associado a um código da tabela de vetores-código (dicionário) da quantização vetorial. Em seguida, para o locutor a ser verificado é calculado um valor de probabilidade P_l (descrito na Equação (4.12), capítulo 4) utilizando-se os parâmetros do modelo de referência deste locutor já calculados durante a fase de treinamento. O locutor é aceito se seu valor de probabilidade é maior que um dado limiar, caso contrário o locutor é rejeitado.

5.2 Escolha dos Parâmetros do Modelo

O treinamento dos HMMs envolve inicialmente a escolha dos valores de N, M, inicialização de π, a_{ij} e $b_j(k)$. Tem sido mostrado em alguns sistemas de reconhecimento

de locutor que a escolha de boas estimativas iniciais é essencial e conduz a melhores resultados [11, 13].

5.2.1 Base de dados

O sistema de verificação de locutor proposto foi avaliado utilizando uma população composta de 10 locutores, cinco mulheres e cinco homens. Nos testes de verificação foram usados entretanto, apenas 6 locutores (3 mulheres e 3 homens). Cada locutor dispõe de uma senha e uma sentença que permitem o seu acesso ao sistema. Os dados de treinamento e teste foram obtidos em uma única sessão de gravação e cada locutor pronunciou sua sentença dez vezes (para verificar os índices de falsa rejeição) e a sentença dos outros cinco locutores 5 vezes (para verificar os índices de falsa aceitação). Cada sentença usada é muito pequena, com duração de 1.5 segundos. As sentenças utilizadas foram:

1. Quero usar a máquina. (Figura 5.2).
2. Desejo usar o sistema. (Figura 5.3).
3. Eu serei um vencedor.(Figura 5.4).
4. Sou usuário do sistema. (Figura 5.5).
5. O dia será belíssimo. (Figura 5.6).
6. Minha senha é secreta. (Figura 5.7).
7. Minha meta é vencer.
8. Farei o máximo hoje.
9. Tudo sairá muito bem
10. A ciência tem futuro.

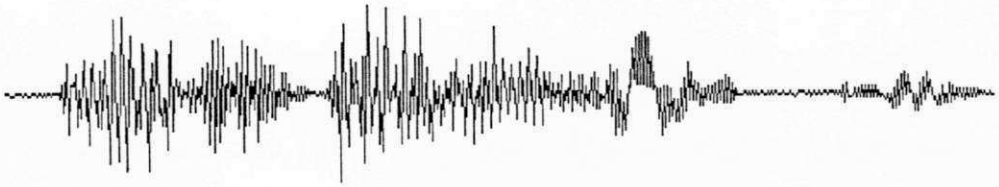


Figura 5.2: Forma de onda da sentença pronunciada pelo locutor 1 (L1)

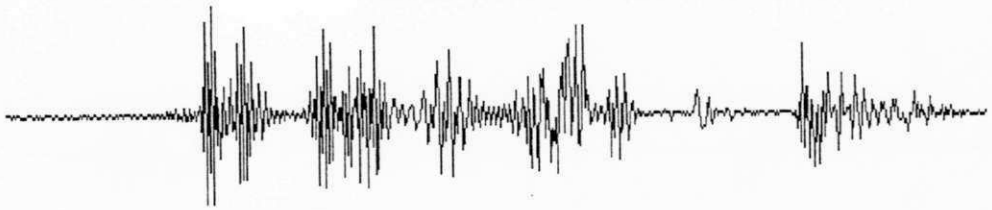


Figura 5.3: Forma de onda da sentença pronunciada pelo locutor 2 (L2)

Figura 5.4: Forma de onda da sentença pronunciada pelo locutor 3 (L3)

Figura 5.5: Forma de onda da sentença pronunciada pelo locutor 4 (L4)

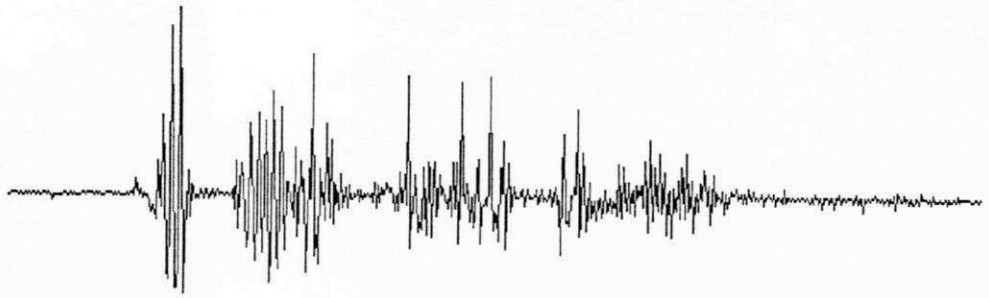


Figura 5.6: Forma de onda da sentença pronunciada pelo locutor 5 (L5)

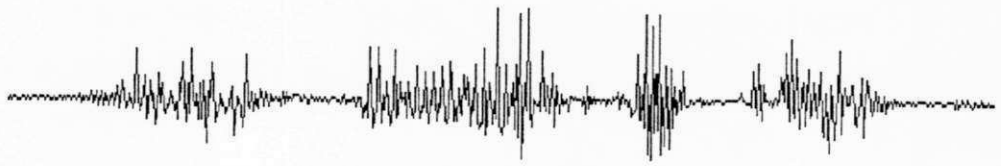


Figura 5.7: Forma de onda da sentença pronunciada pelo locutor 6 (L6)

As Figuras 5.2, 5.3, 5.4, 5.5, 5.6 e 5.7 mostram as formas de onda das sentenças pronunciadas pelos locutores L1, L2, L3, L4, L5 e L6, respectivamente.

A escolha das sentenças obedeceu à apenas algumas restrições para que possuissem, em média, o mesmo tamanho e o mesmo número de intervalos de silêncio. Cada locutor pronunciou um total de 10 (10 repetições de sua sentença) + 5x5 (5 repetições das sentenças dos outros 5 locutores de teste) = 35 repetições das sentenças.

5.3 Algoritmo do quantizador vetorial

O algoritmo utilizado para quantização vetorial foi O LBG, gerado com um dicionário inicial constituído de um conjunto de amostras iniciais da seqüência de treino, tomadas de forma aleatória. Em seguida, foi construído o dicionário do quantizador a partir de uma medida de distância para determinar o vetor mais próximo.

A medida de distância utilizada no quantizador vetorial foi a medida de Distância do Erro Quadrático, cujas características foram descritas no capítulo 3 [19].

A seqüência de voz (seqüência de treino) utilizada para geração do dicionário do quantizador vetorial era composta das dez sentenças, citadas anteriormente, pronunciadas por seus respectivos locutores. A seqüência de treino (Figura 5.8) foi gravada em uma sessão diferente da utilizada para a gravação da sentença de cada locutor individualmente.

Optou-se pela utilização de uma única seqüência de treino para todos os locutores, em lugar de uma seqüência de treino para cada locutor, visando obter um dicionário capaz de representar, de uma forma geral, um conjunto qualquer de homens e mulheres. Isto é possível pois, teoricamente, o conjunto de locutores utilizados neste trabalho representa uma população contendo todas as características de voz de locutores masculinos e femininos. Desta forma, é possível ampliar o número de usuários do sistema de verificação de locutor não sendo necessário para tanto, gerar um novo dicionário. Além disso, a utilização de um único dicionário para todos os locutores reduz o total de memória utilizada, em comparação com o uso de um dicionário para cada locutor.

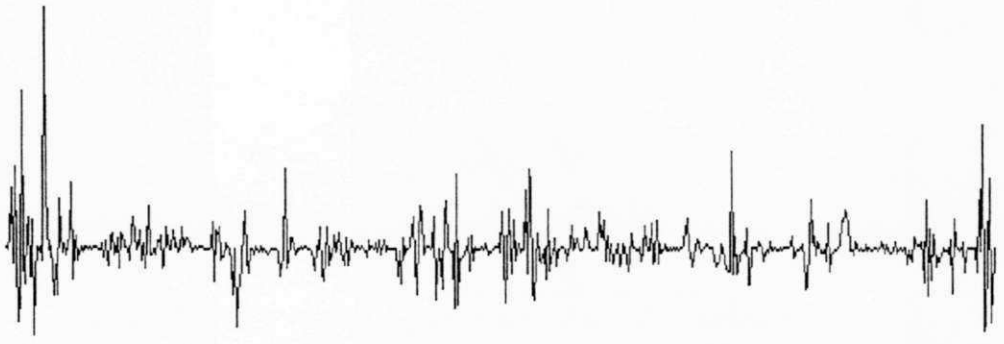


Figura 5.8: Seqüência de treino do quantizador vetorial, gerada pelos locutores L1, L2, L3, L4, L5, L6, L6, L7, L8, L9 e L10, com cada um pronunciando sua respectiva sentença.

5.3.1 Escolha da dimensão do quantizador

Os parâmetros escolhidos para representar o sinal de voz, como dito no capítulo 4, foram os coeficientes LPC (Coeficientes de Predição Linear) obtidos através do algoritmo de Levinson Durbin (descrito no apêndice A) [5, 9, 19]. A ordem do preditor ($p = 12$), neste trabalho, corresponde a dimensão do quantizador. A Figura 5.9 mostra a taxa de erro, para os locutores considerados, em função do número de coeficientes LPC.

Verifica-se através da Figura 5.9 que inicialmente foram utilizados 9 coeficientes LPC + energia do sinal (dimensão do quantizador igual a 9). Porém, os resultados obtidos não foram satisfatórios, havendo erros de falsa rejeição bastante elevados para alguns locutores. Obteve-se taxa de erro de 50% para L3, 40% para L6, 30% para L2,

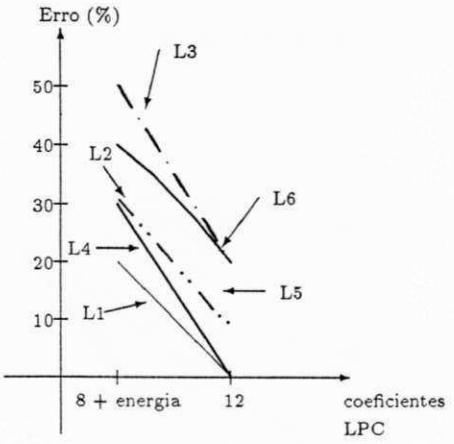


Figura 5.9: Taxa de erro do locutor x dimensão do quantizador (para N=5 e M=256)

L4 e L5 e 20% para L1. Por fim, foram feitos testes utilizando coeficientes LPC de ordem 12 e os resultados obtidos foram bem melhores, como mostra a Figura, obtendo-se uma taxa de falsa rejeição de até 0% para os locutores L1 e L4, 10% para L2 e L5 e 20% para L3 e L6. Não foram feitos testes utilizando um número de coeficientes LPC acima de 12, pois estes valores, segundo a maioria das referências [9, 11, 13, 29], não fornecem resultados muito diferentes dos obtidos com 12 coeficientes LPC, para verificação de locutor.

5.3.2 Escolha do número de níveis do quantizador (símbolos do alfabeto, M)

Foram realizados vários testes para escolha do número de níveis do quantizador (valor de M) de forma a obter um resultado que melhor se adaptasse as características do sistema proposto. Inicialmente, os testes foram realizados para $M = 32$ e $M = 64$, porém os resultados obtidos não foram satisfatórios. Em seguida, foi utilizado $M = 128$. Por fim, foi utilizado $M = 256$, o qual forneceu os melhores resultados.

A Figura 5.10 mostra as taxas de erro em função do número de níveis do quantizador, para a dimensão 12, referentes aos 6 locutores utilizados para teste.

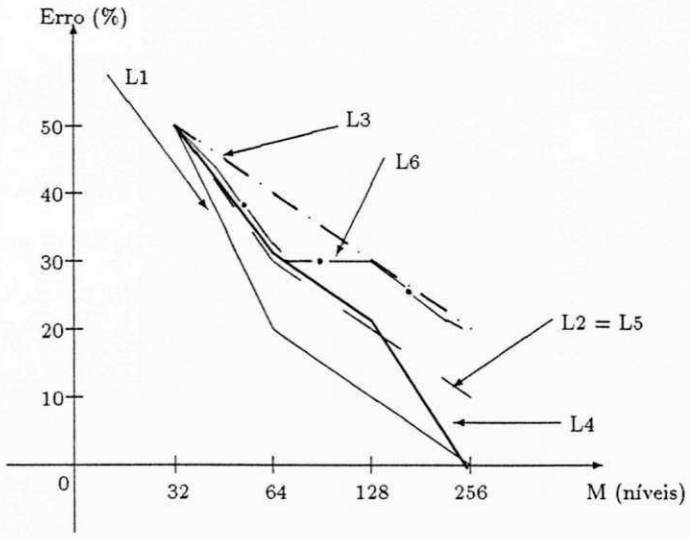


Figura 5.10: Taxa de erro do locutor x número de níveis do quantizador (para $N = 5$ e dimensão=12)

Verifica-se através da Figura 5.10 que para $M = 32$ a taxa de erro é bastante elevada (50%) para todos os locutores. $M = 64$ fornece uma taxa de erro de 30%, o que já representa uma melhora razoável. Porém, para alguns locutores o índice de falsa aceitação (locutor falso é considerado verdadeiro) é considerável. $M = 128$ apresenta melhores resultados para alguns locutores, com uma taxa de erro de 10% para os locutores 1 e 4 e 20% para os locutores 2 e 5, entretanto para os locutores 3 e 6 a taxa de erro é 30%. Os valores de $M = 256$, como mostra a Figura 5.10, apresentam resultados bastante satisfatórios apesar das taxas de erro para os locutores 3 e 6 serem de 20%.

O quantizador com $M = 32$ apresenta uma distorção maior, bem como uma distribuição não uniforme da ocupação das classes (Tabela A1, Apêndice A) e seus resultados, consequentemente, são os piores. Pois, com esse número de níveis há uma compressão muito grande do sinal de voz, fazendo com que redundâncias deste sinal sejam, muitas vezes, eliminadas. $M = 256$ apresenta os melhores resultados, como mostra a tabela A3 (Apêndice A).

Procurou-se obter, com a quantização vetorial, uma compressão de dados que fornecesse uma baixa taxa de bits/amostra ($\leq$ 1bit/amostra). Dentro desse contexto poder-se-ia ainda utilizar um maior número de níveis, por exemplo $M = 512$, 1024 ou 2048. Estes valores de M poderiam fornecer melhores resultados que os obtidos para $M = 256$, entretanto isso levaria a um aumento considerável do volume de dados, acarretando em um custo computacional elevado.

5.4 Escolha do número de estados do HMM (N)

De uma forma geral, existem dois métodos para escolha do valor de N em sistemas de reconhecimento de palavras isoladas. Uma opção seria tomar o número de estados correspondendo ao número de sons de cada palavra pronunciada pelo locutor. Assim, seria necessário utilizar uma quantidade muito grande de estados. A segunda opção seria tomar o número de estados correspondendo ao número médio de observações em uma versão falada das palavras da sentença, também chamado modelo de Bakis [34]. Desta forma, cada estado corresponderia a um intervalo de observação, ou seja, em torno de 30ms (240 amostras) para o sistema proposto. Esta opção também implicaria na utilização de um número de estados bastante elevado, em torno de 50. Como o propósito deste trabalho não é o reconhecimento de palavras isoladas e sim a verificação de locutor dependente do texto, não é necessário verificar a variação dos sons de cada palavra, mas a forma como cada locutor as pronuncia. Portanto, a partir da bibliografia disponível, [1, 11, 13, 22, 26, 27, 33] foram testados valores de N, variando de 2 a 8, (Figura 5.11) para verificar o desempenho do sistema.

Pode-se observar através da Figura 5.11, que valores de N muito pequenos, ($N = 2$)

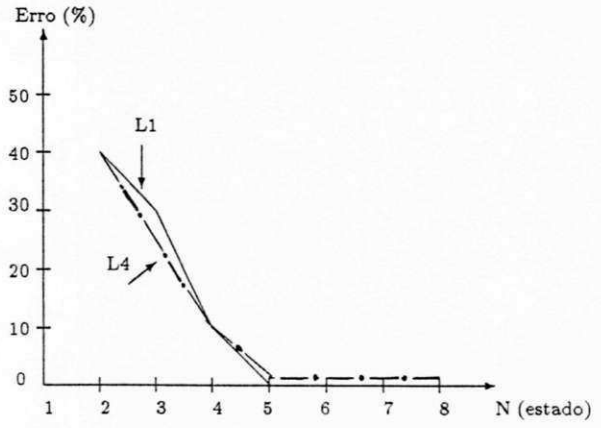


Figura 5.11: Taxa de erro do locutor x número de estados do HMM (para $M = 256$ e dimensão=12)

não apresentam bons resultados, o que já era esperado em virtude desse valor não ser capaz de representar de forma satisfatória as variações do sinal de voz. Entretanto, para $N = 5$, 6, 7 e 8 os resultados são muito bons (considerando os dois locutores, L1 e L4, utilizados para o teste de N). Assim, para evitar um maior número de cálculos optou-se pela utilização de $N = 5$, uma vez que a redução do erro para valores de $N > 5$ não foi significativa, conforme pode ser verificado na Figura 5.11.

5.5 Inicialização de a_{ij}

A distribuição de probabilidade de transição de estado basicamente modela a transição de um estado q_i no instante de tempo t para o estado q_j no instante $t+1$ (a_{ij}), bem como a duração na qual reside um processo em um estado particular (a_{ii}). Na prática, várias são as estimativas utilizadas para a_{ij} [11, 12]. A maioria das estimativas são obtidas através do método de "tentativas", verificando-se qual delas produz o melhor resultado. A partir de um conjunto de estimativas apresentadas em [12, 13, 26, 27]

e fazendo-se um conjunto de testes, utilizou-se neste trabalho a seguinte matriz de transição de estados:

$$\mathrm{A} = \begin{vmatrix} 0.8 & 0.1 & 0.1 & 0.0 & 0.0 \\ 0.0 & 0.2 & 0.8 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.1 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.8 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{vmatrix}$$

Verifica-se através da matriz acima, que as restrições impostas pelo modelo "esquerda-direita" utilizado foram obedecidas. Ou seja, $a_{ij} = 0$ para $j < i$ e $j > i + 2$.

5.6 Inicialização de $\mathrm{b}_j(\mathrm{k})$

Estimativas iniciais de $b_j(k)$ têm uma forte influência nas estimativas finais. Vários métodos tais como segmentação de "K-means" com agrupamento, etc., são utilizados para obter as melhores estimativas iniciais em voz [11, 13]. Todos estes métodos envolvem bastante pré-processamento. Para simplificar, neste trabalho todos os símbolos nos estados são assumidos como sendo "igualmente prováveis" e $b_j(k)$ é inicializado com $1/M$ para todo j, k [13, 26].

$$\mathrm{B} = \begin{vmatrix} 1/M & 1/M & 1/M & \ldots & 1/M \\ 1/M & 1/M & 1/M & \ldots & 1/M \\ 1/M & 1/M & 1/M & \ldots & 1/M \\ 1/M & 1/M & 1/M & \ldots & 1/M \\ 1/M & 1/M & 1/M & \ldots & 1/M \end{vmatrix}$$

Os parâmetros do sistema são portanto,

1. Taxa de amostragem = 8KHz.
2. Comprimento do bloco de voz = 240 amostras (30ms).
3. Comprimento do bloco de superposição = 120 amostras (15ms).
4. Dimensão do vetor de características = 12 (coeficientes LPC).
5. Comprimento da sentença de voz = 12.000 amostras (1.5 segundos).
6. Comprimento da sentença de voz após o janelamento e superposição = 23.760 amostras (2.96 segundos).
7. Comprimento da elocução de treinamento do HMM = 23.760 amostras.
8. Comprimento da elocução para teste de verificação do locutor = 23.760 amostras.
9. Número de estados do HMM = 5 (N).
10. Número de símbolos (níveis do quantizador) = 256 (M).
11. Valor de ϵ (limiar de ajuste) = $1X10^{-6}$

Dentro dos propósitos de investigação do trabalho, foi utilizado um tamanho fixo de amostras para cada sentença, sem considerar a variação do tempo que o locutor levou para pronunciá-la. Acarretando num "corte" da sentença se, p.ex., o locutor demorou um certo tempo para iniciá-la ou se a pronunciou de um forma bastante lenta, ultrapassando assim a quantidade de amostras consideradas. De uma forma geral, essa limitação não afetou em muito o desempenho do sistema, pois os locutores pronunciaram suas sentenças, em média, mantendo um mesmo padrão de repetição.

5.7 Considerações de implementação

A implementação dos algoritmos de HMM, vistos anteriormente, requerem alguns cuidados especiais de forma a obter-se resultados mais precisos. Dois problemas são bastante comuns. Primeiro, os métodos requerem a avaliação de $\alpha_t(i)$ e $\beta_t(i)$ para $1 \leq t \leq T$ e $1 \leq i \leq N$. A partir das equações (4.10) e (4.16) (capítulo 4), é fácil verificar que $T \rightarrow \infty$ (ou T muito grande, i.é., 100 ou mais), cada termo $\alpha_t(i)$ e $\beta_t(i)$ tende rápida e exponencialmente para zero. Na prática, o número de observações necessárias para treinar adequadamente o modelo e/ou computar esta probabilidade irá resultar em "underflow" se as equações (4.10) e (4.16) são avaliadas diretamente. Felizmente, há um método para escalonamento da computação desses valores que não somente soluciona o problema de "underflow", mas também simplifica bastante alguns outros cálculos [11].

O segundo problema é mais sério, e mostra que sérias dificuldades de verificação irão ocorrer se algum elemento de $b_j(k)$ assume um valor zero durante a fase de treinamento [11, 13]. Isto ocorre porque, a fase de reconhecimento envolve a computação de $P_l(\mathbf{O}^l/\lambda_1)$ de $\alpha_t(i)$. Se acontecer o caso em que $\alpha_{t-1}(i)a_{ij}$ é diferente de zero para algum valor de j, e $\mathbf{O}_t = v_k$, então a probabilidade da seqüência associada ao modelo com $b_j(k) = 0$ é $P_l = 0$; assim um erro de reconhecimento deverá ocorrer. Este problema é contornado assumindo que o valor de um $b_j(k)$ nunca poderá ser menor que um dado ϵ. Para tanto, os valores de $b_j(k)$ são reescalonados de forma que $\sum_{i=1}^{N} b_j(k) = 1$. Desta forma, todos os $b_j(k)s$ são comparados com o limiar ϵ e aquele que cai abaixo de ϵ é substituído por ϵ para cada j. Após esta substituição, cada $b_j(k)$ que não foi modificado pelo valor ϵ é reescalonado pela quantidade $1 - R_{bj}\epsilon$ (onde R_{bj} é o número de $b_j(k)s$ modificados para um dado j) normalizando, assim, os $b_j(k)s$. Valores de ϵ entre 10^{-3} e 10^{-7}, são usados para reconhecimento de voz e fornecem baixas taxas de erro [11, 13]. Neste trabalho foi usado $\epsilon = 10^{-6}$.

Para evitar problemas de indeterminação na resolução das equações do modelo, os valores de a_{ij} e π_i iguais a zero assumem também o valor ϵ. Durante a reestimação, para os coeficientes a_{ij} é realizado o mesmo procedimento de reescalonamento feito para os coeficientes $b_j(k)$.

A Figura 5.12 mostra as taxas de erro para os dois locutores usados nos testes dos valores de ϵ (correspondendo aos locutores 1 e 4, respectivamente). Verifica-se que, comprovando a teoria de [11, 13], os resultados obtidos para $\epsilon = 10^{-6}$ são os melhores, porém quando ϵ diminui a taxa de erro volta a crescer. Demostrando assim, que existe um valor limite de ϵ que fornece os melhores resultados.

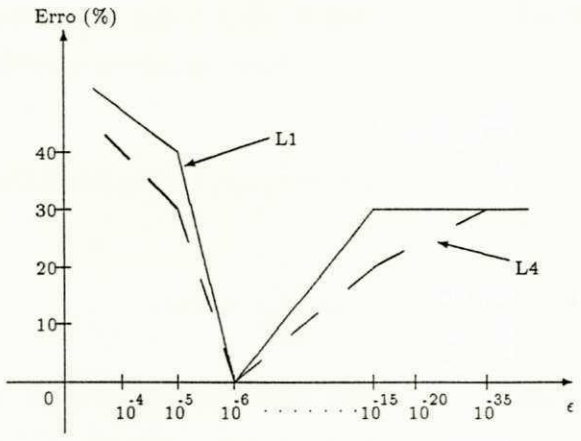


Figura 5.12: Taxa de erro do locutor x ϵ (para $N = 5$, $M = 256$ e dimensão=12)

5.7.1 Escalonamento

Para entender porque o escalonamento é requerido para a técnica de reestimação de HMMs, consideremos a definição de $\alpha_t(i)$ (eq.(4.10), capítulo 4). Pode ser visto que $\alpha_t(i)$ consiste de uma soma de um grande número de termos, cada um da forma

$$(\prod_{s=1}^{t-1} a_{q_s q_{s-1}} \prod_{s=1}^{t} b_{q_s}(O_s)) \tag{5.1}$$

Com $q_t = S_1$. Sendo cada **a** e **b** muito menor que 1 (geralmente significativamente menor que 1), pode-se ver que a medida que **t** torna-se grande (i.é., 10 ou mais), cada

termo de $\alpha_t(i)$ tenderá exponencialmente para zero. Para **t** suficientemente grande (i.é., 100 ou mais) a taxa de variação de $\alpha_t(i)$ irá exceder a precisão da máquina (quando é utilizada dupla precisão). Assim, uma forma de solucionar este problema é incorporar a técnica de escalonamento. O princípio no qual está baseado o escalonamento usado neste trabalho, consiste em multiplicar $\alpha_t(i)$ por algum coeficiente de escalonamento independente de **i** (i.é, depende somente de **t**) tal que este permaneça dentro da faixa dinâmica do computador para $1 \leq t \leq T$. A proposta consiste em desempenhar uma operação similar em $\beta_t(i)$ e então, no final da computação, remover o efeito total do escalonamento.

O coeficiente de escalonamento, esc_t, possui a forma

$$esc_t = (\sum_{i=1}^{N} \alpha_t(i))^{-1} \tag{5.2}$$

Os termos $\alpha_t(i)$ e $\beta_t(i)$ escalonados são, em seguida, utilizados nas fórmulas de reestimação de a_{ij} e $b_j(k)$, descritas no capítulo 4.

A única variação real no procedimento do HMM em virtude do escalonamento, reside na computação de $P_l(\mathbf{O}^l/\lambda_l)$. Não podemos meramente somar os termos $\alpha_T(i)$ desde que estes tenham sido escalonados anteriormente. Entretanto, nós podemos usar a propriedade

$$\prod_{t=1}^{T} esc_t \sum_{i=1}^{N} \alpha_T(i) = C_T \sum_{i=1}^{N} \alpha_T(i) = 1 \tag{5.3}$$

Assim tem-se

$$\prod_{t=1}^{T} esc_t P_l(\mathbf{O}^l/\lambda_l) = 1 \tag{5.4}$$

ou

$$P_l(\mathbf{O}^l/\lambda_l) = \frac{1}{\prod_{t=1}^{T} esc_t} \tag{5.5}$$

ou

$$log[P_l(\mathbf{O}^1/\lambda_1)] = -\sum_{t=1}^{T} log(esc_t) \tag{5.6}$$

O log de $P_l(\mathbf{O}^1/\lambda_1)$ pode ser computado, mas $P_l(\mathbf{O}^1/\lambda_1)$ não, visto que o cálculo de $P_l(\mathbf{O}^1/\lambda_1)$ poderia levar a um resultado fora da faixa dinâmica da máquina [11].

5.8 Avaliação dos Resultados Experimentais

Os resultados obtidos são mostrados na tabela 5.1. Para melhorar o desempenho do sistema foi empregado um procedimento de decisão sequencial, no qual as decisões para cada linha limite são adiadas até o futuro teste de entrada. Ao invés do uso de um único limiar para aceitar ou rejeitar um determinado locutor, dois limiares dividem o limite em três decisões: aceita se a probabilidade é maior ou igual ao maior limiar; rejeita se é menor que o limiar mais baixo; e solicita um nova entrada se a menor distância está entre os dois limiares. Tal método evita erros em casos mais próximos [7].

Na tabela 5.1 são mostrados os resultados do sistema onde as intersecções de uma das linhas com colunas indicam as ocorrências do teste de verificação de cada locutor L1, L2, L3, L4, L5 e L6 com ele mesmo e entre locutores distintos, por exemplo, locutor L1 pronunciando a sentença dos demais locutores, ou seja, tentando se passar por L2, L3, L4, L5 e L6. Os testes de verificação de um locutor com ele mesmo foram repetidos 10 vezes, ao passo que cada locutor com os demais foi repetido 5 vezes.

A tabela 5.1 mostra que o sistema apresenta resultados bastante satisfatórios para alguns locutores e alguns erros mais elevados para outros locutores. Verifica-se, através da tabela, que os locutores 1 e 4 apresentam 0% de falsa rejeição (a=10), ou seja, dentre as dez repetições de suas sentenças correspondentes, todas foram aceitas. Para os locutores 2 e 5, dentre as dez repetições de suas sentenças 9 foram aceitas (a=9) e para uma foi solicitada uma outra repetição, pois seu valor de probabilidade estava entre os dois limiares (r=1). Para o locutor 3, dentre as dez repetições de sua sentença 8 foram

	L1	L2	L3	L4	L5	L6
L1	**a=10**	na=4,a=1	na=4,r=1	na=5	na=5	na=5
L2	na=4,r=1	**a=9,r=1**	na=5	na=4,r=1	na=5	na=5
L3	na=4,r=1	na=4,a=1	**a=8,r=1,na=1**	na=5	na=5	na=5
L4	na=4,r=1	na=5	na=5	**a=10**	na=5	na=5
L5	na=5	na=5	na=5	na=4,a=1	**a=9,r=1**	na=5
L6	na=5	na=5	na=5	na=4,r=1	na=5	**a=8,r=2**

Tabela 5.1: Resultados obtidos para o sistema proposto, onde **a** = aceita o locutor (verdadeiro ou falso); **r** = repete a sentença e **na** = não aceita o locutor (verdadeiro ou falso).

aceitas (a=8), uma solicitou outra repetição (r=1) e outra não aceitou o locutor, ou seja, apresentou uma falsa rejeição (na=1). Para o locutor 6, dentre as dez repetições de sua sentença 8 foram aceitas (a=8) e duas solicitaram outra repetição (r=2).

Os índices de falsa aceitação para todos os locutores foram, de uma forma geral, baixos. Dentre as 5 repetições da sentença do locutor 2 feitas pelo locutor 1, uma foi aceita incorretamente (falsa aceitação, a=1) e 4 não foram aceitas (na=4). Para as 5 repetições da sentença do locutor 3, uma solicitou repetição (r=1) e 4 não foram aceitas (na=4). Para as 5 repetições das sentenças dos locutores 4, 5 e 6, respectivamente, nenhuma delas foi aceita (na=5). Para os demais locutores os resultados foram semelhantes, não sendo necessário comentá-los. As tabelas 5.2, 5.3, 5.4 e 5.5 mostram os índices de falsa rejeição, falsa aceitação e repetição, respectivamente, para os locutores utilizados para o teste do sistema.

Locutor	L1	L2	L3	L4	L5	L6
falsa rejeição	0%	0%	10%	0%	0%	0%

Tabela 5.2: Índices de falsa rejeição para cada locutor

Locutor	L1	L2	L3	L4	L5	L6
L1	-	20%	0%	0%	0%	0%
L2	0%	-	0%	0%	0%	0%
L3	0%	20%	-	0%	0%	0%
L4	0%	0%	0%	-	0%	0%
L5	0%	0%	0%	20%	-	0%
L6	0%	0%	0%	0%	0%	-

Tabela 5.3: Índices de falsa aceitação para cada locutor

Locutor	L1	L2	L3	L4	L5	L6
repetição	0%	10%	10%	0%	10%	20%

Tabela 5.4: Índices de repetição por locutor da sua própria sentença

Locutor	L1	L2	L3	L4	L5	L6
L1	-	0%	20%	0%	0%	0%
L2	20%	-	0%	20%	0%	0%
L3	20%	0%	-	0%	0%	0%
L4	20%	0%	0%	-	0%	0%
L5	0%	0%	0%	0%	-	0%
L6	0%	0%	0%	20%	0%	-

Tabela 5.5: Índices de repetição por locutor da sentença de outro locutor

Alguns dos erros ocorrem devido às limitações inerentes do modelo (citadas posteriormente) e devido a uma série de problemas tais como: ruído originário do ambiente de gravação (conversação, portas abrindo e fechando, ar condicionado, telefone, etc.). A energia dessas diferentes fontes está concentrada em certas faixas de freqüência e não podem ser tratadas como ruído branco, para que assim seus efeitos possam ser reduzidos mais facilmente. Seria necessário a utilização de filtros adptativos, que não foram usados neste trabalho para evitar uma maior complexidade dos algoritmos, bem como um maior tempo de computação. Outro fator que afeta o desempenho do sistema são as características inerentes de cada locutor. Por exemplo, a mesma sentença pronunciada por um único locutor apresenta, algumas vezes, características bem diferentes. Para demonstrar esse fato, as Figuras 5.13 e 5.14 mostram as formas de onda de duas elocuções da mesma sentença pronunciadas pelo locutor 5 e verifica-se que as mesmas apresentam características bem diferentes.

Para testar os índices de falsa aceitação só foram utilizadas 5 repetições de cada sentença por locutor, em virtude das limitações do total de memória disponível. Resultados mais significativos poderiam ter sido apresentados com a utilização de uma maior quantidade de repetições para cada locutor.

Figura 5.13: Terceira elocução da sentença do locutor 5

Figura 5.14: Oitava elocução da sentença do locutor 5

O maior problema com modelos "esquerda-direita" é que uma simples seqüência de observação, às vezes, não pode ser usada para treinar o modelo usando o princípio tratado neste trabalho. Isto ocorre porque a natureza transiente dos estados do modelo só permite um número pequeno de observações para um estado. Assim, estimativas realizáveis são possíveis somente com múltiplas seqüências de observação. Para tanto, assume-se que cada seqüência de observação é independente da outra e que as freqüências individuais de ocorrência para cada uma das seqüências são somadas [35, 36]. Portanto, pode-se melhorar o desempenho do modelo utilizando-se múltiplas seqüências de observação ($\mathbf{O}_k^l$) para cada l-ésimo locutor com $1 \leq k \leq Kl$, sendo necessário um número bastante elevado de repetições para cada locutor (Kl em torno de 100). Tal procedimento poderia melhorar os resultados, mas acarretaria num maior número de dados e uma maior complexidade dos algoritmos utilizados.

Embora o uso de HMM tenha contribuído bastante para os avanços na área de Verificação de Locutor, há algumas limitações inerentes para este tipo de modelo. A maior limitação é assumir que observações sucessivas (blocos de voz) são independentes, e portanto a probabilidade de uma seqüência de observações $P(O_1, O_2, ..., O_T)$ pode

ser escrita como um produto de probabilidades de observações individuais [11].

Outra limitação é assumir que a probabilidade de está em um dado estado no instante de tempo t, depende somente do estado no instante de tempo anterior, $t-1$, o que é, em alguns casos, inapropriada para sons de voz onde dependências muitas vezes se estendem através de muitos estados [11].

Apesar dessas limitações, as pesquisas têm mostrado que HMM fornece ótimos resultados para vários sistemas de Reconhecimento de Locutor, apesar de sua utilização ainda ser restrita nessa área sendo mais utilizado em reconhecimento de fala.

Capítulo 6

Conclusões

Até um certo tempo atrás pouco se ouvia falar em Modelos de Markov para reconhecimento de locutor, muito embora estes tenham sido estudados desde os anos 60. Com o passar do tempo, percebeu-se que as técnicas de reconhecimento mais usuais apesar de apresentarem bons resultados, tinham alguns problemas. Como por exemplo, o alto custo computacional do método utilizando programação dinâmica. Além disso, apresentavam sérias dificuldades quando da execução de tarefas mais complicadas (p.ex. reconhecimento de locutor independente do texto).

Além de conseguir solucionar os problemas acima citados, os HMMs conseguem modelar inerentemente as características temporais do sinal de entrada bem como, não é necessário uma distribuição a priori das entradas para estimação dos parâmetros, o que não é o caso, usualmente, em outras técnicas estatísticas.

A partir desse conjunto de vantagens e por se tratar de uma técnica ainda não muito explorada em verificação de locutor dependente do texto, optou-se pela utilização de HMMs tendo consciência, também, dos problemas inerentes desta técnica de reconhecimento de padrões.

Os resultados apresentados no capítulo anterior, demonstram que o desempenho da técnica utilizada foi muito bom, mesmo diante das adversidades também já citadas no capítulo 5. Poder-se-ia tentar melhorar alguns resultados utilizando-se, p.ex., uma maior taxa de bits/amostra (desde que a taxa de bits/amostra seja < 1), usando

um maior número de níveis do quantizador vetorial, p. ex., M=512, 1024 ou 2048. Entretanto, isto levaria a um aumento considerável do volume de dados, acarretando em um custo computacional mais elevado. Outra opção seria utilizar coeficientes cepstrais [9] em substituição aos coeficientes LPC utilizados para representação de cada vetor de observação. Outra tentativa poderia ser a utilização de um maior número de estados, o que muitas vezes (como mostrado na Figura 5.11) não iria modificar muito os resultados obtidos.

Partindo-se do que foi apresentado durante esta dissertação, verifica-se que os HMMs necessitam de uma larga quantidade de parâmetros e que muitas vezes se torna dispendioso verificar o seu comportamento variando-se apenas um de seus parâmetros e mantendo-se os demais fixos. Portanto, foi necessário considerar como verdade algumas suposições já definidas em outros trabalhos, como por exemplo o "chute" inicial para as matrizes $[a_{ij}]$ (matriz transição de estados) e $[b_j(k)]$ (matriz de probabilidades), bem como o uso do algoritmo LBG para quantização vetorial e o uso do algoritmo de reestimação de "Baum-Welch" para obter os parâmetros do modelo.

Sabendo-se que os modelos "esquerda-direita" funcionam, na maioria das situações, muito bem para os sistemas de reconhecimento de voz e locutor, os mesmos foram utilizados neste trabalho apesar de suas limitações (p. ex., a necessidade, em alguns casos, da utilização de múltiplas seqüências de observação). Fato este que não foi tratado neste trabalho, devido às limitações de memória (pois são necessárias, aproximadamente, 100 repetições de cada sentença para cada locutor). Além disso, seria necessário abusar da boa vontade dos locutores utilizados, que nem sempre estariam dispostos a repetir sua sentença em torno de 100 vezes.

Algumas limitações do próprio HMM, que portanto não podem ser modificadas e sim melhor tratadas, afetam o seu desempenho. Por exemplo, assumir que observações sucessivas (blocos de voz) são independentes, como também assumir que a probabilidade de está em um dado estado no instante de tempo t, depende somente do estado no instante de tempo anterior, o que muitas vezes se torna inapropriada para sons de voz onde dependências estatísticas se estendem, na maioria das vezes, ao longo de muitos estados [11].

Mesmo diante das restrições citadas, os resultados apresentados nesta dissertação quando comparados com outros trabalhos já publicados [28, 29] demonstram que os Modelos de Markov Escondidos podem fornecer resultados muito bons para verificação de locutor dependente do texto.

A verificação de locutor não se justifica apenas por possibilitar uma interação mais confortável entre o homem e a máquina, mas sobretudo pela segurança que pode proporcionar. Pode-se observar que os resultados apresentados neste trabalho satisfizeram essas exigências, obtendo-se índices de falsa aceitação pequenos para todos os locutores, ou seja, os impostores não conseguiram se passar pelos locutores verdadeiros e índices de falsa rejeição também pequenos, significando que os locutores verdadeiros foram aceitos, na maioria das vezes, pelo sistema.

A principal contribuição deste trabalho foi demonstrar que os HMMs podem ser usados, obtendo-se bons resultados, não só em reconhecimento de fala onde o seu uso é mais comum, como também em reconhecimento de locutor, mais especificamente, verificação de locutor, onde as pesquisas ainda são restritas.

Como encerramento desta dissertação, serão sugeridos diversos trabalhos de aprimoramento e aplicação dos resultados obtidos:

a) Prosseguimento deste trabalho, a nível de Doutorado, pretendendo-se, inicialmente, realizar uma avaliação mais minunciosa dos parâmetros do Modelo de Markov Escondido p.ex., número de estados (N), número de símbolos (M), matriz de transição de estados $[a_{ij}]$, vetor de probabilidade inicial (π_i), matriz de probabilidades $[b_j(k)]$.

b) Estudo e avaliação de outros parâmetros espectrais que caracterizem o locutor.

c) Verificar a melhoria do seu desempenho quando da utilização de múltiplas seqüências de observação (descritas no capítulo anterior).

d) Combinação de parâmetros espectrais e temporais para o modelamento da identidade vocal.

Seguindo-se ainda o uso, junto ao HMM, de outras técnicas de quantização vetorial para codificação dos vetores de observação em substituição à quantização vetorial, p.ex., utilizando-se Redes Neuronais [37] e a comparação dos resultados com outros que utilizam técnicas mais usuais para reconhecimento de locutor (p.ex. Alinhamento Dinâmico no Tempo (DTW), Coeficientes de Predição Linear (LPC), Quantização Vetorial (QV)).

Por fim, sugere-se a implementação e otimização do sistema proposto para aplicações em tempo real, por exemplo, para o controle de acesso, através da senha verbal, a um ambiente restrito.

Apêndice A

Algoritmos utilizados

Para todos os algoritmos, foram utilizados os seguintes parâmetros:

1. Taxa de amostragem = 8KHz.
2. Comprimento do bloco de voz = 240 amostras (30ms).
3. Comprimento do bloco de superposição = 120 amostras (15ms).
4. Dimensão do vetor de características = 12 (coeficientes LPC).
5. Comprimento da sentença de voz = 12.000 amostras (1.5 segundos).
6. Comprimento da sentença de voz após o janelamento e superposição = 23.760 amostras (2.96 segundos).
7. Comprimento da elocução de treinamento do HMM = 23.760 amostras.
8. Comprimento da elocução para teste de verificação do locutor = 23.760 amostras.
9. Número de estados do HMM = 5 (N).
10. Número de símbolos (níveis do quantizador) = 256 (M).
11. Valor de ϵ (limiar de ajuste) = $1X10^{-6}$

A.1 Algoritmo de Levinson-Durbin

$R_r(i) = \sum_{k=1}^{p} c_k R_r(|i-k|) \quad 1 \leq i \leq p$

$E^{(0)} = R_r(0)$

$k_i = \frac{\{R_r(i) - \sum_{j=1}^{i-1} c_j^{(i-1)} R_r(i-j)\}}{E^{(i-1)}}$

$c_i^{(i)} = k_i$

$c_j^{(i)} = c_j^{(i-1)} - k_i c_{i-j}^{(i-1)} \quad 1 \leq j \leq i-1$

$E^{(i)} = (1 - k_i^2) E^{(i-1)} \quad 1 \leq i \leq p$

As equações são resolvidas recursivamente para $i = 1, 2, ..., p$ e a solução final é dada por:

$c_j = c_j^p \quad 1 \leq j \leq p$

Onde:

1. $E^{(i)}$ - energia do sinal de erro na i-ésima iteração;
2. $R_r(i)$ - função de autocorrelação a curto prazo com atraso i;
3. k_i - i-ésimo coeficiente de reflexão;
4. $c_j^{(i)}$ - valor do j-ésimo termo a ser determinado, na i-ésima iteração.

A.1.1 Listagem dos resultados obtidos pelo algoritmo

Resultados obtidos para os coeficientes LPC, utilizando o algoritmo de Levinson-Durbin com $p = 12$, para o primeiro vetor de observação (O_1), correspondente a primeira elocução das sentenças de cada um dos 6 locutores utilizados para teste do desempenho do HMM.

Para $1 \leq i \leq p$,

Locutor1

c_i = {-0.073272, -0.090811, 0.089345, -0.167509, -0.078622, -0.209981, -0.172869, -0.164146, -0.016351, -0.113089, -0.202120, 0.038541}

Locutor2

c_i = {0.265195, -0.420437, 0.764533, 0.153486, -0.201648, -0.185682, -0.469318, 0.111605, -0.029406, -0.094617, -0.154798, -0.042327}

Locutor3

c_i = {-0.355514, -0.159712, 1.367586, 0.575595, -0.083949, -1.123632, -0.627773, 0.023298, 0.428609, 0.134060, 0.002949, -0.170609}

Locutor4

c_i = {-0.543503, 0.230617, 0.259027, -0.164399, 0.276080, 0.102992, -0.023984, -0.263758, -0.343110, -0.288388, 0.051286, 0.032559}

Locutor5

c_i = {0.933509, 0.138016, 0.240065, -0.509754, -0.183555, 0.037540, 0.105740, 0.119810, 0.046307, -0.033533, -0.120451, -0.038033}

Locutor6

c_i = {-0.071916, -0.353073, 0.079637, -0.133923, -0.017692, -0.098317, -0.134598, -0.271867, -0.022395, -0.149009, -0.068456, -0.182145}

A.2 Algoritmo LBG - Quantização Vetorial

A.2.1 Geração do dicionário inicial

O dicionário inicial é gerado a partir da escolha aleatória de M vetores, onde M corresponde ao número de níveis do quantizador, da seqüência de treino.

A.2.2 Geração do dicionário do quantizador vetorial

O dicionário é gerado da seguinte forma:

1. Inicialmente, cada vetor x da seqüência de treino é comparado com todo o dicionário inicial atráves da medida de distância do erro quadrático obtendo-se um vetor y, que de todos os vetores do dicionário, é o que melhor representa x.

2. Cálculo do centróides de cada partição. Depois de ter lido todo o arquivo de amostras, o dicionário deve ser atualizado pelos centóides das partições.

3. Em seguida, o dicionário de comparação de cada vetor x não será mais o dicionário inicial e sim o dicionário já atualizado.

A.2.3 Resultados obtidos para o quantizador vetorial

As tabelas A.1a, A.1b, A.2 e A.3 apresentam os resultados do quantizador para os níveis 32, 64, 128 e 256, respectivamente, para a dimensão = 12.

Número de iterações = 4.
Número de vetores lidos = 9603.
Distorção total por dimensão = 1.144788e-03.
Ocupação dos vetores para cada classe:

$c[0] = 472$	$c[1] = 1590$	$c[2] = 461$	$c[3] = 1151$	$c[4] = 455$	$c[5] = 616$
$c[6] = 311$	$c[7] = 121$	$c[8] = 206$	$c[9] = 111$	$c[10] = 145$	$c[11] = 78$
$c[12] = 750$	$c[13] = 34$	$c[14] = 172$	$c[15] = 12$	$c[16] = 40$	$c[17] = 43$
$c[18] = 25$	$c[19] = 31$	$c[20] = 55$	$c[21] = 280$	$c[22] = 228$	$c[23] = 100$
$c[24] = 16$	$c[25] = 8$	$c[26] = 46$	$c[27] = 8$	$c[28] = 617$	$c[29] = 563$
$c[30] = 677$	$c[31] = 181$				

(a)

Número de iterações = 4.

Número de vetores lidos = 9603.

Distorção total por dimensão = 7.731805e-04.

Ocupação dos vetores para cada classe:

$c[0] = 289$	$c[1] = 1264$	$c[2] = 334$	$c[3] = 8$	$c[4] = 351$	$c[5] = 381$
$c[6] = 156$	$c[7] = 48$	$c[8] = 60$	$c[9] = 97$	$c[10] = 34$	$c[11] = 52$
$c[12] = 172$	$c[13] = 24$	$c[14] = 91$	$c[15] = 6$	$c[16] = 10$	$c[17] = 23$
$c[18] = 18$	$c[19] = 21$	$c[20] = 39$	$c[21] = 187$	$c[22] = 85$	$c[23] = 24$
$c[24] = 10$	$c[25] = 7$	$c[26] = 31$	$c[27] = 7$	$c[28] = 258$	$c[29] = 355$
$c[30] = 449$	$c[31] = 164$	$c[32] = 147$	$c[33] = 42$	$c[34] = 65$	$c[35] = 7$
$c[36] = 9$	$c[37] = 38$	$c[38] = 19$	$c[39] = 63$	$c[40] = 54$	$c[41] = 80$
$c[42] = 33$	$c[43] = 136$	$c[44] = 112$	$c[45] = 37$	$c[46] = 85$	$c[47] = 79$
$c[48] = 199$	$c[49] = 295$	$c[50] = 415$	$c[51] = 67$	$c[52] = 240$	$c[53] = 183$
$c[54] = 150$	$c[55] = 251$	$c[56] = 40$	$c[57] = 156$	$c[58] = 181$	$c[59] = 15$
$c[60] = 7$	$c[61] = 424$	$c[62] = 14$	$c[63] = 111$		

(b)

Tabela A.1: Resultados obtidos para quantizador vetorial de dimensão = 12 e a) número de níveis = 32, b) número de níveis = 64

Número de iterações = 4.

Número de vetores lidos = 9603.

Distorção total por dimensão = 6.333121e-04.

Ocupação dos vetores para cada classe:

$c[0] = 105$	$c[1] = 164$	$c[2] = 117$	$c[3] = 50$	$c[4] = 135$	$c[5] = 128$
$c[6] = 57$	$c[7] = 37$	$c[8] = 31$	$c[9] = 76$	$c[10] = 27$	$c[11] = 46$
$c[12] = 69$	$c[13] = 23$	$c[14] = 24$	$c[15] = 6$	$c[16] = 10$	$c[17] = 23$
$c[18] = 18$	$c[19] = 20$	$c[20] = 26$	$c[21] = 63$	$c[22] = 54$	$c[23] = 23$
$c[24] = 10$	$c[25] = 7$	$c[26] = 9$	$c[27] = 7$	$c[28] = 111$	$c[29] = 135$
$c[30] = 128$	$c[31] = 61$	$c[32] = 50$	$c[33] = 38$	$c[34] = 46$	$c[35] = 7$
$c[36] = 9$	$c[37] = 33$	$c[38] = 14$	$c[39] = 37$	$c[40] = 16$	$c[41] = 48$
$c[42] = 33$	$c[43] = 23$	$c[44] = 57$	$c[45] = 32$	$c[46] = 51$	$c[47] = 40$
$c[48] = 102$	$c[49] = 109$	$c[50] = 130$	$c[51] = 43$	$c[52] = 69$	$c[53] = 61$
$c[54] = 70$	$c[55] = 74$	$c[56] = 25$	$c[57] = 83$	$c[58] = 87$	$c[59] = 9$
$c[60] = 7$	$c[61] = 76$	$c[62] = 14$	$c[63] = 71$	$c[64] = 40$	$c[65] = 109$
$c[66] = 158$	$c[67] = 73$	$c[68] = 175$	$c[69] = 276$	$c[70] = 50$	$c[71] = 64$
$c[72] = 215$	$c[73] = 104$	$c[74] = 159$	$c[75] = 136$	$c[76] = 117$	$c[77] = 68$
$c[78] = 35$	$c[79] = 171$	$c[80] = 110$	$c[81] = 114$	$c[82] = 57$	$c[83] = 51$
$c[84] = 50$	$c[85] = 103$	$c[86] = 93$	$c[87] = 147$	$c[88] = 131$	$c[89] = 42$
$c[90] = 73$	$c[91] = 157$	$c[92] = 153$	$c[93] = 151$	$c[94] = 169$	$c[95] = 85$
$c[96] = 75$	$c[97] = 30$	$c[98] = 111$	$c[99] = 93$	$c[100] = 85$	$c[101] = 201$
$c[102] = 86$	$c[103] = 158$	$c[104] = 120$	$c[105] = 125$	$c[106] = 54$	$c[107] = 131$
$c[108] = 46$	$c[109] = 51$	$c[110] = 59$	$c[111] = 133$	$c[112] = 47$	$c[113] = 84$
$c[114] = 55$	$c[115] = 131$	$c[116] = 10$	$c[117] = 55$	$c[118] = 45$	$c[119] = 27$
$c[120] = 40$	$c[121] = 55$	$c[122] = 27$	$c[123] = 32$	$c[124] = 37$	$c[125] = 164$
$c[126] = 96$	$c[127] = 140$				

Tabela A.2: Resultados obtidos para quantizador vetorial de dimensão = 12 e número de níveis = 128

Número de iterações = 4,
número de vetores lidos = 9603,
distorção total por dimensão = 5.538295e-04 e
ocupação dos vetores para cada classe:

$c[0] = 52$	$c[1] = 52$	$c[2] = 41$	$c[3] = 26$	$c[4] = 53$	$c[5] = 61$
$c[6] = 48$	$c[7] = 33$	$c[8] = 22$	$c[9] = 61$	$c[10] = 20$	$c[11] = 29$
$c[12] = 14$	$c[13] = 20$	$c[14] = 14$	$c[15] = 6$	$c[16] = 8$	$c[17] = 20$
$c[18] = 18$	$c[19] = 18$	$c[20] = 19$	$c[21] = 40$	$c[22] = 42$	$c[23] = 20$
$c[24] = 10$	$c[25] = 3$	$c[26] = 8$	$c[27] = 7$	$c[28] = 59$	$c[29] = 46$
$c[30] = 84$	$c[31] = 26$	$c[32] = 13$	$c[33] = 24$	$c[34] = 21$	$c[35] = 7$
$c[36] = 9$	$c[37] = 21$	$c[38] = 14$	$c[39] = 32$	$c[40] = 14$	$c[41] = 30$
$c[42] = 26$	$c[43] = 16$	$c[44] = 32$	$c[45] = 24$	$c[46] = 42$	$c[47] = 23$
$c[48] = 75$	$c[49] = 33$	$c[50] = 14$	$c[51] = 23$	$c[52] = 17$	$c[53] = 24$
$c[54] = 27$	$c[55] = 8$	$c[56] = 23$	$c[57] = 47$	$c[58] = 61$	$c[59] = 4$
$c[60] = 6$	$c[61] = 57$	$c[62] = 14$	$c[63] = 28$	$c[64] = 26$	$c[65] = 36$
$c[66] = 7$	$c[67] = 29$	$c[68] = 47$	$c[69] = 72$	$c[70] = 37$	$c[71] = 42$
$c[72] = 133$	$c[73] = 69$	$c[74] = 101$	$c[75] = 88$	$c[76] = 52$	$c[77] = 45$
$c[78] = 15$	$c[79] = 87$	$c[80] = 41$	$c[81] = 67$	$c[82] = 27$	$c[83] = 44$
$c[84] = 49$	$c[85] = 75$	$c[86] = 47$	$c[87] = 98$	$c[88] = 73$	$c[89] = 34$
$c[90] = 48$	$c[91] = 61$	$c[92] = 66$	$c[93] = 55$	$c[94] = 33$	$c[95] = 47$
$c[96] = 56$	$c[97] = 25$	$c[98] = 54$	$c[99] = 29$	$c[100] = 42$	$c[101] = 30$
$c[102] = 34$	$c[103] = 55$	$c[104] = 46$	$c[105] = 61$	$c[106] = 16$	$c[107] = 46$
$c[108] = 14$	$c[109] = 41$	$c[110] = 48$	$c[111] = 59$	$c[112] = 37$	$c[113] = 34$
$c[114] = 21$	$c[115] = 63$	$c[116] = 8$	$c[117] = 25$	$c[118] = 31$	$c[119] = 13$
$c[120] = 28$	$c[121] = 16$	$c[122] = 10$	$c[123] = 20$	$c[124] = 15$	$c[125] = 68$
$c[126] = 33$	$c[127] = 53$	$c[128] = 44$	$c[129] = 23$	$c[130] = 34$	$c[131] = 37$
$c[132] = 62$	$c[133] = 71$	$c[134] = 51$	$c[135] = 43$	$c[136] = 30$	$c[137] = 40$
$c[138] = 56$	$c[139] = 12$	$c[140] = 90$	$c[141] = 68$	$c[142] = 124$	$c[143] = 71$
$c[144] = 57$	$c[145] = 99$	$c[146] = 27$	$c[147] = 54$	$c[148] = 40$	$c[149] = 26$
$c[150] = 42$	$c[151] = 114$	$c[152] = 69$	$c[153] = 49$	$c[154] = 55$	$c[155] = 27$

$c[156] = 22$	$c[157] = 61$	$c[158] = 64$	$c[159] = 57$	$c[160] = 45$	$c[161] = 72$
$c[162] = 114$	$c[163] = 69$	$c[164] = 7$	$c[165] = 44$	$c[166] = 11$	$c[167] = 6$
$c[168] = 9$	$c[169] = 11$	$c[170] = 14$	$c[171] = 27$	$c[172] = 29$	$c[173] = 20$
$c[174] = 20$	$c[175] = 46$	$c[176] = 14$	$c[177] = 13$	$c[178] = 25$	$c[179] = 21$
$c[180] = 25$	$c[181] = 50$	$c[182] = 83$	$c[183] = 4$	$c[184] = 104$	$c[185] = 44$
$c[186] = 80$	$c[187] = 41$	$c[188] = 53$	$c[189] = 10$	$c[190] = 17$	$c[191] = 35$
$c[192] = 58$	$c[193] = 3$	$c[194] = 7$	$c[195] = 7$	$c[196] = 8$	$c[197] = 30$
$c[198] = 68$	$c[199] = 28$	$c[200] = 49$	$c[201] = 32$	$c[202] = 30$	$c[203] = 55$
$c[204] = 50$	$c[205] = 36$	$c[206] = 39$	$c[207] = 12$	$c[208] = 44$	$c[209] = 87$
$c[210] = 71$	$c[211] = 9$	$c[212] = 8$	$c[213] = 70$	$c[214] = 97$	$c[215] = 50$
$c[216] = 5$	$c[217] = 40$	$c[218] = 5$	$c[219] = 5$	$c[220] = 4$	$c[221] = 6$
$c[222] = 17$	$c[223] = 15$	$c[224] = 14$	$c[225] = 34$	$c[226] = 13$	$c[227] = 19$
$c[228] = 22$	$c[229] = 39$	$c[230] = 14$	$c[231] = 45$	$c[232] = 12$	$c[233] = 25$
$c[234] = 50$	$c[235] = 20$	$c[236] = 52$	$c[237] = 52$	$c[238] = 20$	$c[239] = 71$
$c[240] = 33$	$c[241] = 32$	$c[242] = 39$	$c[243] = 8$	$c[244] = 111$	$c[245] = 12$
$c[246] = 34$	$c[247] = 39$	$c[248] = 56$	$c[249] = 11$	$c[250] = 19$	$c[251] = 23$
$c[252] = 28$	$c[253] = 20$	$c[254] = 25$	$c[255] = 13$		

Tabela A.3: Resultados obtidos para quantizador vetorial de dimensão = 12 e número de níveis = 256

A.3 Algoritmo para cálculo do HMM referente a cada locutor

Para cada locutor, foi calculado um Modelo de Markov Escondido (HMM) utilizando o algoritmo de reestimação de "Baum-Welch". Os fluxogramas referentes as fases de treinamento (Figura A.1) e verificação (Figura A.2) são mostrados em seguida.

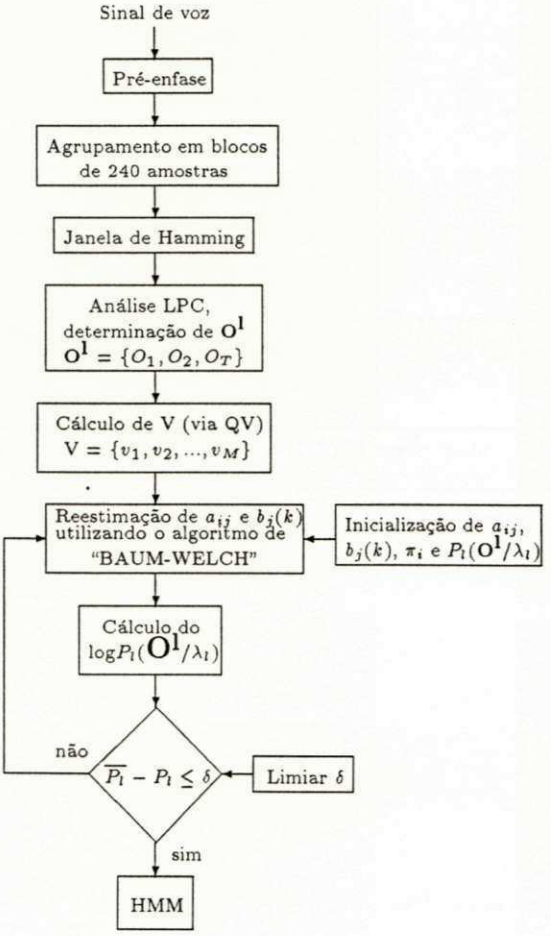


Figura A.1: Fase de treinamento

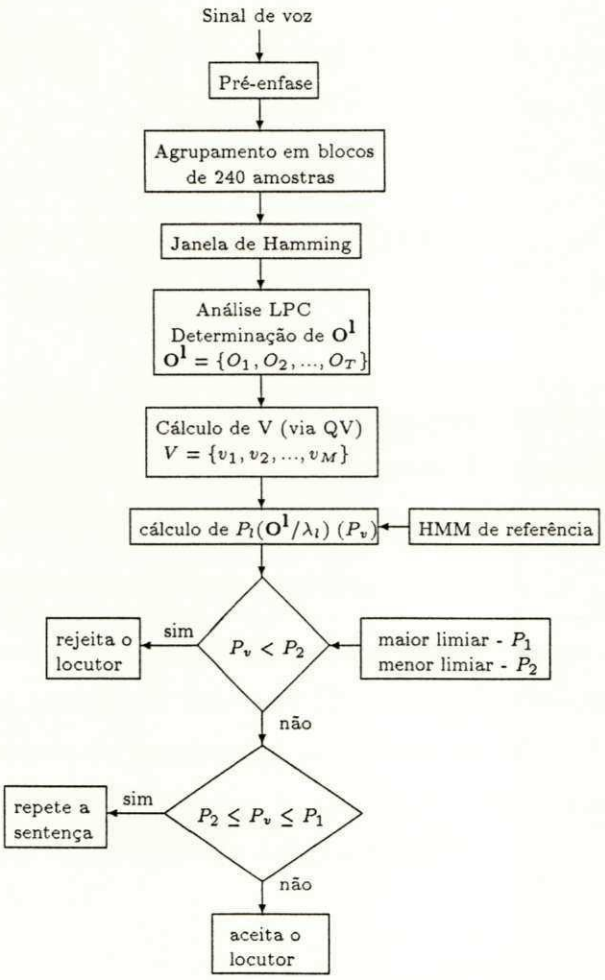


Figura A.2: Fase de verificação

Apêndice B

Ambiente de trabalho

A aquisição dos sinais de voz e suas reproduções foram realizadas em uma estação de trabalho SUN SPARCstation da SUN Microsystem Incorporation, do Laboratório de Automação e Processamento de Sinais (LAPS), no DEE/UFPb.

Todas as simulações foram realizadas em tempo não real.

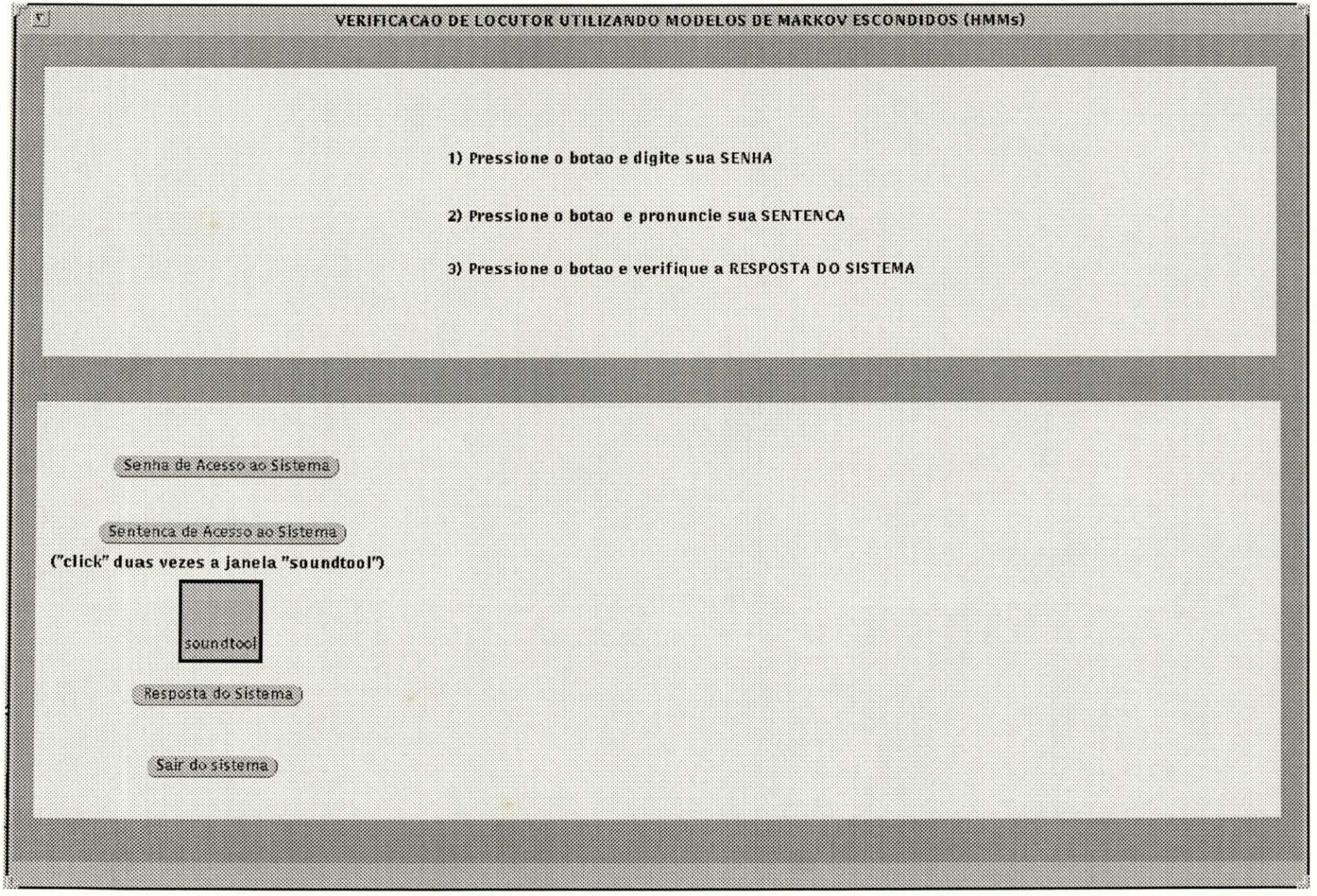
Todos os programas utilizados neste trabalho (algoritmo para realização da pré-ênfase, algoritmo para geração da Janela de Hamming, algoritmo de Levinson Durbin, algoritmo LBG, algoritmo de reestimação de "Baum-Welch"), determinando as fases de treinamento e verificação do sistema foram escritos em linguagem C, compatível com o ambiente SunOs "Unix Like" da estação SUN.

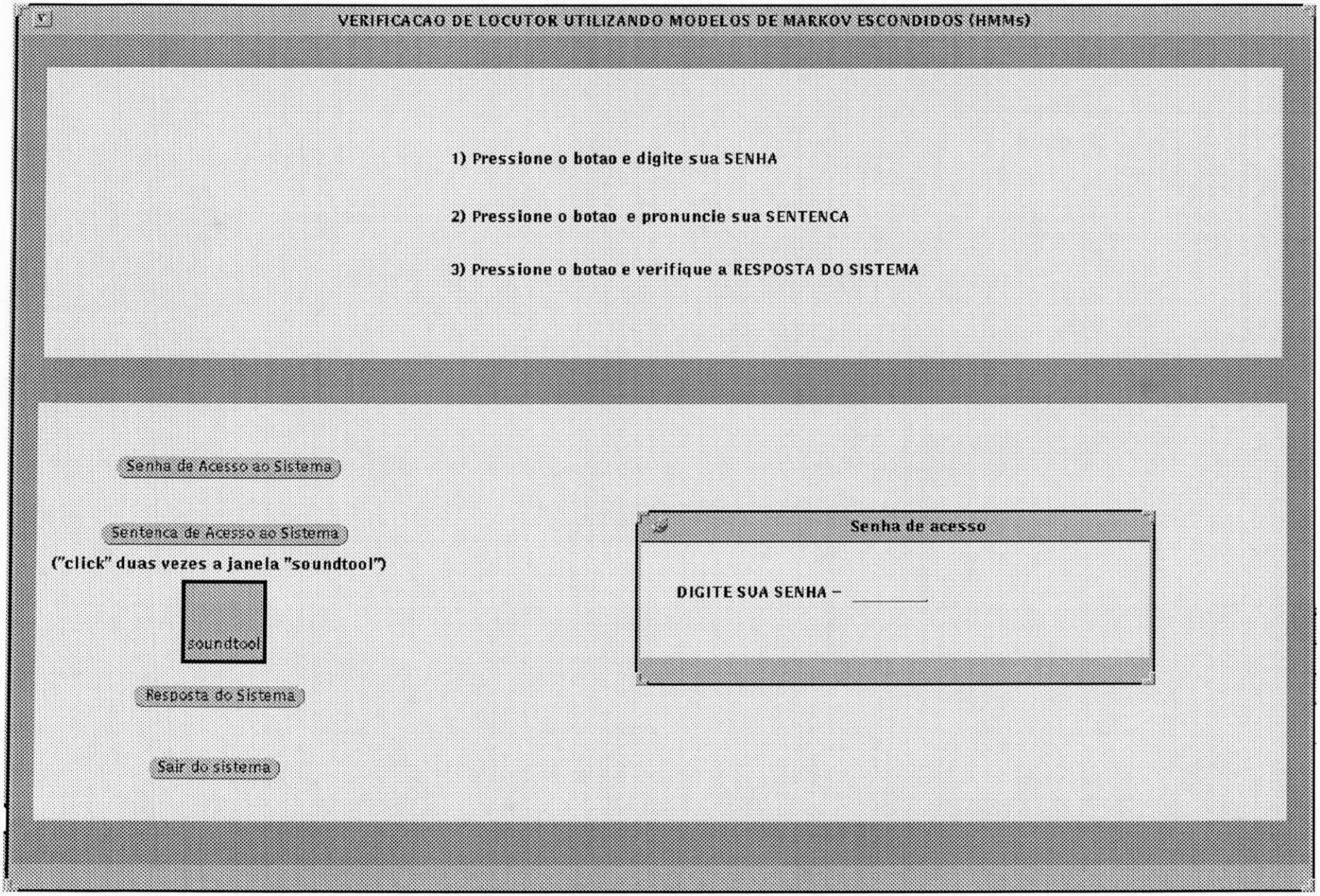
Apêndice C

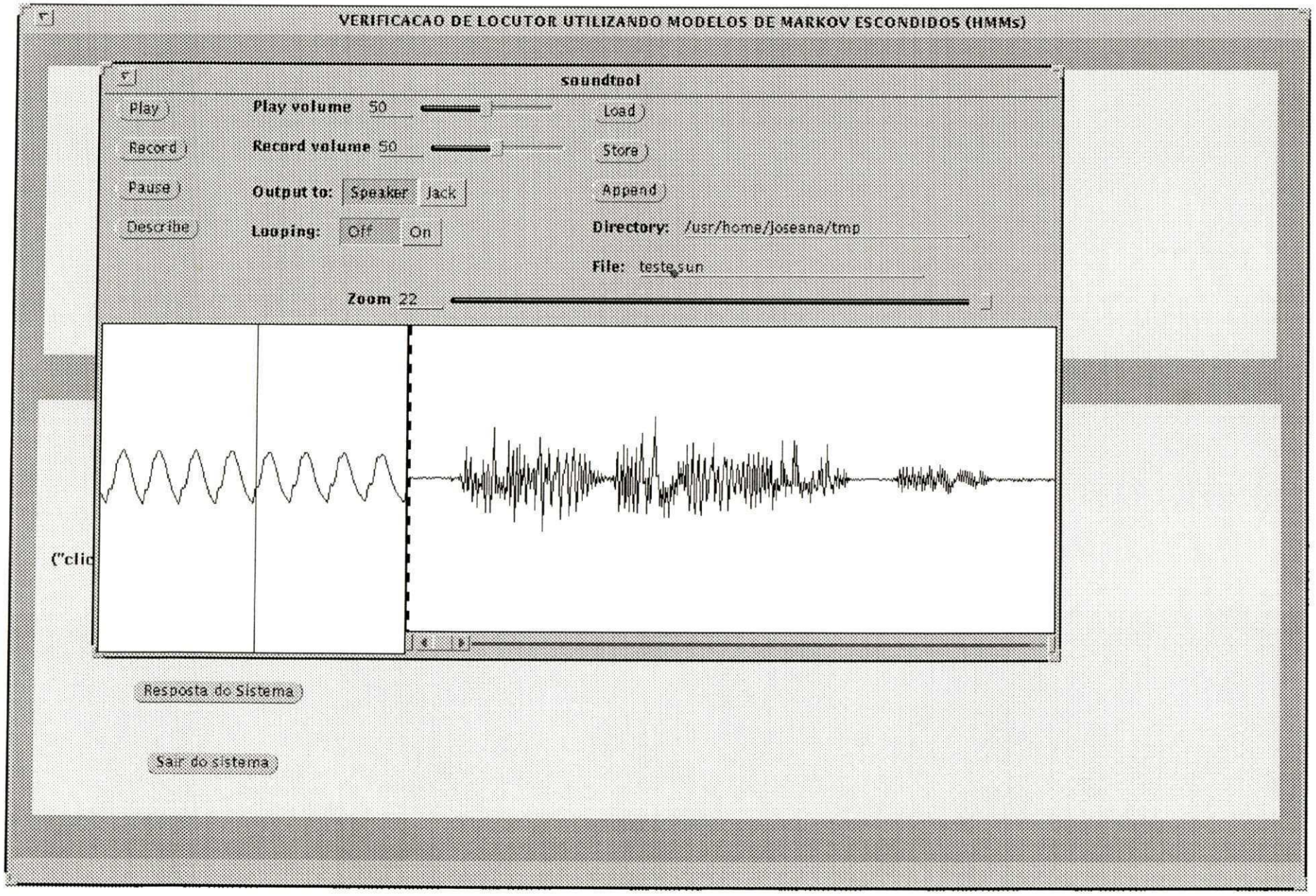
Interface do Sistema

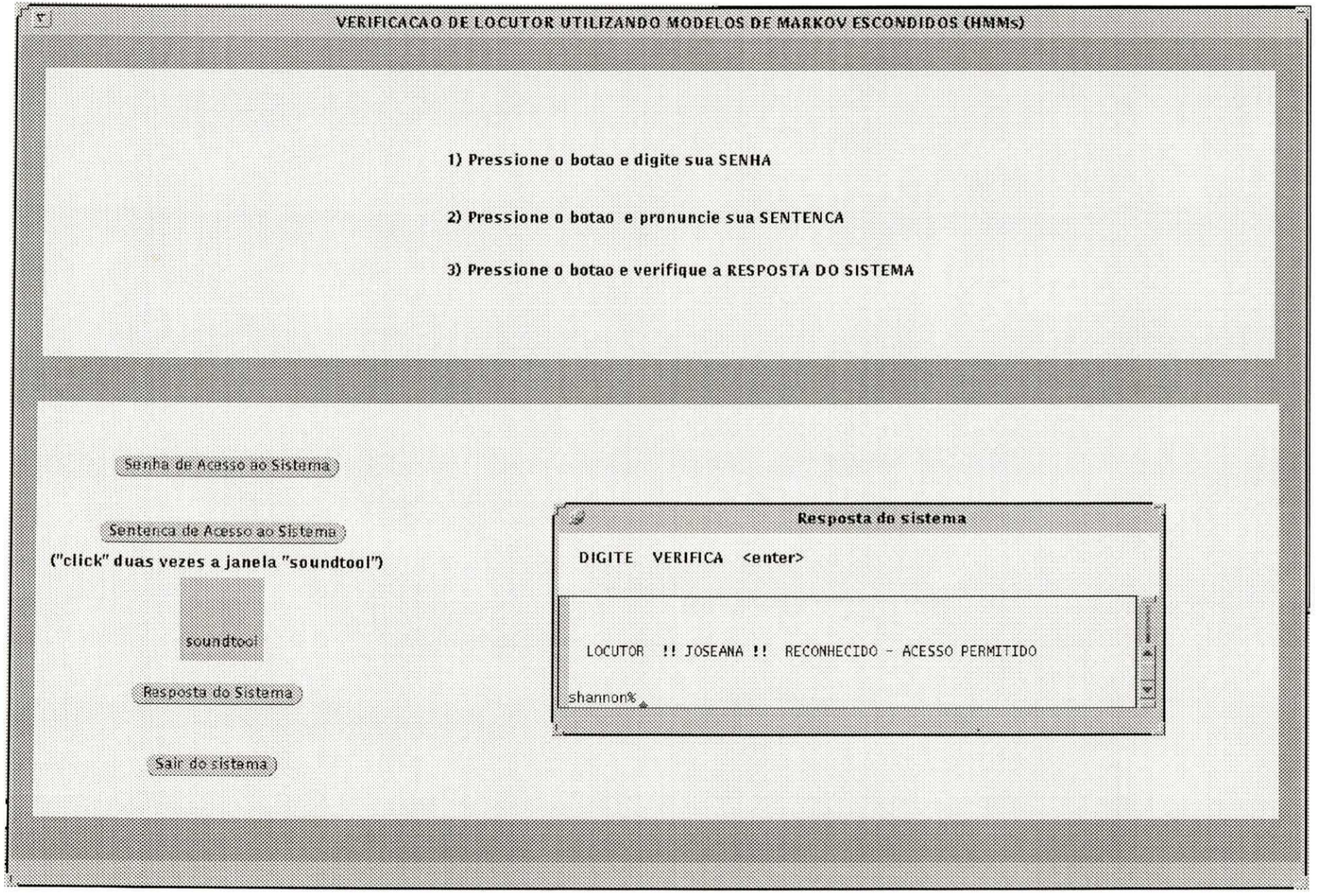
A interface utilizada no sistema de verificação de locutor proposto neste trabalho, foi implementada utilizando a ferramenta gráfica GUIDE para o ambiente OPENWINDOWS, implementada em uma estação de trabalho SUN. Nas páginas seguintes, serão apresentados os "Menus" correpondentes a cada opção do sistema de acordo com a seguinte seqüência:

1. "Menu" Principal.
2. "Menu 1" - Digitar a Senha de Acesso.
3. "Menu 2" - Pronunciar a Sentença de Acesso.
4. "Menu 3" - Resposta do Sistema.


VERIFICACAO DE LOCUTOR UTILIZANDO MODELOS DE MARKOV ESCONDIDOS (HMMs)
1) Pressione o botao e digite sua SENHA
2) Pressione o botao e pronuncie sua SENTENCA
3) Pressione o botao e verifique a RESPOSTA DO SISTEMA
Senha de Acesso ao Sistema
Sentenca de Acesso ao Sistema
("click" duas vezes a janela "soundtool")
soundtool
Resposta do Sistema
Sair do sistema


VERIFICACAO DE LOCUTOR UTILIZANDO MODELOS DE MARKOV ESCONDIDOS (HMMs)
1) Pressione o botao e digite sua SENHA
2) Pressione o botao e pronuncie sua SENTENCA
3) Pressione o botao e verifique a RESPOSTA DO SISTEMA
Senha de Acesso ao Sistema
Sentenca de Acesso ao Sistema
("click" duas vezes a janela "soundtool")
soundtool
Resposta do Sistema
Sair do sistema
Senha de acesso
DIGITE SUA SENHA -


VERIFICACAO DE LOCUTOR UTILIZANDO MODELOS DE MARKOV ESCONDIDOS (HMMs)
soundtool
Play)
Play volume 50
Load)
Record)
Record volume 50
Store)
Pause)
Output to: Speaker Jack
Append
Describe
Looping: Off On
Directory: /usr/home/joseana/tmp
File: teste.sun
Zoom 22
("clic
Resposta do Sistema)
Sair do sistema)


VERIFICACAO DE LOCUTOR UTILIZANDO MODELOS DE MARKOV ESCONDIDOS (HMMs)
1) Pressione o botao e digite sua SENHA
2) Pressione o botao e pronuncie sua SENTENCA
3) Pressione o botao e verifique a RESPOSTA DO SISTEMA
Senha de Acesso ao Sistema
Sentenca de Acesso ao Sistema
("click" duas vezes a janela "soundtool")
soundtool
Resposta do Sistema
Sair do sistema
Resposta do sistema
DIGITE VERIFICA <enter>
LOCUTOR !! JOSEANA !! RECONHECIDO - ACESSO PERMITIDO
shannon%

Referências

[1] Fagundes, R. D. R. and Alens, N (1993). "Reconhecimento de Voz, Linguagem Contínua, Usando Modelos de Markov". *11º Simpósio Brasileiro de Telecomunicações - SBT*, Setembro.

[2] Fechine, J. M. and Aguiar Neto, B. G. (1993). "Modelamento de Identidade Vocal Utilizando Modelos de Markov Escondidos". *XVI Congresso Nacional de Matemática Aplicada e Computacional - CNMAC*, Setembro.

[3] Davis et al, K. H. (1952). "Automatic Recognition of Spoken Digits". *Journal of the Acoustical Society of America*, 24(6):637–642.

[4] Koenig, W. (1946). "The Sound Spectrograph". *Journal of the Acoustical Society of America*, 17:19–49.

[5] Vieira, M. N. (1989). "Módulo Frontal para um Sistema de Reconhecimento Automático de Voz". *Universidade de Campinas - Dissertação de Mestrado*, Dezembro.

[6] Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals.* Prentice Hall, USA.

[7] O'Shaughnessy, D. (1986). "Speaker Recognition". *IEEE transactions on Acoustics, Speech, and Signal Processing Magazine*, pages 4–17, October.

[8] Doddington, G.R. (1985). "Speaker Recognition - Identifying People by their Voices". *Proceedings IEEE*, 73(11):1651–1664, November.

[9] Furui, S. (1981). "Cepstral Analysis Technique for Automatic Speaker Verification". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, April.

[10] Bennani, Y., Fogelman Soulie, F. and Gallinari, P. (1990). "Text-Dependent Speaker Identification Using Learning Vector Quantization".

[11] Rabiner, L. R. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". *Proceedings IEEE*, 77(2):257–286, February.

[12] Levinson, S. E., Rabiner, L. R. and Sondhi M. M. (1983). "An Introduction to the Application of the Theory of Probabilist Functions of a Markov Process to Automatic Speech Recognition". *The Bell System Technical Journal*, 62(4):1035–1068, April.

[13] Rabiner, L. R., Levinson, S. E. and Sondhi, M. M. (1982). "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition". *The Bell System Technical Journal*, 62(4):1075–1105, April.

[14] Flanagan, L. J. (1978). *Speech Analysis Synthesis and Perception.* Murray Hill 2ª Edição, New Jersey.

[15] Fellbaum, K. (1984). *Sprachsignalverarbeitung and Sprachübertragung.* Springer-verlag, Berlin.

[16] Atal, B. S. and Hanauer, S. L. (1971). "Speech Analysis and Synthesis by Linear prediction of the Speech Wave". *J. Acoust. Soc. Am.*, 50(2 (Part 2)):637–655, August.

[17] Rabiner, L. R. (1968). "Digital Formant Synthesizer for Speech Synthesis Studies". *J. Acoust. Soc. Am.*, 43(4):822–828, April.

[18] Winham, G. and Steiglitz, K. (1970). "Input Generators for Digital Sound Synthesis". *J. Acoust. Soc. Am.*, 47(2):665–666, February.

[19] Silva, A. J. S. (1992). "Quantização Vetorial: Aplicações a um Vocoder LPC". *Universidade Federal da Paraíba - Dissertação de Mestrado*, Dezembro.

[20] Makhoul, J. (1975). "Linear Prediction: A Tutorial Review". *Proceedings IEEE*, 63:561–580.

[21] Liu, J. (1989). *Zur Untersuchung und Optimierung von Spracherkennungssystemen für isoliert gesprochene Wörter.* VDI VERLAG, Düsseldorf.

[22] Rabiner L. R., Juang, B. H., Levinson, S. E. and Sondhi, M. M. (1985). "Recognition of Isolated Digits Using Hidden Markov Models With Continuos Mixture Densities". *ATT Technical Journal*, 64(6):1211–1234, July-August.

[23] Makhoul, J., Roucos, S. and Gish, H. (1985). "Vector Quantization in Speech Coding". *Proceedings IEEE*, 73(11):1551–1588, November.

[24] Linden, Y. and Gray, R. M. (1980). "An Algorithm for Vector Quantizer Design". *IEEE transactions on Acoustics, Speech, and Signal Processing*, 28(1):84–95, January.

[25] Gray, R. M. (1984). "Vector Quantization". *IEEE transactions on Acoustics, Speech, and Signal Processing Magazine*, 1:4–29, April.

[26] Satish, L. and Gururaj, B. I. (1993). "Use of Hidden Markov Models for Partial Discharge Pattern Classification". *IEEE Transactions on Electrical Insulation*, 28(2):172–182, April.

[27] Rabiner L. R., Juang, B. H., Levinson, S. E. and Sondhi, M. M. (1985). "Some Properties of Continuos Hidden Markov Model Representations". *ATT Technical Journal*, 64(6):1251–1270, August.

[28] Bennani, Y., Folgelman Soulie, F. and Gallinari, P. (1990). "A Connectionist Approach for Automatic Speaker Identification". *Proceedings IEEE*, pages 265–268.

[29] Savic, M. and Gupta K.S. (1990). "Variable Parameter Speaker Verification System Based On Hidden Markov Modeling". *Proceedings IEEE*, pages 281–284.

[30] Buzo, A., Gray, Jr., A. H., Gray, R. M. and Markel, J. D. (1980). "Speech Coding Based Upon Vector Quantization". *IEEE transactions on Acoustics, Speech and Signal Processing*, ASSP-28(5):562–574, October.

[31] Juang, B. H., Wong, D. Y. and Gray, Jr., A. H. (1982). "Distortion Performance of Vector Quantization for LPC Voice Coding". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-30(2):294–304, April.

[32] Rabiner, L. R. and Juang, B. H. (1986). "An Introduction to Hidden Markov Models". *IEEE Acoustics, Speech, and Signal Processing Magazine*, 3(1):4–16, February.

[33] Rabiner, L. R. and Levinson, S. E. (1985). "A Speaker-Independent, Syntax-Directed, Connected Word Recognition System Based On Hidden Markov Models and Level Building". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33(3):561–573, Juny.

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

[35] Juang, B. H. (1985). "Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains". *ATT Technical Journal*, 64(6):1235–1249, July-August.

[36] Juang, B. H. and Levinson, S. E. (1986). "Maximum likelihood estimation for multivariate mixture observations Markov chains". *IEEE Transactions Informat. Theory*, IT-32(2):307–309, March.

[37] Oglesby, J. and Mason, J. S. (1990). "Optimisation of Neural Models for Speaker Identification". *Proceedings IEEE*, pages 261–264.