Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística

Joseana Macêdo Fechine

Tese de Doutorado submetida à 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 Doutor em Ciências no domínio da Engenharia Elétrica.

Área de Concentração: Processamento da Informação

Benedito Guimarães Aguiar Neto - Dr.-Ing.
Orientador

Campina Grande, Paraíba, Brasil

Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística

Joseana Macêdo Fechine

Benedito Guimarães Aguiar Neto - Dr.-Ing.
Orientador

Abraham Alcaim - Ph.D
Componente da Banca

Adrião - Ph.D
Componente da Banca

Marcus Antônio Brasileiro - Ph.D
Componente da Banca

Marcelo Sampaio de Alencar - Ph.D
Componente da Banca

Campina Grande, Paraíba, Brasil

Dedico este trabalho a Deus em primeiro lugar, aos meus pais, José e Ana Ildaíza, aos meus irmãos, Vicente, Geovane e Guilhermino e aos meus sobrinhos, Mariana, Melina e Gabriel.

"Todos nós temos uma soma de deveres a cumprir.

A vida exige de cada um o direito de lutar e vencer."

J.S. Nobre

Agradecimentos

A realização deste trabalho recebeu o apoio de muitos que me ensinaram a ter perseverança para seguir com o presente estudo. A todos o meu agradecimento e a certeza de que as palavras e gestos de incentivo não foram inúteis, sabendo que o verdadeiro agradecimento consiste no reconhecimento daqueles que contribuem para o sucesso de outros, ainda que involuntariamente.

Em especial agradeço a Deus, pelo amor infinito.

A minha família, pela paciência, apoio e incentivo sempre presentes.

Ao professor Benedito Guimarães Aguiar Neto, pela orientação deste trabalho, estímulo e dedicação sempre prestados, que muito me enriqueceram intelectualmente, fortalecendo o meu desenvolvimento profissional.

Ao meu amigo Francisco Madeiro Bernardino Júnior, pelo apoio, incentivo e pela valorosa contribuição.

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

Aos amigos Paulo Márcio, Rinaldo, Waslon Terllizzie, Eustáquio, ... e demais colegas do LAPS.

Aos demais amigos que compuseram a amostra de locutores: Isabel, Sissi, Suzete, professora Maria de Fátima, Rute, Marta, Socorro, Camila, Vânia, Yuska, Ellaine, Claudia, Vivian, Natasha, Mariana, Renata, Karina, Josemar, Leonel, Bruno, Denis, Antônio Neto, Alynthor, Luiz Gonzaga Júnior, Felipe, Avishek, Murali, Sérgio, Towar e Edmar.

A todos que fazem a COPELE, em especial à Ângela, Pedrinho e Eleonôra, pelo apoio constante.

A Joab e Antonio Carlos da ATECEL, pelas palavras de incentivo e apoio, como também pela gentileza no atendimento das minhas solicitações.

As minhas grandes amigas Kátia, Magna, Kíssia e Kenia, que tanto me apoiaram em todos os momentos.

A todos os meus amigos, que direta ou indiretamente me incentivaram no decorrer deste trabalho.

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

Resumo

Este trabalho trata da aplicação de uma técnica híbrida (paramétrica e estatística), que utiliza Análise por Predição Linear, Quantização Vetorial, Redes Neurais e Modelos de Markov Escondidos, para o desenvolvimento de um sistema de reconhecimento (identificação) automático da identidade vocal, visando obter alternativas para os algoritmos tradicionais. Com o objetivo de se obter um sistema mais rápido e robusto, é realizada uma etapa de pré-identificação, seguida da identificação. A primeira etapa utiliza a freqüência fundamental (F_0) como parâmetro de separação prévia dos locutores em grupos gerais, de acordo com o sexo. O método proposto para estimação da F_0 se mostra eficiente (99% de classificação correta), fornecendo estimativas representativas de cada locutor, reduzindo assim o número de locutores a participar da etapa posterior. A etapa de identificação utiliza Modelos de Markov Escondidos (HMMs) de Densidades Discretas e Quantização Vetorial Paramétrica, com parâmetros acústicos obtidos a partir da Análise por Predição Linear (coeficientes LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados). Os coeficientes Cepestrais, seguido dos Delta Cepestrais, proporcionam maiores taxas de identificação. Em se tratanto do projeto do dicionário do quantizador vetorial, são avaliados três algoritmos: LBG (Linde-Buzo-Gray), KMVVT (Kohonen Modificado com Vizinhança Centrada em Torno do Vetor de Treino) e SSC (Competitivo no Espaço Sináptico). O algoritmo SSC apresenta-se como o mais adequado para o projeto dos dicionários, levando a maiores taxas de identificação. A modelagem por HMMs se constitui em uma etapa de "refinamento" do processo de identificação, sendo utilizada quando as medidas de distorção obtidas pela comparação do padrão de teste do locutor a ser identificado (vetor de características acústicas) com os padrões de referência (dicionários do quantizador vetorial) indicarem "similaridade" entre os padrões vocais. A técnica aplicada neste trabalho proporciona a obtenção de um sistema de reconhecimento automático da identidade vocal que apresenta taxa média de identificação elevada (97,8%) e significativa, baixas taxas médias de falsa aceitação (0,8%) e de falsa rejeição (1,5%), bem como alta confiabilidade (99,2%). O sistema de identificação de locutor desenvolvido é, portanto, capaz de discriminar, de forma eficiente, os locutores a partir das suas características vocais apresentando, independentemente do sexo do locutor, pequenas variações intralocutor e grandes variações interlocutor.

Abstract

This work presents an investigation concerning the use a hybrid system (parametric and statistic) composed by Linear Prediction, Vector Quantization, Neural Networks and Hidden Markov Models (HMMs) with discrete densities applied to speaker identification. Several parameters, such as coefficients obtained by Linear Prediction Coding (LPC, Cepstrum, Weighted Cepstrum, Delta Cepstrum and Delta Weighted Cepstrum methods) are used to represent each speaker. In order to achieve a robust identification, a two-step system is designed, consisting of a pre-identification stage followed by an identification stage (main stage). The first stage uses the pitch (or fundamental frequency) to distinguish two subgroups (male and female). The proposed method to estimate the pitch produces high pre-identification rate (99%), reducing the set of speakers to be identified in subsequent stage. The main stage is divided into two substages. The first uses vector quantization with codebooks designed by LBG (Linde-Buzo-Gray), KMVVT (Modified Kohonen's Algorithm with Neighborhood Centered in the Training Vector) and SSC (Synaptic Space Competitive) algorithms. Results show that the codebooks of acoustic patterns designed by SSC lead to higher identification rates when compared to the ones designed by KMVVT and LBG. Additionally, this work presents a comparative study of the linear predictive analysis methods applied to speaker identification. Cepstrum and Delta Cepstrum coefficients produce better results when compared to other coefficients. The second substage uses HMMs when the acoustic patterns indicate that speakers present similar vocal characteristics. Thus, the second substage is a refinement of the main stage. The system produces high mean identification rate (97,8%), small mean false acceptance rate (0,8%) and mean false rejection rate (1,5%), as well as high confiability (99,2%). The results show that the speaker identification system which is able to efficiently descriminate the vocal characteristics of the speakers (female and male), with a small intra-speaker and a large inter-speaker variation.

Índice

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas

SRAL - Sistema de Reconhecimento Automático de Locutor

HMM - Hidden Markov Model (Modelo de Markov Escondido)

VQ - Vector Quantization (QV - Quantização Vetorial)

DTW - Dynamic Time Warping (Alinhamento Dinâmico no Tempo)

E_{seg} - Energia por segmento (segmental)

TCZ - Taxa de Cruzamento por Zero

NTP - Número Total de Picos

DNP - Diferença entre os Picos

PPOS - Número de Picos Positivos

PNEG - Número de Picos Negativos

RAL - Reconhecimento Automático de Locutor

LBG - Algoritmo para projeto de dicionários conhecido como Linde-Buzo-Gray

KMVVT - Algoritmo de Kohonen Modificado com Vizinhança Centrada em Torno do Vetor de Treino

AMDF - Average Magnitude Difference Function (Função da Média de Diferenças de Amplitudes)

FFT - Fast Fourier Transform (Transformada Rápida de Fourier)

LPC - Linear Prediction Coding

CEP - Coeficientes Cepestrais

CEP-P - Coeficientes Cepestrais Ponderados

DCEP - Coeficientes Delta Cepestrais

DCEP-P - Coeficientes Delta Cepestrais Ponderados

LFi - i-ésimo locutor feminino

LMi - i-ésimo locutor masculino

Ei - i-ésima elocução

F_0 - Freqüência Fundamental

P_0 - Período de Pitch (Período da Freqüência Fundamental)

$s(n)$ - sinal de voz

N_A - tamanho do "quadro" de amostras do sinal

$\mu_{s(n)}$ - média do sinal $s(n)$

$sgn[s(n)]$ - número de vezes que o sinal $s(n)$ inverte a polaridade

ρ_1 - primeiro coeficiente de correlação

$c_{s(n)s(n-1)}$ - covariância entre $s(n)$ e $s(n-1)$

$\sigma_{s(n)}$ - desvio padrão de $s(n)$

$R_{ss}(1)$ - primeiro coeficiente de autocorrelação

$S(z)$ - transformada Z do sinal $s(n)$

$G(z)$ - transformada Z do modelo do pulso glotal $g(n)$

$A_s(n)$ e $A_f(n)$ - intensidade da excitação dos sinais de voz e de ruído, respectivamente.

$V(z)$ - transformada Z do modelo do trato vocal $v(n)$

$R(z)$ - transformada Z do modelo da radiação $r(n)$

$H(z)$ - transformada Z da função de transferência $h(n)$

$U(z)$ - transformada Z do sinal de excitação $u(n)$

T - período de amostragem

K - tamanho do vetor de características acústicas do sinal

F_1, F_2 e F_3 - três primeiras freqüências formantes

P - período do sinal

$d(n)$ - diferença entre amostras do sinal

E_{seg1} e E_{seg2} - energia de cada metade do "quadro" em análise

E_1, E_2, E_3 - limiares de energia

$suso'$ - decisão surdo-sonoro inicial para o quadro em análise

$suso_{-1}$ - decisão surdo-sonoro do último quadro

$suso_{-2}$ - decisão surdo-sonoro do penúltimo quadro

$suso$ - decisão surdo-sonoro do quadro atual

max = amplitude máxima da AMDF

min = amplitude mínima da AMDF

$minp$ = posição do mínimo da AMDF

c_k - k-ésimo coeficiente LPC (coeficiente do filtro)

G - ganho do filtro

$\widetilde{s}(n)$ - estimativa de $s(n)$

$e(n)$ - erro de predição

$vs(n)$ - sinal de voz selecionado e ponderado

$\widetilde{v}s(n)$ - aproximação de $vs(n)$

$Erro(n)$ - Erro quadrático

$R_r(k)$ - função de autocorrelação para curtos intervalos

$ce_i(n)$ - n-ésimo coeficiente Cepestral no i-ésimo bloco de amostras

X_i - i-ésimo bloco do espectro de potência do sinal

$jp(n)$ - janela de ponderação

$cp_i(n)$ - n-ésimo coeficiente Cepestral Ponderado no i-ésimo bloco de amostras

$\Delta ce_i(n)$ - n-ésimo coeficiente Delta Cepestral no i-ésimo bloco de amostras

$\Delta cp_i(n)$ - n-ésimo coeficiente Delta Cepestral Ponderado no i-ésimo bloco de amostras.

ϕ - constante de normalização

M - tamanho do dicionário

$\vec{x}$ - vetor de entrada

$\hat{\vec{x}}$ - vetor de reprodução

W - alfabeto de reprodução

$\vec{w}_i$ - vetores do alfabeto de reprodução

S - partição do espaço vetorial

C_i - células do quantizador

$q(x)$ - quantizador de $\vec{x}$

$||D_M||$ - medida de distorção do quantizador vetorial

$d(\vec{x}, \hat{\vec{x}})$ - distorção - erro médio quadrático

$\hat{A}_0$ - alfabeto de reprodução inicial

$\vec{w}_{i^*}$ - neurônio vencedor

$\eta(n)$ -taxa de aprendizagem na n-ésima iteração

$\mathcal{O}_i$ - função que define a vizinhança em torno do neurônio vencedor

Δw_{ij} - modificação introduzida na j-ésima componente (sinapse) do neurônio

$r(n)$ - raio de vizinhança, medido na grade bidimensional

$d_g(\cdot)$ - distância medida na grade

q_i - i-ésimo estado do HMM

N - número de estados do HMM

$\mathcal{A} = [a_{ij}]$ - matriz transição de estados do HMM

$\mathcal{B}$= $[b_j(k)]$ - matriz de função de probabilidade das observações do HMM

$\pi = \ \pi_i$ - vetor de probabilidade do estado inicial do HMM

L - número de locutores

$\mathbf{O^l}$ - vetor de observações do l-ésimo locutor

λ_l - modelo do HMM referente ao l-ésimo locutor

$\overline{P_l}$ - probabilidade associada ao l-ésimo locutor

$\alpha_t(i)$ - probabilidade de avanço (*forward probability*)

$\beta_t(i)$ - probabilidade de retrocesso (*backward probability*)

$\delta_t(i)$ - maior valor de probabilidade ao longo de um único caminho

q_t^* - seqüência de estados ótima

$L(Z)$ - transformada Z do filtro $l(n)$

a_p - fator de pré-ênfase

$s_p(n)$ - sinal de voz após a pré-ênfase

$J(n)$ - janela (Retangular, Hamming ou Hanning)

$s'(n)$ - seqüência de voz filtrada

f_s - freqüência de amostragem

esc_t - coeficiente de escalonamento

L_I e L_S - limites inferior e superior, respectivamente, da freqüência fundamental

F_{0F} - freqüência fundamental feminina

F_{0M} - freqüência fundamental masculina

LF e LM - locutor feminino e masculino, respectivamente

C.V. - Coeficiente de Variação

s - desvio padrão da amostra

n_A - tamanho da amostra

EP_m - erro padrão da média

1-α - *nível de confiança* ou *grau de confiança*

μ - média populacional

$\overline{x}$ - média amostral

σ - desvio padrão populacional

σ^2 - variância populacional

Capítulo 1

Introdução

1.1 Comunicação Vocal Homem-Máquina

O ser humano sempre buscou meios de comunicação que facilitassem a interação com a máquina. Em função disso e do crescente desenvolvimento tecnológico de *hardware* para o processamento digital de sinais, o meio de comunicação mais adequado seria a fala humana. Tal meio de comunicação proporciona uma cômoda adaptação do usuário e a capacidade de transmitir uma grande quantidade de informações com pouca interação. Os métodos tradicionais de identificação de pessoas requerem a apresentação de um objeto (chave, cartão, etc.) ou uma mensagem fornecida através de um teclado (senha, etc.). Muitos desses métodos são impraticáveis em sistemas de telecomunicações e apresentam a desvantagem de não serem diretamente dependentes da pessoa, visto que as pessoas podem perder seu cartão ou esquecer sua senha.

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

A partir, unicamente da voz, é possível 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 grande quantidade de outras características.

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

Com o desenvolvimento tecnológico foi 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 à 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, 3, 4].

Os primeiros trabalhos descrevendo máquinas que podiam, de alguma forma, reconhecer com certo sucesso a pronúncia de determinadas palavras datam de 1952 [5]. Uma grande quantidade de trabalhos sobre o assunto surgiu nos anos 60, graças às descobertas de algumas propriedades da voz através do uso de espectógrafos [6] 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 século XVIII, quando foram elaborados curiosos engenhos acústicos que produziam sons semelhantes à voz e eram "tocados" à maneira de um instrumento musical [2].

Além da facilidade de comunicação, a voz oferece muitas outras vantagens na interação com as máquinas como, por exemplo, a velocidade: a maioria das pessoas pode falar facilmente a taxas de 200 palavras por minuto; por outro lado, poucas pessoas podem digitar, em um teclado, mais de 60 palavras por minuto [7].

A entrada vocal é bastante adequada para aplicações em que uma ou mais das seguintes condições se aplicam: as mãos do usuário estão ocupadas; mobilidade é exigida durante o processo de entrada de dados; os olhos do operador devem permanecer fixos sobre um *display*; um instrumento óptico ou algum objeto é rastreado; é incoveniente o uso de teclado em um ambiente, dentre outras. Por não requererem nem as mãos nem os olhos do usuário para sua operação, os sistemas de entrada vocal podem ser utilizados em diversas aplicações, como por exemplo: controle de tráfego aéreo, auxílio a deficientes físicos, controle de qualidade e inspeção e controle de acesso a ambientes restritos [3].

A identificação da voz tem a conveniência da facilidade de coleção de dados. Outra vantagem dessa técnica, quando comparada com outras técnicas, por exemplo, o exame de fundo de olho, impressões digitais e assinaturas, se refere a sua facilidade de utilização em sistemas em que se exige o reconhecimento à distância; por exemplo transações bancárias por telefone. Além disso, a voz não pode ser perdida nem tão

pouco esquecida, diferentemente dos outros métodos de identificação, tais como cartões magnéticos e senhas numéricas [3].

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 subáreas principais [1]:

1. Resposta Vocal;
2. Reconhecimento de Fala;
3. Reconhecimento de Locutor.

A Figura 1.1 mostra uma descrição geral do processamento da voz, para a tarefa de reconhecimento, com ênfase ao reconhecimento de locutor (objeto de estudo deste trabalho) e a relação entre as suas subáreas [8].

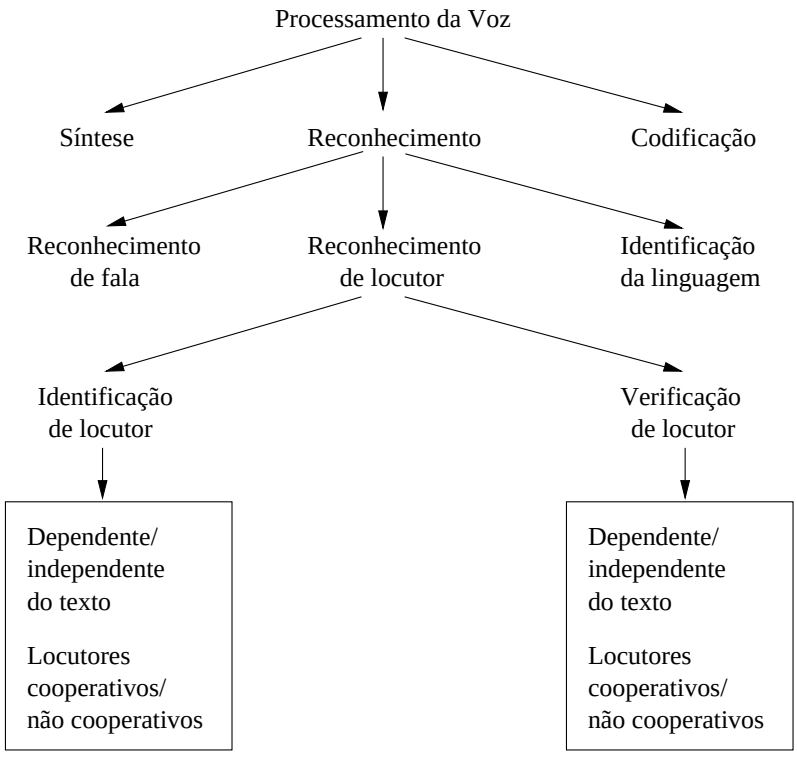


Figura 1.1: Descrição geral do processamento da voz.

1.1.1 Sistemas de Resposta Vocal

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 [1].

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

Todos os métodos de codificação de forma de onda conhecidos (PCM [1], PCM diferencial, PCM diferencial adaptativo, etc.) e métodos de análise-síntese (técnicas de codificação preditiva linear) podem ser usados para armazenar os elementos de texto. A escolha do método a ser utilizado é uma função da qualidade da reprodução das mensagens e da capacidade de armazenamento exigidos pelo sistema. A qualidade da voz depende, essencialmente, do método de codificação utilizado. Os valores dos parâmetros derivados dessa representação são, então, usados para controlar um sintetizador de voz que modela a produção da voz humana.

Alguns problemas que ainda devem ser solucionados quanto à síntese de voz incluem, entonação incorreta de frases e pronúncia errônea de palavras mais complexas, ou de combinações de palavras [9].

1.1.2 Sistemas de Reconhecimento de Fala

Nos sistemas de reconhecimento de fala a comunicação vocal é feita do homem para a máquina. O reconhecimento de fala, pode ser subdividido em um grande número de subáreas dependendo de alguns fatores, tais como, tamanho do vocabulário, população de locutores, etc [1].

A tarefa básica no reconhecimento de fala é reconhecer uma determinada elocução de uma sentença ou "entender" um texto falado (ou seja, responder de forma correta ao que está sendo falado) [1]. O conceito de entendimento, ao invés de reconhecimento, é de grande importância para sistemas que tratam com entrada de voz contínua com grande vocabulário, enquanto que o conceito de reconhecimento exato é de maior importância para sistemas de palavras isoladas, vocabulário limitado e pequeno número

[1] Modulação por Codificação de Pulsos

de usuários [1, 10].

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

De uma forma geral, os sistemas de reconhecimento automático de fala podem ser considerados como pertencentes a uma das seguintes categorias [1]:

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

Os sistemas de reconhecimento de palavras isoladas podem ser definidos como aqueles sistemas que exigem uma pausa curta antes e depois das sentenças que devem ser reconhecidas [11].

O modo de entrada de palavras conectadas pode ser conveniente para o usuário porque se assemelha à maneira mais natural de se falar, contudo esse tipo de comunicação tem algumas limitações em vista do presente estágio da tecnologia de reconhecimento de fala [1].

Os sistemas dependentes do locutor são caracterizados por serem treinados para obedecerem às características específicas da voz dos seus usuários [1].

Os sistemas de reconhecimento independente do locutor, ou sistemas "insensíveis" ao locutor, podem ser definidos como aqueles que não estão presos às características específicas da voz do usuário [1].

1.1.3 Sistemas de Reconhecimento de Locutor

O objetivo de um sistema de reconhecimento de locutor é reconhecer um locutor a partir da sua voz, sendo bastante útil em aplicações de segurança, como por exemplo o controle de acesso a ambientes restritos (utilização da voz para abrir e fechar portas) e

o controle de acesso de dados em computadores. Em criminalística, pode ser utilizado com o mesmo propósito que hoje é dado às impressões digitais [1]. Nesse contexto, os Sistemas de Reconhecimento Automático de Locutor (SRALs) constituem uma das principais áreas da comunicação vocal homem-máquina [1].

Nos sistemas de reconhecimento de locutor, da mesma forma que nos sistemas de reconhecimento de fala, a comunicação vocal é feita do homem para a máquina.

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

Duas classes de aplicações são desenvolvidas baseadas em sistemas de reconhecimento de locutor: identificação de locutor e verificação de locutor. Aplicações para identificação de locutor buscam responder a seguinte questão: "Quem é você ?", enquanto que aplicações para verificação de locutor buscam responder: "Você é mesmo quem alega ser ?" [8, 12, 13].

A identificação de locutor é um processo de determinação da identidade de um locutor dentre vários locutores, pela comparação do sinal de voz deste locutor (sinal de entrada) com os demais, escolhendo o que proporcionar o melhor "casamento" com o sinal de voz de entrada [1, 14].

A verificação de locutor tem por objetivo determinar, automaticamente, se a identidade de um pretenso locutor é verdadeira ou não [1, 15].

A Figura 1.2 mostra a representação geral de um problema de reconhecimento de locutor [16].

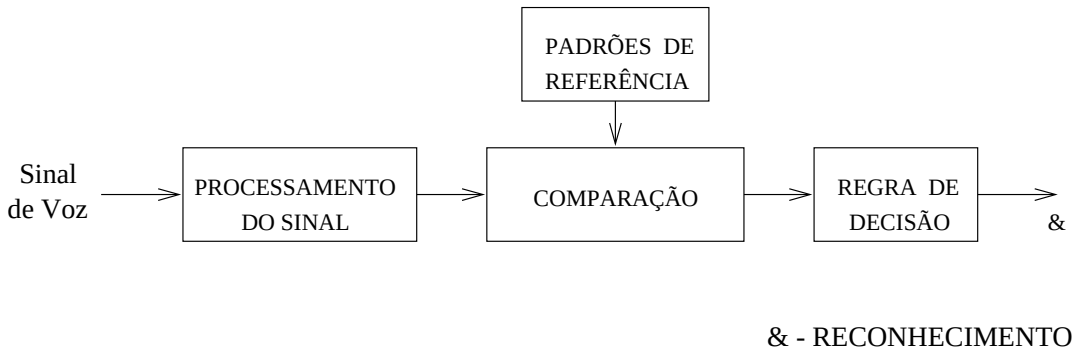


Figura 1.2: Modelo genérico para um sistema de reconhecimento de locutor.

O reconhecimento de locutor é uma tarefa de reconhecimento de padrões. Em

essência requer um mapeamento entre identificação de voz e locutor, tal que cada possível forma de onda de entrada é identificada com seu locutor correspondente.

Para a implementação de um sistema de reconhecimento de locutor deve-se obter, para cada locutor, um conjunto de parâmetros representativos da sua voz. Os parâmetros obtidos irão compor um modelo (ou padrão) representativo do locutor. Nesse sistema o locutor será aceito ou rejeitado, a partir da comparação dos seus parâmetros (padrão) de teste com os parâmetros já armazenados (padrões de referência), utilizando-se uma regra de decisão.

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** [1]. Portanto, 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.

Na verificação de locutor, uma identidade é alegada pelo usuário e a decisão requerida pelo sistema é estritamente binária, isto é, consiste simplesmente em aceitar ou rejeitar a identidade alegada.

A literatura aborda, com diferentes termos, a verificação de locutor, incluindo denominações tais como: verificação da voz, autenticação do locutor, autenticação da voz e verificação do locutor [8].

O problema da identificação de locutor difere significativamente do problema da verificação de locutor, uma vez que, nesse caso, o sistema é requisitado a fazer uma identificação entre todos locutores. Assim, em vez de uma única comparação entre um conjunto de medidas e um padrão de referência armazenado, torna-se necessário um número de comparações igual ao número de locutores. Este tipo de reconhecimento pode ocorrer de duas formas: conjunto-aberto (o locutor pode não estar entre a população) e conjunto-fechado (sabe-se a priori que o locutor é um membro da população).

Descrições gerais de sistemas para o reconhecimento de locutor têm sido mostradas em [3, 8, 13, 17, 18, 19].

Todas as tarefas de reconhecimento de padrões, inclusive o reconhecimento de locutor, utilizam duas fases: treinamento (Figura 1.3) e reconhecimento (Figura 1.4).

Na fase de treinamento é estabelecido um dicionário de padrões de referência de voz, aos quais são atribuídos rótulos que identificam o locutor. Na fase de reconhecimento são obtidos padrões de teste que são comparados com todos os padrões de referência e então, utilizando-se uma regra de decisão, é identificado aquele mais semelhante ao padrão de entrada desconhecido.

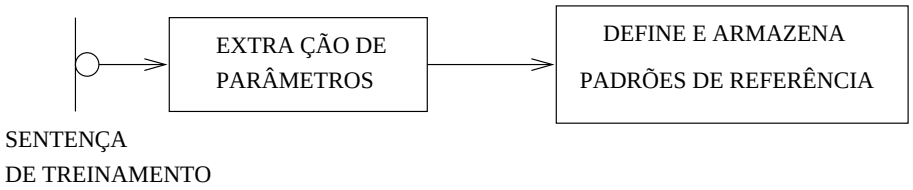


Figura 1.3: Fase de Treinamento de um SRAL.

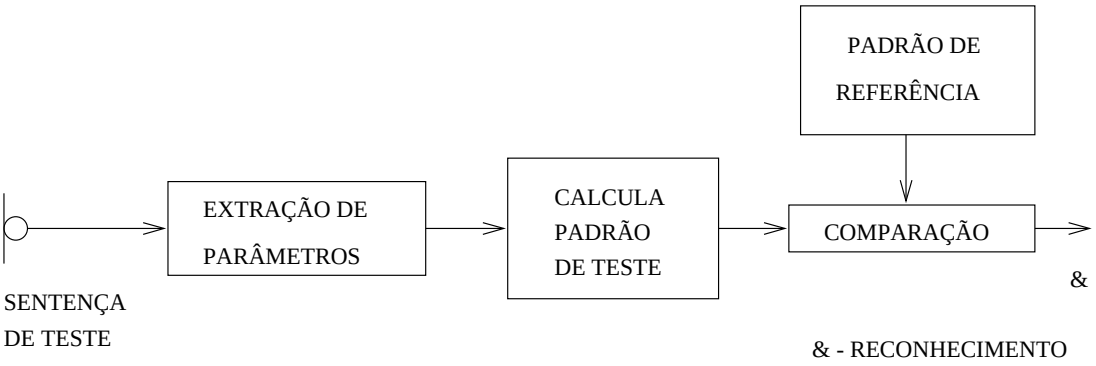


Figura 1.4: Fase de Reconhecimento de um SRAL.

O reconhecimento de locutor também pode ser dependente ou independente do texto. SRAL dependente do texto requer que o locutor pronuncie uma frase ou uma dada senha pré-determinada e o sistema independente do texto não requer a exigência do caso anterior. Na área da criminalítica, por exemplo, é de maior interesse o uso de SRAL independente do texto, uma vez que na maioria das aplicações os locutores a serem identificados são não cooperativos. Em outras situações se torna mais adequado uso do SRAL dependente do texto, a exemplo das aplicações que envolvem acesso a ambientes restritos, neste caso os locutores são cooperativos.

Alguns fatores externos podem contribuir para erros em um sistema de reconhecimento automático de locutor. A Tabela 1.1 apresenta alguns dos fatores humanos e de ambiente que contribuem para esses erros. Esses fatores geralmente são externos aos algoritmos ou são melhor corrigidos por meios que não envolvam necessariamente os

algoritmos (*e.g.*, o uso de microfones de melhor qualidade). Esses fatores são importantes e, em alguns casos, não importa o quão bom o algoritmo para reconhecimento de locutor possa ser, o erro humano (*e.g.*, o erro de leitura e às vezes de elocução) pode limitar o seu desempenho [8].

Tabela 1.1: Fontes externas de erro para um SRAL.

Erro de elocução ou de leitura das frases pré-definidas
Estado emocional
Variação da posição do microfone (intra ou inter-sessões)
Ambiente acústico pobre ou inconsistente (*e.g.*, ruído)
Erro de "casamento" do canal (*e.g.*, microfones diferentes para treinamento e teste)
Problemas de saúde (*e.g.*, resfriado que pode alterar as características do trato vocal)
Idade (*e.g.*, a forma do trato vocal pode ser alterada com a idade)

Portanto, para o projeto de um SRAL eficiente, deve-se minimizar, o máximo possível, os erros externos ao sistema e, em seguida, utilizar técnicas que possam representar, com eficiência, as características vocais que diferenciam os locutores.

1.2 Motivação

A comunicação oral é, sem dúvida alguma, a forma mais natural de comunicação humana. Em virtude da interação homem-máquina se tornar cada vez mais comum, surge uma demanda natural por sistemas capazes de reconhecer o que está sendo dito, bem como quem está falando [20]. O interesse nessa área se deve ao número de aplicações, bem como à existência de várias questões teóricas que ainda não foram respondidas [21].

Sistemas automáticos de verificação e identificação de locutor são provavelmente os métodos mais econômicos e naturais para solucionar os problemas de uso autorizado de computadores e sistemas de comunicação e controle de acesso. Com a disponibilidade das linhas telefônicas e microfones acoplados aos computadores, o custo de um sistema de reconhecimento de locutor está relacionado, basicamente, ao projeto do *software*.

Sistemas biométricos reconhecem a pessoa pelo uso de traços (feições) distintos. A voz, assim como outras características biométricas, não pode ser esquecida ou perdida,

diferentemente dos métodos de controle de acesso baseados em objetos (cartões, chaves, etc.) ou mensagens fornecidas através do teclado (senha, etc.). Além disso, os sistemas de reconhecimento de locutor, através da fala, podem ser projetados de tal forma que se tornem robustos, mesmo diante de ruído e variações do canal [19, 22], de alterações humanas (*e.g.*, resfriados) e de ambientes de gravação [8].

Com o objetivo de obter-se sistemas de reconhecimento automático de locutor eficientes, diversas técnicas têm sido utilizadas, dentre as quais destacam-se: Modelos de Markov Escondidos (HMMs - *Hidden Markov Models*) [23, 24, 25, 26], Redes Neurais Artificiais [27, 28, 29], Quantização Vetorial (VQ - *Vector Quantization*) [30, 31, 32, 33, 34], Análise por Predição Linear [35, 36] e Alinhamento Dinâmico no Tempo (DTW - *Dynamic Time Warping*) [16].

Apesar do sucesso obtido com a maioria dessas técnicas, o uso de Modelos de Markov Escondidos se torna 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, conseqüentemente, podem formar uma base teórica muito forte para uso em um grande grupo de aplicações (*e.g.*, modelagem do sinal de voz), tendo a 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, na fase de reconhecimento, em comparação com outros métodos (*e.g.*, DTW) [23, 37, 38, 39].

Entretanto, mesmo diante do sucesso alcançado com HMM, torna-se interessante investigar a utilização conjunta dessas técnicas, de forma a possibilitar o projeto de um sistema automático de reconhecimento da identidade vocal, para a língua portuguesa, capaz de modelar eficientemente as características vocais dos locutores, apresentando pequenas variações intralocutor e grandes variações interlocutor.

1.3 Objetivos do Trabalho

Tradicionalmente, os paradigmas para reconhecimento de padrões são divididos em três componentes: extração e seleção de características; escolha dos padrões e classificação. Embora essa divisão seja conveniente para o projeto do sistema, esses componentes não são independentes. Uma escolha inadequada de algum poderá comprometer, bastante, o desempenho do sistema [8]. O que não poderia ser diferente para

o reconhecimento de locutores.

Dentro desse contexto, o trabalho, aqui apresentado, trata do desenvolvimento de um sistema híbrido, que utiliza métodos paramétrico e estatístico, para o reconhecimento (identificação) automático da identidade vocal de locutores, em um grupo fechado (dependente do texto), para a língua portuguesa, que apresente, a partir da técnica utilizada, desempenho elevado.

Com o objetivo de tornar a tarefa de reconhecimento mais eficiente e rápida, o sistema é composto de dois estágios: pré-identificação e identificação.

No estágio de pré-identificação os locutores são separados em dois grupos gerais de acordo com o sexo (homens e mulheres), utilizando a freqüência fundamental. A deteção do Período Fundamental (período da freqüência fundamental), ou a estimação da freqüência fundamental de vibração das cordas vocais, torna mais rápida, portanto, a fase final da identificação, pois os locutores só serão analisados dentro dos seus respectivos subgrupos (masculino ou feminino). Tal procedimento poderá diminuir as taxas de erro do sistema (quando o locutor feminino é considerado masculino e vice-versa).

O segundo estágio, a identificação propriamente dita, é subdividido em duas etapas da seguinte forma:

Primeira etapa: a regra de decisão baseia-se em uma medida de distorção, obtida a partir da comparação do vetor de teste (vetor de parâmetros acústicos) com o conjunto de padrões de referência (vetores-código do dicionário). A construção dos padrões acústicos representativos dos locutores (padrões de referência), um padrão para cada locutor, é levada a efeito a partir da Quantização Vetorial (QV) Paramétrica. Os parâmetros são obtidos através da análise por predição linear, sendo realizada uma análise comparativa do desempenho de diversos tipos de coeficientes obtidos a partir dessa análise (coeficientes LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados), de forma a determinar qual(is) o(s) tipo(s) de coeficiente que melhor representa(m) as características vocais dos locutores. Na construção dos padrões acústicos, dicionários do QV, são avaliados três métodos: o primeiro utiliza o algoritmo LBG [40], o segundo o algoritmo KMVVT (Kohonen Modificado com Vizinhança Centrada em Torno do Vetor de Treino) e o terceiro método utiliza o algoritmo SSC (Competitivo no Espaço Sináptico), os dois últimos propostos por Vilar França et al [41, 42, 43]. O algoritmo SSC se mostrou mais adequado para o projeto do dicionários, sendo portanto o escolhido.

Segunda etapa: a regra de decisão baseia-se em uma medida de probabilidade, obtida a partir da compração do vetor de teste com o novo conjunto de padrões de referência. Estes padrões representativos dos locutores são obtidos a partir da Modelagem por Modelos de Markov Escondidos (HMMs) de Densidades Discretas (os parâmetros representativos dos locutores são transformados, a partir da QV, em um conjunto de observações discretas), um HMM associado a cada locutor do sistema.

Na tarefa de reconhecimento (identificação), são utilizadas, portanto, duas medidas para discriminação de locutores: a medida de distorção obtida a partir da quantização vetorial, seguida da probabilidade obtida do HMM. Esta última é utilizada como parâmetro de "refinamento" do processo, sendo aplicada quando a medida de distorção indicar "similaridade" entre as características vocais dos locutores.

1.4 Organização do Trabalho

Esta descrição do trabalho desenvolvido foi estruturada em sete capítulos. O presente capítulo tem por objetivo permitir ao leitor uma visão mais ampla da comunicação vocal homem-máquina, ao mesmo tempo que procura focalizar sua atenção no objeto de estudo deste trabalho. Além disso, esta seção apresenta uma breve visualização dos demais capítulos deste documento, mostrando nos parágrafos a seguir uma descrição rápida desses capítulos.

No Capítulo 2 é descrito o mecanismo de produção da voz e o seu modelo correspondente, o qual possibilitará a obtenção dos parâmetros necessários à representação dos sinais de voz, visando a realização da tarefa de reconhecimento (identificação) automático da identidade vocal de locutores.

No Capítulo 3 é realizada a descrição das técnicas analisadas para extração das características vocais representativas dos locutores.

No Capítulo 4 são apresentadas as técnicas a serem utilizadas no processo de reconhecimento (identificação) de locutor, especificando os elementos necessários à modelagem dos sinais de voz de cada locutor, para sua posterior identificação.

O Capítulo 5 faz a descrição do sistema de reconhecimento (identificação) automático da identidade vocal de locutores.

No Capítulo 6 é realizada a apresentação e análise dos resultados obtidos. Por

fim, os resultados, as conclusões e sugestões para trabalhos futuros são comentados no Capítulo 7.

O Anexo A apresenta resultados complementares aos apresentados no Capítulo 5.

No Anexo B é realizada uma descrição geral da interface projetada para o sistema de reconhecimento (identificação) automático da identidade vocal de locutores.

Capítulo 2

O Mecanismo de Produção da Voz

2.1 Introdução

Os sinais de voz são compostos de uma seqüência de sons que servem como uma representação simbólica da mensagem produzida pelo locutor para o ouvinte. A composição desses sons é governada pelas regras de linguagem. O estudo científico da linguagem e a forma como essas regras são usadas na comunicação humana é denominada *lingüística*. A ciência que estuda as características da produção do som pelo homem, especialmente para a descrição, classificação e transcrição da voz, é denominada *fonética* [1].

A voz é um sinal produzido como resultado de várias transformações que ocorrem em diferentes níveis: semântico, lingüístico, articulatório e acústico. As diferenças nessas transformações aparecem como diferenças nas propriedades acústicas do sinal de voz. Diferenças relacionadas com os locutores são um resultado da combinação das diferenças anatômicas inerentes ao trato vocal (características inerentes) e daquelas relacionadas ao movimento dinâmico do trato vocal, ou seja, a forma como a pessoa fala (características instruídas). Em reconhecimento de locutor, todas essas diferenças podem ser usadas para discriminar os locutores entre si [8].

Para gerar o som desejado, o locutor exerce uma série de controles sobre o aparelho fonador, representado na Figura 2.1, produzindo a configuração articulatória e a excitação apropriadas. A Figura 2.1 evidencia as características importantes do sistema vocal humano. O trato vocal, nome genérico dado ao conjunto de cavidades e estruturas que participam diretamente da produção sonora, 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. 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 [1].

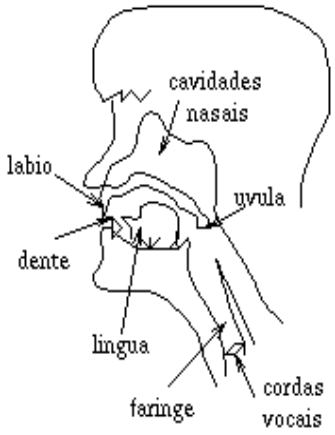


Figura 2.1: Anatomia do aparelho fonador.

Na Figura 2.2 é apresentado um modelo mecânico para a produção de voz. Nesse modelo os tratos oral e nasal são representados por tubos acusticamente acoplados.

O diagrama completo inclui o sistema subglotal composto dos pulmões, brônquios e traquéia. O sistema subglotal funciona como uma fonte de energia para produção da voz. A voz é a onda acústica radiada do sistema quando o ar é expelido dos pulmões [1].

O trato vocal e o trato nasal podem ser vistos como tubos de seção transversal não uniforme. O som se propaga através desses tubos e o espectro de freqüência é modelado pela seletividade de freqüência do tubo. Esse 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 de 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 em função das variações da forma

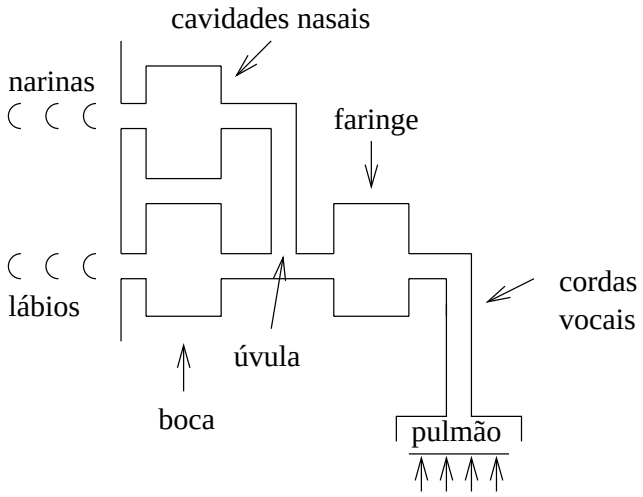


Figura 2.2: Modelo acústico do aparelho fonador.

assumida pelo trato vocal. Assim, as propriedades espectrais do sinal de voz variam com o tempo e com a forma do trato vocal [44].

Se o ouvinte decodificar de forma correta a seqüência de sons emitida, a cadeia de comunicação se completará fechando o ciclo, que compreende desde a concepção da idéia até sua completa assimilação pelo interlocutor.

Em virtude das limitações dos orgãos humanos de produção de voz e o sistema auditivo, a comunicação humana típica está limitada na faixa de 7-8 kHz [1].

Diante do exposto, faz-se necessário realizar análises acústicas, compreender as formas de excitação do aparelho fonador, bem como avaliar os parâmetros temporais do sinal de voz, de forma a tornar possível a obtenção de um modelo para a produção da voz, o qual é fundamental para a implementação de um sistema de reconhecimento automático da identidade vocal de locutores.

2.2 Análises Acústicas Elementares

As características espectrais do sinal de voz são variantes no tempo (ou não estacionárias), visto que o sistema físico varia com o tempo. Como resultado, o sinal de voz pode ser dividido em segmentos que possuem propriedades acústicas semelhantes para curtos intervalos de tempo. Inicialmente, os sinais de voz são, tipicamente, particionados dentro de duas categorias básicas: (1) *vogais* que quase não apresentam restrição

à passagem do ar através do trato vocal e (2) *consoantes* que apresentam uma maior restrição à passagem do ar e são, em geral, mais "fracas" em amplitude e podem ser semelhantes a uma fonte de ruído. Algumas das diferenças entre vogais e consoantes são evidentes visualizando a forma de onda no tempo da palavra *aplausos* pronunciada por um locutor masculino (Figura 2.3).

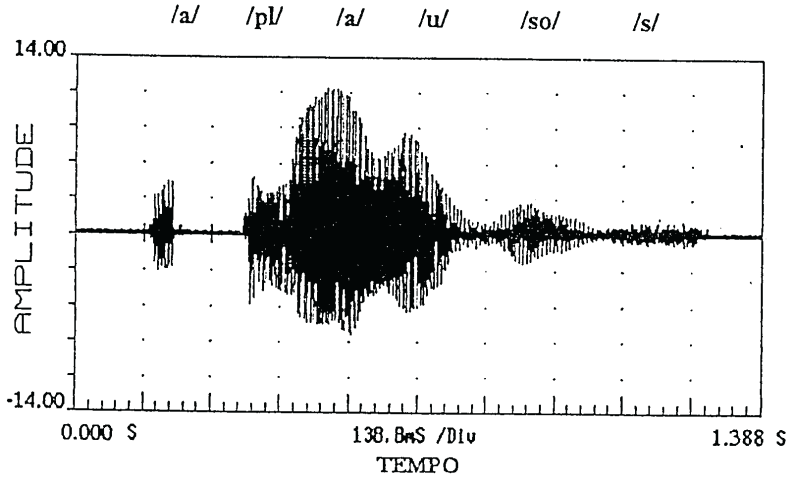


Figura 2.3: Forma de onda no tempo da palavra *aplausos*.

Para a engenharia elétrica é interessante observar as formas de onda, para verificar o que estas podem revelar sobre os aspectos acústicos e psicológicos da voz. A Figura 2.3 apresenta as características básicas do sinal de voz tais como: periodicidade, intensidade, duração, etc. Uma das mais importantes características da voz, bastante evidente na Figura 2.3, é que a voz não é constituída por sons discretos bem definidos.

As variações evidentes na forma de onda da voz são uma conseqüência direta dos movimentos do sistema articulatório da voz, o qual raramente permanece fixo por um considerável período de tempo [1].

Para o propósito da comunicação humana, é de interesse observar o sinal acústico produzido pelo locutor, com o objetivo de determinar os paralelos entre a comunicação humana e a eletrônica [1].

2.3 Formas de Excitação: Classificação dos Sons da Voz

A Figura 2.3 ilustra a forma de onda típica de um sinal de voz, que é contínua no tempo e em amplitude. Um aspecto muito importante a ser observado é que o sinal apresenta trechos que se repetem quase periodicamente e trechos basicamente aleatórios, sem nenhuma periodicidade. Assim, os sons da voz podem ser classificados em 3 classes distintas de acordo com o modo de excitação. As classes são as seguintes [1]: sons sonoros, sons surdos e sons explosivos.

2.3.1 Sons Sonoros

O fluxo de ar vindo dos pulmões é controlado pela abertura e fechamento das cordas vocais, ou dobras vocais, que são ligamentos semelhantes a dois lábios que podem ser tensionados e(ou) 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 subglótica 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. Dessa 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 [1, 45].

Quanto mais rápida a repetição, mais alta a freqüência e mais aguda é a voz, como nas vozes femininas e infantis; quanto mais lentamente essas repetições se reproduzem, mais grave é a voz, como no caso das vozes masculinas [45].

As vogais, cujo grau de nasalização é determinado pelo abaixamento da úvula, são exemplos típicos de sons sonoros. A Figura 2.4 mostra a forma de onda para a vogal /a/, na palavra **a**plausos. Algumas consoantes, como /l/ 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 (ou período de *pitch*), P_0, é dado por

$$P_0 = \frac{1}{F_0} \tag{2.1}$$

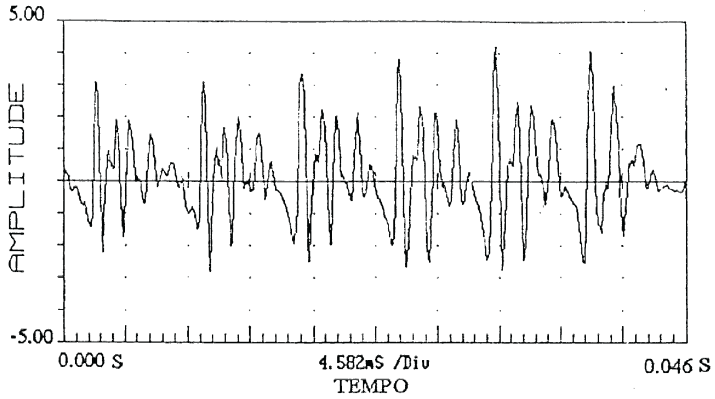


Figura 2.4: Forma de onda da vogal não nasalizada /a/ na palavra **a**plausos.

Em processamento de voz os termos *pitch* e freqüência fundamental são utilizados como sinônimos, embora o conceito de *pitch* seja mais abrangente. A rigor o *pitch* de um determinado estímulo sonoro (não necessariamente um sinal de voz), corresponde à freqüência, em Hz, de um tom senoidal que está "afinado" com o estímulo, segundo a percepção auditiva de um determinado indivíduo. Como, na percepção de voz, o *pitch* dos sons sonoros geralmente corresponde ao valor da freqüência fundamental, para as pessoas com audição normal, os dois termos passaram a ser empregados indistintamente [2].

A freqüência fundamental dos sons sonoros fica entre 80-120 Hz (para homens) e 350 Hz (para crianças), sendo 240 Hz um valor típico para mulheres [46].

2.3.2 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 um 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 fonema /s/ em aplauso**s** (Figura 2.5), 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 dessa forma tem características ruidosas com concentração relativa de energia nas mais altas componentes de freqüência do espectro de sinais de voz [1, 2].

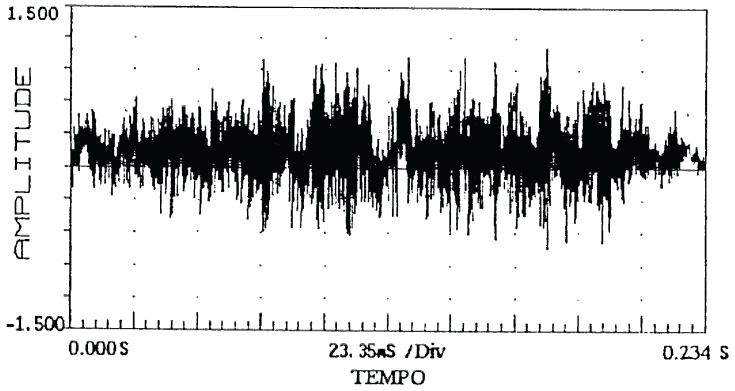


Figura 2.5: Forma de onda do fonema /s/ na palavra aplauso**s**.

2.3.3 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/, /t/, /k/, dentre outros [1, 2]. A Figura 2.6 mostra a forma de onda do explosivo /p/, em a**p**lausos.

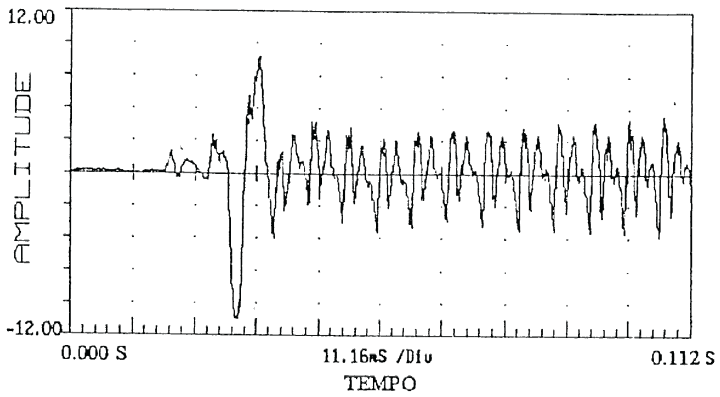


Figura 2.6: Forma de onda do fonema /p/ na palavra a**p**lausos.

2.3.4 Sons com excitação mista

Os sons fricativos sonoros, como /j/, /v/ e /z/, são produzidos combinando-se vibração das cordas vocais e excitação turbulenta. Nos períodos em que a pressão glótica atinge um 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 do ar e as ondas de pressão apresentam comportamento mais suave [1, 2]. A Figura 2.7 mostra o fonema fricativo sonoro /z/ em aplau**so**s.

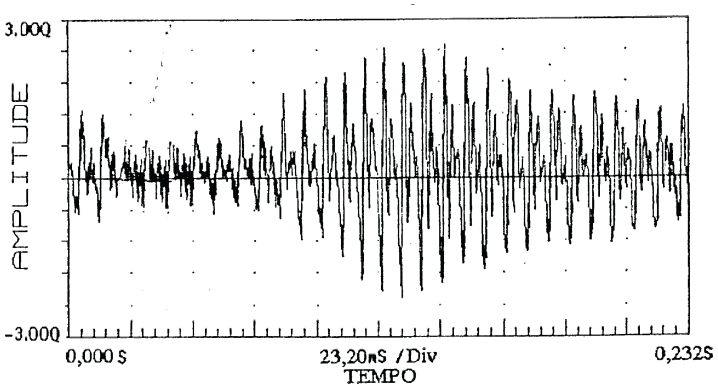


Figura 2.7: Forma de onda do fonema /z/ na palavra aplau**so**s.

Os sons oclusivos (ou explosivos) sonoros, como /d/ e /b/, 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. A Figura 2.8 mostra a forma de onda do fonema explosivo sonoro /b/ em **b**ola.

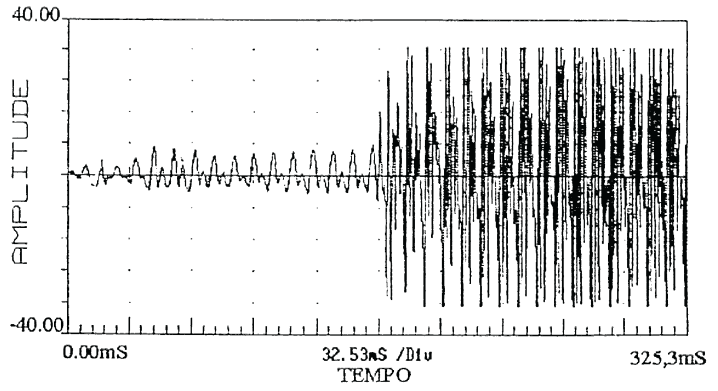


Figura 2.8: Forma de onda do fonema /b/ na palavra **b**ola.

2.4 Parâmetros Temporais do Sinal de Voz

O gráfico amplitude-*versus*-tempo de um sinal permite a avaliação de muitas características importantes que permitem uma completa descrição do mesmo. A partir do uso de parâmetros temporais torna-se possível identificar os sons básicos da fala. Dentre esses parâmetros destacam-se: a Energia do Sinal, a Taxa de Cruzamento por Zero, o Coeficiente de Correlação Normalizado, o Número Total de Picos, dentre outros.

A partir da Figura 2.3 é possível perceber uma combinação de características inerentes ao processo de produção da fala. Em alguns intervalos, o sinal apresenta níveis elevados de energia além de uma certa periodicidade e, em outros, tem a aparência de um sinal aleatório com níveis de amplitude bastante reduzidos.

Os parâmetros temporais extraídos do sinal de voz neste trabalho são: Energia do Sinal, Coeficiente de Correlação Normalizado, Taxa de Cruzamento por Zero, Número Total de Picos e a Diferença entre os Picos [1].

A energia e a taxa de cruzamento por zero são parâmetros tradicionais na análise de voz. Os outros parâmetros: número total de picos (e a diferença entre os picos) da forma de onda e o coeficiente de correlação são propostos para auxiliar a deteção de categorias de sons como fricativos surdos e fricativos sonoros, por exemplo [1].

Uma característica importante dos sinais de voz é que suas propriedades estatísticas podem ser consideradas invariantes no tempo, para curtos intervalos, até 32 ms, sendo um valor típico, 16 ms. Assim sendo, para se obter os parâmetros temporais do sinal é necessário particioná-lo em segmentos (ou blocos de amostras), visando trabalhar com o sinal dentro dos seus limites de estacionariedade [1, 20, 27, 47].

2.4.1 Energia por segmento

A energia por segmento (segmental), E_{seg}, é definida por

$$E_{seg} = N_A \cdot E\{[s(n) - \mu_{s(n)}]^2\}. \tag{2.2}$$

Para sinais ergódicos [1] e estacionários no sentido amplo [2], com média nula, como a

[1] Para um processo estocástico ergódico, as suas médias estatísticas são iguais as suas médias temporais.

[2] Um processo estocástico estacionário no sentido amplo possui uma média constante e uma função de autocorrelação que depende apenas da diferença entre os intervalos de medição.

voz, E_{seg} é definida por [1, 2]:

$$E_{seg} = N_A \cdot E\{[s(n)]^2\} = \sum_{n=0}^{N_A-1} [s(n)]^2 \quad \text{e} \tag{2.3}$$

$$E_{seg}(dB) = 10.\log[E_{seg}], \tag{2.4}$$

em que $s(n)$ é o sinal de voz, $\mu_{s(n)}$ a média de $s(n)$ e N_A o tamanho da janela (bloco de amostras do sinal) em análise. A energia é obtida, portanto, simplesmente, somando-se os quadrados das amplitudes das N_A amostras do sinal contido na janela em análise, devendo refletir as variações de amplitude do sinal de voz entre intervalos ou janelas.

A amplitude do sinal de voz varia consideravelmente com o tempo. Considerando-se que a amplitude dos segmentos surdos é muito menor que a dos segmentos sonoros, a utilização do parâmetro energia tem importância fundamental na diferenciação entre os sons surdos e sonoros.

Freqüentemente, a energia é maior nos sons surdos do que nos intervalos de silêncio mas, em alguns casos, essa afirmação não é totalmente correta. Quando o segmento em análise representa um som fricativo, sua energia pode estar muito próxima do nível de energia do ruído, único sinal existente nos intervalos de silêncio, o que pode causar erros de interpretação do sinal desejado. Neste caso, outros parâmetros temporais são utilizados para auxiliar numa tomada de decisão correta.

A energia do sinal de voz está concentrada na região de freqüências mais baixas do espectro, que compreende a faixa de 500 a 800 Hz. No entanto, mesmo contendo baixos valores de energia, as componentes de freqüências mais altas são importantes pois determinam, em grande parte, a inteligilibilidade da voz.

As freqüências abaixo de 500 Hz contribuem muito pouco para a compreensão da fala, mas têm um efeito importante na naturalidade da voz reproduzida. Em alguns sistemas de comunicações, a compreensão é fundamental e a naturalidade da voz é secundária, o que justifica o uso de larguras de faixa mais estreitas que em sistemas telefônicos ou principalmente de radiodifusão, por exemplo, nos quais a naturalidade da voz é prioritária [1, 48, 49].

2.4.2 Taxa de Cruzamento por Zero

A taxa de cruzamento por zero, TCZ, é outro parâmetro bastante utilizado em aplicações de processamento digital de sinais de voz, que utilizam métodos de análise

no domínio do tempo. Esse parâmetro indica o número de vezes que as amostras de um sinal, em um determinado segmento, cruzam o zero (limiar) tomado como referência.

Esse parâmetro é, geralmente, definido por [1, 2]:

$$TCZ = N_A \cdot E\{\text{sgn}[s(n)] - \text{sgn}[s(n-1)]\} = \sum_{n=1}^{N_A-1} |\text{sgn}[s(n)] - \text{sgn}[s(n-1)]|, \tag{2.5}$$

em que:

$$\text{sgn}[s(n)] = \begin{cases} 1 & \text{, se } s(n) \geq 0 \\ -1 & \text{, se } s(n) < 0 \end{cases} \tag{2.6}$$

Isto significa que a contagem é efetuada sempre que sucessivas amostras do sinal tiverem polaridades contrárias.

Essa medida pode ser interpretada como uma forma simples de se determinar o conteúdo de freqüência de um sinal.

Ao contrário da energia, altas taxas de cruzamento por zero caracterizam os sons surdos e taxas mais reduzidas indicam a presença de sons sonoros. Em geral, o número de cruzamentos por zero é bastante eficaz na identificação de consoantes fricativas surdas.

Pode-se atribuir, portanto, as seguintes propriedades à taxa de cruzamento, relativas ao sinal de voz [1, 2]:

- sinais com conteúdo harmônico de alta freqüência (como os sons surdos) possuem altos valores de TCZ;
- sinais com conteúdo harmônico de baixa freqüência (como os sons sonoros) possuem baixos valores de TCZ;
- é importante ter cuidado ao se analisar os sinais em função das afirmativas anteriores, pois as definições do que sejam valores altos ou valores baixos de TCZ, são um tanto imprecisas;
- em geral, os sons sonoros apresentam valores de TCZ entre 0 e 30, com média em torno de 14 ou 15, e os sons surdos encontram-se, tipicamente, na faixa de 10 a 100, com média entre 48 e 49.

2.4.3 Coeficiente de Correlação Normalizado

O Coeficiente de Correlação Normalizado é bastante útil na distinção entre os sons sonoros e surdos. Os valores do coeficiente de correlação normalizado para sons sonoros são muito próximos da unidade, pois esses sons são altamente correlacionados devido à concentração de energia, do sinal que os constitui, nas baixas freqüências do espectro. Já para os sons surdos, o valor desse parâmetro aproxima-se de zero. Os valores típicos para os intervalos de silêncio variam com o ambiente, mas encontram-se entre os valores obtidos para os sons sonoros e surdos [1].

Para sinais de voz, é comum utilizar apenas o primeiro coeficiente de correlação (ρ_1), pois este por si só fornece as informações necessárias para o auxílio na identificação dos diversos sons.

O primeiro coeficiente de correlação de uma variável aleatória, $s(n)$, é dado por [1, 2, 27, 50]

$$\rho_{s(n)s(n-1)} = \frac{c_{s(n)s(n-1)}}{\sigma_{s(n)} \cdot \sigma_{s(n-1)}}. \tag{2.7}$$

Sendo:

$$c_{s(n)s(n-1)} = E\{s(n)s(n-1)\} - \mu_{s(n)} \cdot \mu_{s(n-1)}, \tag{2.8}$$

$$\sigma^2_{s(n)} = E\{[s(n)]^2\} - \mu^2_{s(n)} \quad \text{e} \quad \sigma^2_{s(n-1)} = E\{[s(n-1)]^2\} - \mu^2_{s(n)}. \tag{2.9}$$

Como o sinal de voz, a curtos intervalos de tempo, possui média nula,

$$c_{s(n)s(n-1)} = E\{s(n)s(n-1)\} = R_{ss}(1), \tag{2.10}$$

$$\sigma^2_{s(n)} = E\{[s(n)]^2\} \quad = \quad \sigma^2_{s(n-1)} = E\{[s(n-1)]^2\} = \sigma^2_s, \tag{2.11}$$

em que $c_{s(n)s(n-1)}$ é a covariância entre $s(n)$ e $s(n-1)$, σ_s é o desvio padrão e $R_{ss}(1)$ é o primeiro coeficiente de autocorrelação.

Substituindo as Equações (2.10) e (2.11) na Equação (2.7) obtém-se, para um segmento de N_A amostras

$$\rho_1 = \frac{\sum_{n=1}^{N_A}[s(n).s(n-1)]}{\sqrt{[\sum_{n=1}^{N_A} s(n)^2].[\sum_{n=0}^{N_A-1} s(n)^2]}}. \tag{2.12}$$

2.4.4 Número Total de Picos

O número total de picos, NTP, não é um parâmetro tradicional na análise dos sinais de voz, mas é útil na identificação de sons surdos como as consoantes fricativas de pequena itensidade. Esse parâmetro mede o número de picos encontrados dentro do intervalo de voz em análise [1, 2, 51].

2.4.5 Diferença entre os Picos

Ocasionalmente, verifica-se o fato de sons fricativos sonoros poderem ser confundidos facilmente com vogais de pequena itensidade. Essa dificuldade pode ser razoavelmente contornada através da diferença entre os picos, DNP, dada por [1, 2]

$$DNP = PPOS - PNEG, \tag{2.13}$$

em que PPOS e PNEG correspondem ao número de picos positivos e negativos, respectivamente.

2.5 Modelo para Produção da Voz

Para um modelamento detalhado do processo de produção da voz os seguintes efeitos devem ser considerados [1]:

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. 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 glotais, ou sinal aleatório (ruído). Os parâmetros da fonte e sistema são escolhidos de forma a se obter na saída o sinal de voz desejado [1].

Colocando-se todos os componentes necessários, obtém-se o modelo da Figura 2.9 [1].

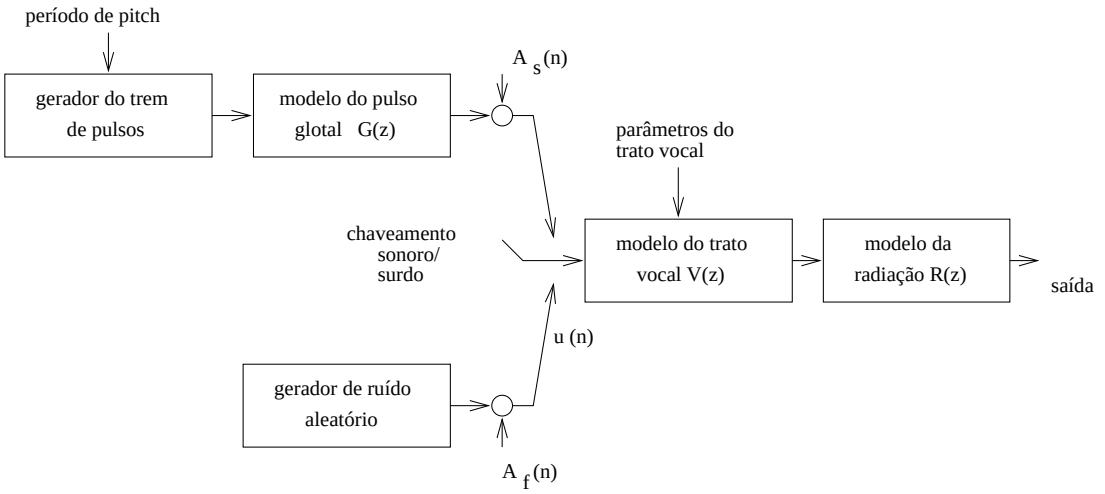


Figura 2.9: Modelo discreto da produção da fala.

Na figura acima, $u(n)$ é o sinal de excitação, $A_s(n)$ 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 grande variedade de formas. 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 da análise por predição linear é conveniente combinar o pulso glotal, radiação e componentes do trato vocal todos juntos e então representá-los com uma simples função de transferência

$$H(z) = G(z)V(z)R(z). \tag{2.14}$$

Sendo:

$G(z)$ - Transformada-z do modelo do pulso glotal;

$V(z)$ - Transformada-z do modelo do trato vocal;

$R(z)$ - Transformada-z do modelo da radiação.

Esse modelo apresenta algumas limitações [1, 2]:

1. 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. Os parâmetros do modelo são considerados constantes ao longo de curtos intervalos de tempo, tipicamente, 10 a 20 ms. A função de transferência $H(z)$, então, serve para definir a estrutura do modelo cujos parâmetros variam muito pouco com o tempo.

2. 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 [52] e pode ser aplicado quando necessário.

3. O modelo da Figura 2.9 requer que o pulso glotal seja espaçado por um múltiplo inteiro do período de amostragem, T. Witham e Steiglitz [53] têm considerado formas de eliminação desta limitação em situações requerendo controle preciso de *pitch* (período fundamental) [1].

Nenhuma das deficiências deste modelo limita seriamente a sua aplicabilidade.

2.6 Discussão

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. 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 compreensão dos fenômenos físicos associados à produção da fala é fundamental para a determinação de um modelo apropriado para representação dos sons da voz, que será de suma importância para a obtenção dos padrões acústicos representativos dos locutores.

Capítulo 3

Métodos para Extração de Parâmetros Representativos dos Locutores

3.1 Introdução

O reconhecimento de locutor é uma técnica que se baseia na identificação de certas características intrínsicas da pessoa (como a voz), diferentemente daquelas que usam artefatos para identificação (como chaves, cartões magnéticos, senhas numéricas, dentre outros). Essa distinção faz com que o reconhecimento de locutor seja, provavelmente, mais confiável. 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, controle de acesso de dados em computador, controle automático de transações telefônicas (*e.g.*, reservas de vôo ou banco por telefone), dentre outras.

Um sistema de reconhecimento automático de locutor consiste, basicamente, na extração e seleção dos parâmetros vocais, seguida do processo de classificação. O vetor de características é a interface entre essas duas fases e deve conter toda a informação relevante à fase subseqüente, de classificação, ser insensível às variações irrelevantes devido às alterações das características acústicas quando da elocução de uma sentença e ao mesmo tempo ter uma baixa dimensionalidade, visando minimizar a demanda de tempo computacional na etapa de classificação [54].

Uma forma de selecionar as características acústicas (parâmetros) para reconhecimento automático de locutor é examinar que características se correlacionam com a percepção humana de similaridade de voz. As seguintes características são úteis na discriminação das características acústicas de 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, a freqüência F_0 pode contribuir para uma estimação dessas características [2, 45].

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 e podem ser afetadas pelas condições de saúde (*e.g.*, gripes que congestionam as passagens nasais). Essas características são menos susceptíveis à imitação de impostores do que as características instruídas.

As características instruídas se referem ao movimento dinâmico do trato vocal, ou seja, a forma como o locutor fala e 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. Portanto, um SRAL eficiente deve modelar as características inerentes em detrimento às características instruídas. Características estatísticas, por exemplo, refletem mais as características inerentes do que as instruídas e são adequadas para reconhecimento de locutor, principalmente para o caso independente do texto [16].

A escolha das características únicas do locutor na análise de um sinal de voz incorre na melhoria da qualidade do reconhecimento. Os sistemas de reconhecimento de locutor têm usado diversas características [55], como formantes, intensidade, coeficientes obtidos a partir da análise por predição linear (Coeficientes LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais, Delta Cepestrais Ponderados, etc.) [35, 51, 56, 57, 58, 59, 60], entre outros. Dentre essas últimas características, segundo Reynolds [22] e outros [14, 21], uma das mais utilizadas são os coeficientes Cepestrais.

O uso da freqüência fundamental se constitui em uma grande ajuda visando o reconhecimento de locutor, pois permite a separação prévia dos locutores em grupos gerais de acordo com o sexo e a idade (homens, mulheres e crianças), facilitando assim,

a fase final do reconhecimento, pois os locutores só serão analisados dentro dos seus respectivos grupos.

O sinal de voz ocupa a faixa de freqüências compreendida entre 80 e 12000 Hz do espectro de freqüências. A freqüência fundamental da voz humana está situada entre 80 e 350 Hz, estando o valor típico para os sons produzidos pelos homens, mulheres e crianças, em torno de 120 Hz, 240 Hz e 350 Hz, respectivamente [1, 3, 45, 49]. No português brasileiro, a voz do falante apresenta sua freqüência fundamental em torno de 105 Hz para o sexo masculino e 213 Hz para o sexo feminino; as crianças, antes da puberdade, apresentam uma freqüência média de 290 Hz [45].

Pelo exposto, verifica-se que a freqüência fundamental pode ser utilizada apenas como um parâmetro de classificação de locutores em grupos gerais, sendo de grande importância quando combinada a outros parâmetros, no sentido de se obter uma classificação mais precisa de locutores dentro de um mesmo subgrupo e dessa forma, uma melhor caracterização da identidade vocal de um locutor.

A análise por predição linear é uma das mais importantes técnicas para análise de voz. Esse método tem sido a técnica predominante para estimar os parâmetros básicos da voz, que são utilizados para representação 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 computacional [2, 56, 61].

Apesar da literatura ser vasta na descrição de algoritmos [27, 47], a quantidade de trabalhos que apresentam a influência dos diversos parâmetros envolvidos num sistema de reconhecimento de locutor é relativamente pequena [21]. Portanto, se faz necessário uma avaliação criteriosa para a determinação dos parâmetros que irão representar as características vocais dos locutores.

No contexto deste trabalho, inicialmente é feita uma pré-classificação dos locutores em grupos gerais de acordo com o sexo, utilizando a freqüência fundamental. Em seguida é utilizada a técnica da análise por predição linear para obtenção dos parâmetros representativos do locutor, seguida de uma análise comparativa, de forma a determinar qual(is) parâmetro(s) melhor se adapta(m) à tarefa de reconhecimento automático da identidade vocal dos locutores. Ou seja, o objetivo é obter um conjunto de parâmetros que proporcione a separabilidade máxima de um locutor, em relação aos demais locutores do sistema.

A seguir, será descrita a forma de obtenção da freqüência fundamental e dos parâmetros representativos do locutor.

3.2 Freqüência Fundamental

A deteção do Período Fundamental, P_0, (período de pitch), ou a estimação da freqüência fundamental de vibração das cordas vocais, é um processo essencial em uma grande variedade de sistemas de processamento de voz.

A freqüência F_0 é estimada pela média dos pulsos pseudo-periódicos produzidos pelo movimento das cordas vocais, durante os períodos de análise.

Há algumas razões que tornam difícil a estimativa da freqüência fundamental, como por exemplo, o fato da forma de onda da excitação glótica não ser um trem de pulsos perfeito, a interação entre o sistema vocal e a excitação glótica, a dificuldade de estabelecer o início e o fim do período correspondente à freqüência fundamental em segmentos com voz e a dificuldade de distinguir os segmentos com baixos níveis de energia das zonas de silêncio (segmentos com ausência de voz) [16].

Em razão das dificuldades citadas, surgiram vários detetores de freqüência fundamental. No geral, o detetor seleciona os segmentos com voz e as zonas de silêncio e durante os primeiros faz a medição da freqüência fundamental. Portanto, para a estimação da freqüência fundamental, torna-se útil utilizar, inicialmente, um detetor surdo-sonoro.

Na literatura, os algoritmos de deteção do Período Fundamental são agrupados nas seguintes categorias [62]: método no Domínio do Tempo, método no Domínio da Freqüência e Métodos Híbridos. Neste trabalho foi abordado o método no Domínio do Tempo.

3.2.1 Métodos no Domínio do Tempo

A idéia básica consiste em realizar o pré-processamento do sinal de voz "quase periódico", visando reduzir a estrutura formante e, em seguida, utilizar métodos simples no domínio do tempo para fazer a estimativa da freqüência fundamental [62].

Alguns métodos práticos no domínio do tempo são: Taxa de Cruzamentos por Zero (*Zero Crossings*), Método de Medições de Picos e Vales, Método da Função de

Autocorrelação e o Método da Função da Média de Diferenças de Amplitudes (AMDF - *Average Magnitude Difference Function*) [2, 62].

Um estudo dos métodos e a comparação entre eles podem ser encontrados em [62]. Devido a sua simplicidade e eficiência reportadas em [2, 62], neste trabalho foi utilizado o Método da Função da Média de Diferenças de Amplitudes.

Método da Função da Média de Diferenças de Amplitudes (AMDF)

A AMDF considera a idéia de que se o sinal (neste trabalho o sinal de voz), $s(n)$, é periódico de período P, a seqüência $d(n)$, definida como [2]

$$d(n) = s(n) - s(n+k), \tag{3.1}$$

é zero para $k = 0, +P, -P, +2P, -2P, ...$

Tomando-se pequenos intervalos do sinal, correspondentes à voz, $d(n)$ será mínimo a intervalos múltiplos do período mas, dificilmente, será zero.

A definição da AMDF é dada pela equação

$$AMDF(k) = \frac{1}{F} \sum_{n=0}^{k_{max}-1} |s(n) - s(n+k)|, \qquad k = 0, 1, 2, ..., k_{max}. \tag{3.2}$$

sendo $AMDF(k)$ o valor da AMDF para um atraso k e F é escolhido apropriadamente. Pode-se utilizar $F = k_{max} = N_A/2$ (N_A é comprimento do quadro) e eliminar a divisão por F, por ser desnecessária. Dessa forma, a Equação (3.2) pode ser reescrita como

$$AMDF(k) = \sum_{n=0}^{\frac{N_A}{2}-1} |s(n) - s(n+k)|, \qquad k = 0, 1, 2, ..., N_A/2. \tag{3.3}$$

A função AMDF de $d(n)$, como função de k, deve ser mínima sempre que k for um múltiplo do período. Para os sons sonoros, a AMDF apresenta vales acentuados nos atrasos correspondentes ao Período Fundamental. Já para os sons surdos, estes vales não são observados.

A Figura 3.1 apresenta dois exemplos típicos da AMDF. Verifica-se que a função AMDF é mínima no período correspondente à freqüência fundamental e que não há mínimos comparáveis nos segmentos sem voz. Logo, para detetar o período da freqüência fundamental é suficiente detetar o primeiro mínimo da função AMDF.

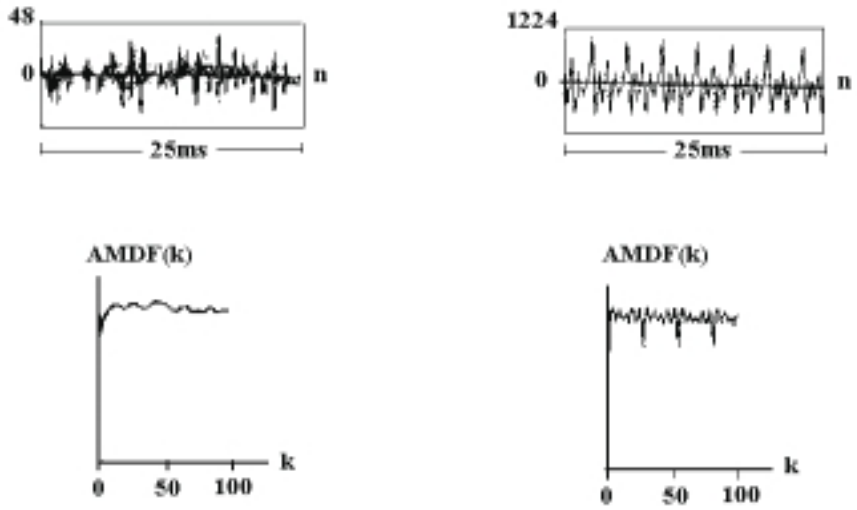


Figura 3.1: Exemplos típicos da AMDF: a) AMDF para um quadro do fricativo surdo /ch/; b) AMDF para um quadro sonoro /a/.

Esse método é de fácil implementação. Assim, a técnica da AMDF estabelece um bom compromisso entre complexidade computacional e precisão dos resultados [62].

Para deteção da freqüência Fundamental, faz-se necessário a separação dos sons sonoros da voz, pois a mesma só existirá nos intervalos da fala que contêm esses sons. Portanto, antes da estimação de F_0 é necessário a implementação do detetor surdo-sonoro.

3.2.2 Detetor Surdo-Sonoro

A qualidade de um detetor Surdo-Sonoro está ligada a sua capacidade de reconhecer a voz na presença de ruído (uma vez que nos segmentos de baixa energia do sinal, o ruído pode mascarar a existência dos sons surdos), detetando os fonemas falados [1, 51]. A Figura 3.2 mostra a configuração do detetor Surdo-Sonoro utilizado.

O detetor utilizado deve ser capaz de determinar se o segmento em análise representa uma das seguintes categorias: um som surdo, um som sonoro ou um intervalo de silêncio. O detetor utiliza parâmetros temporais (descritos no Capítulo 2) para a determinação, no segmento em análise, das categorias citadas acima.

O processo para a deteção surdo-sonoro de uma elocução, consiste em calcular, inicialmente, o valor da energia para cada segmento selecionado. Em seguida, o valor

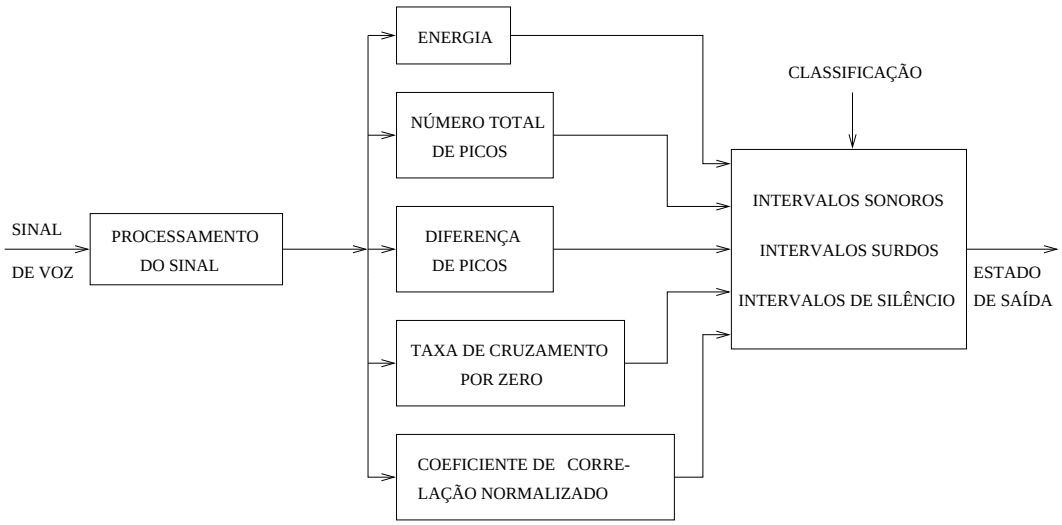


Figura 3.2: Configuração do detetor utilizado na decisão surdo-sonoro.

calculado para a energia segmental é comparado com três valores de limiares previamente estabelecidos (E_1, E_2, E_3), que delimitam quatro faixas de energia. Essas faixas são divididas de acordo com a Tabela 3.1 [2]:

Tabela 3.1: Limiares de decisão que delimitam quatro faixas de energia do detetor Surdo-Sonoro.

FAIXAS	LIMIARES
1	$E_1 > E_{seg}$
2	$E_1 \leq E_{seg} < E_2$
3	$E_2 \leq E_{seg} < E_3$
4	$E_3 \leq E_{seg}$

Cada faixa utiliza, também, testes com os demais parâmetros temporais, com a finalidade de classificar o segmento de voz em análise. Os estados de saída do detetor surdo-sonoro podem ser: som sonoro, som surdo ou silêncio.

Se a energia segmental (E_{seg}) estiver abaixo do limiar definido por E_1, o segmento de voz em análise será classificado como silêncio. Se a energia for superior ao maior valor do limiar, E_3, o segmento em análise será classificado como som sonoro.

Se a energia segmental estiver situada entre E_1 e E_3, a classificação do segmento em análise dependerá dos processamentos que serão efetuados nas faixas 2 e 3 de energia. Na faixa 2 é possível encontrar os três estados de saída para o segmento em análise.

Na faixa 3, pode-se encontrar somente sons surdos ou sonoros em virtude do alto valor de limiar de energia definido por E_2 em relação aos valores de ruído presentes.

É importante ressaltar que os valores dos limiares dos parâmetros utilizados, foram atribuídos segundo testes realizados com arquivos de voz, obtendo-se assim, faixas de valores para os parâmetros, sobre as quais são realizadas as decisões quanto à natureza do som em análise.

A função dos limiares é, portanto, atuar como chave, definindo o estado de saída do detetor, ou selecionando o processamento adequado para a tomada de decisão. Os outros parâmetros são utilizados em escalas intermediárias de importância. Pode-se iniciar a análise do algoritmo, sabendo-se que as faixas 1 ($E_{seg} < E_1$) e 4 ($E_{seg} \geq E_3$) estabelecem diretamente o tipo de sinal obtido na saída do detetor, pois o parâmetro energia é suficiente na indicação da ausência de sinal de voz (silêncio) ou da presença de um som sonoro, respectivamente.

Na faixa 2 ($E_1 \leq E_{seg} < E_2$), observa-se que um quadro com valores de TCZ (Taxa de Cruzamento por Zero) e NTP (Número Total de Picos) baixos, são classificados como som sonoro. O parâmetro DNP (Diferença entre os Picos) é utilizado quando há uma transição entre os quadros, por exemplo, de um som sonoro e silêncio, pois, os parâmetros TCZ e NTP não oferecem uma decisão segura devido aos quadros de silêncio possuírem valores de TCZ e NTP muito próximos dos valores encontrados para sons sonoros nos períodos de transição. Isso é feito devido à transição entre os estados ser um processo complexo e de difícil definição.

A transição entre os estados de silêncio e de sons surdos, apresenta maior simplicidade de identificação. Nesse caso, utiliza-se o parâmetro ρ_1 (primeiro coeficiente de correlação), pois o mesmo possibilita definir com segurança o estado real. Após a verificação de que os parâmetros TCZ e NTP estão na faixa de limiares estabelecida, verifica-se se o valor de ρ_1 está abaixo do limiar utilizado nesta faixa. Caso esteja, o som é classificado como silêncio. Para os quadros com altos valores de TCZ, o estado de saída é classificado como surdo, independentemente de outros processamentos.

Se o quadro em análise não se enquadra em nenhum dos casos anteriores, a classificação será igual a dos últimos quatro quadros, caso os mesmos sejam iguais. Se os últimos quadros não forem iguais, o quadro em questão será considerado indefinido.

Para a faixa 3 ($E_2 \leq E_{seg} < E_3$) do algoritmo, verifica-se uma simplicidade maior na tomada de decisão.

3.2.3 Estimação da Freqüência Fundamental

A Figura 3.3 apresenta o diagrama de blocos para a deteção (estimação) da Freqüência Fundamental, desenvolvido por [2] e que se constitui em uma modificação do algoritmo proposto em [62].

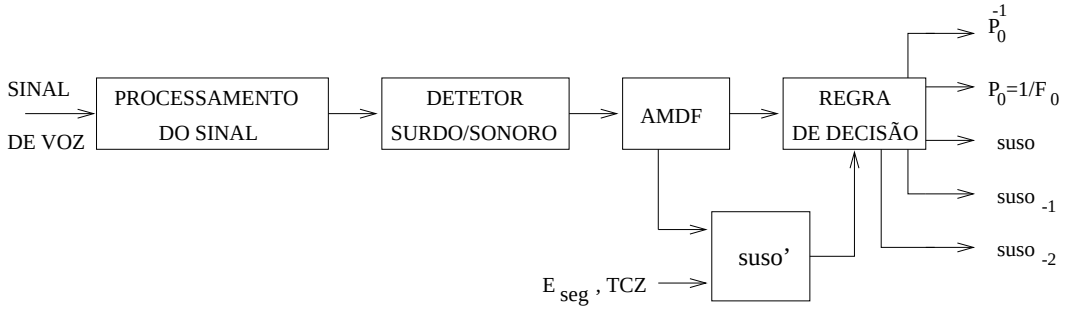


Figura 3.3: Diagrama de blocos do Detetor de Período (Freqüência) Fundamental.

Na Figura 3.3 $suso'$ é a decisão surdo-sonoro inicial para o quadro em análise, E_{seg} é a energia segmental, TCZ é o número de cruzamentos por zero, $suso_{-1}$, $suso_{-2}$ e $suso$ correspondem à decisão surdo-sonoro do último e do penúltimo quadros e do quadro atual, respectivamente e P_0, P_0^{-1} e F_0 representam o Período Fundamental do quadro atual e do quadro anterior e a Freqüência Fundamental, respectivamente.

Inicialmente é utilizado o detetor surdo-sonoro, visando separar os intervalos de silêncio, sons surdos e sons sonoros e, para os últimos, estimar a F_0. Mesmo assim, com o objetivo de minimizar os erros provenientes do detetor surdo-sonoro, no algoritmo de estimação da F_0 são detectadas, a princípio, três categorias sonoras que não necessitam ser processadas, são elas: coarticulações; silêncio e fricativos surdos, divididos em dois grupos: o primeiro grupo, dos fricativos surdos fortes, apresenta um grande número de cruzamentos por zero; o segundo grupo, dos fricativos surdos fracos, apresenta um elevado valor de número total de picos (NTP) e uma energia total (E_{seg}) abaixo do menor limiar.

Se esses segmentos não são detectados, significa que é possível que o quadro seja sonoro, procede-se a execução da extração do período da freqüência fundamental (período de Pitch).

Para o cálculo de F_0 procede-se então, de acordo com o algoritmo a seguir:

1. Inicialmente, atribui-se o valor "1" à variável $suso'$ nos quadros supostamente sonoros, isto é, que apresentam um número moderado de cruzamentos por zero ou uma alta energia.

2. Em seguida, a seqüência de voz, $s(n)$, é filtrada por um filtro passa-baixas não recursivo, com um corte em torno de 1 kHz, gerando a seqüência $s'(n)$. Essa filtragem reduz a influência das componentes de alta freqüência, dando um aspecto mais suave à AMDF da seqüência $\{s(n)\}$ que é calculada pela Equação (3.3).

3. Após o cálculo da AMDF são determinados quatro parâmetros:

   (a) max = amplitude máxima da AMDF;

   (b) min = amplitude mínima da AMDF;

   (c) $minp$ = posição do mínimo da AMDF, isto é, o provável Período Fundamental;

   (d) $naux = max/min$.

4. O algoritmo executa um entre quatro caminhos possíveis, a partir do valor da variável su, dada por:

$$su = suso' + 2suso_{-1} + 4suso_{-2} \tag{3.4}$$

   O procedimento realizado em cada caminho é descrito a seguir [2]:

   (a) **caminho 1** ($su = 0, 2$ ou 4)

   Se o vale da AMDF for pouco profundo, a decisão $suso$ será igual a 0, confirmando $suso'$. Se o vale da AMDF for profundo, $suso$ será igual a 1 e o Período Fundamental será dado por $minp$.

   (b) **caminho 2** ($su = 1$)

   Verifica-se se o vale da AMDF é profundo o bastante para confirmar a decisão $suso'$, que é igual a 1, não confirmando, o algoritmo faz $suso = 0$ e $P = 0$.

   (c) **caminho 3** ($su = 3, 5$, ou 7)

   O algoritmo faz $pitch = minp$ e $suso = 1$.

   (d) **caminho 4** ($su = 6$)

   A decisão $suso'$ é igual a zero, mas o algoritmo estende o período fundamental do quadro anterior para o quadro atual e faz $suso = suso' = 0$.

Esta estratégia procura evitar a ocorrência de um *pitch* nulo dentro de uma grande seqüência de intervalos sonoros, embora ela possa gerar um erro (aceitável) na transição dos trechos sonoros para os trechos surdos.

5. Após a determinação do Período Fundamental, é verificado, finalmente, se ele não corresponde à metade, dobro ou triplo do Período Fundamental anterior, fazendo-se as devidas correções, quando necessário.

Após ter sido calculado F_0 (ou P_0) em cada bloco de amostras, calcula-se então a Freqüência Fundamental média correspondente à elocução completa. Essa medida corresponde à média aritmética das F_0 obtidas ao longo dos segmentos sonoros.

3.3 Análise por Predição Linear

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 e de valores presentes e passados de uma entrada hipotética de um sistema cuja saída é o sinal dado. No domínio da freqüência é equivalente a modelar o sinal por um espectro de pólos e zeros [61].

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

A filosofia da predição linear está intimamente relacionada com o modelo de produção da voz (apresentado no Capítulo 2), que mostra como o sinal de voz pode ser modelado como a saída de um sistema linear variante no tempo excitado por pulsos quase periódicos (para os sons sonoros), ou ruído aleatório (para sons não sonoros). As técnicas de predição fornecem um método robusto, realizável e correto para estimação dos parâmetros que caracterizam o sistema linear variante com o tempo [61].

A forma particular do modelo digital de produção de voz que é apropriada para a utilização da predição linear está descrita na Figura 3.4.

Os métodos de predição linear estão disponíveis na literatura de engenharia há um longo tempo e têm sido vastamente empregados, principalmente em sistemas de controle, automação, telecomunicações e teoria da informação e codificação [1].

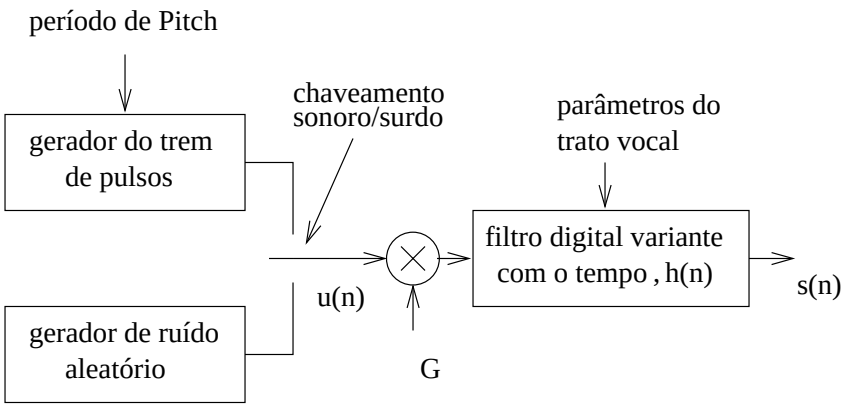


Figura 3.4: Diagrama de blocos para o modelo simplificado de produção de voz.

As técnicas de predição linear também podem ser aplicadas à quantização para reduzir a taxa de bits na representação digital do sinal de voz [1].

O princípio básico da predição linear leva a um conjunto de técnicas de análise que podem ser usadas para estimar parâmetros da fala. Esse conjunto geral de técnicas é freqüentemente denominado de Análise por Codificação Preditiva Linear ou Análise LPC (*Linear Prediction Coding*) [61].

Muitos sistemas de reconhecimento de fala e de locutor têm, tradicionalmente, utilizado os parâmetros obtidos da análise LPC, em virtude das vantagens que esses propiciam em termos de generalização da envoltória espectral, independência do *pitch* das harmônicas, e a sua habilidade para modelar, razoavelmente bem, os picos espectrais [54].

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

Para o reconhecimento de locutor, os coeficientes do preditor formam o vetor de características representativo de um dado locutor. Esses coeficientes podem ser obtidos a partir da análise LPC, denominados coeficientes LPC, ou a partir de técnicas derivadas dessa análise. Dentre os coeficientes utilizados, destacam-se: coeficientes LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados,

os quais foram utilizados neste trabalho. Esses parâmetros visam capturar informação espectral suficiente para permitir a tarefa de reconhecimento [19, 35].

A seguir serão descritas as formas de obtenção dos coeficientes citados.

3.3.1 Coeficientes LPC

Vários são os métodos conhecidos para a determinação dos coeficientes LPC. Dentre esses: o método da covariância [63]; o método da autocorrelação [64]; a formulação do filtro inverso [1]; a formulação da estimação espectral [1]; a formulação da máxima verossimilhança [1] e a formulação do produto interno [1].

Um estudo dos vários métodos e a comparação entre eles podem ser encontrados em [1]. Devido as suas características, neste trabalho foi utilizado o Método da Autocorrelação, discutido em seguida. No modelo descrito na Figura 3.4, os efeitos da radiação, trato vocal, e excitação glotal são representados por um filtro digital variante no tempo cuja função de transferência, $H(z)$, tem a seguinte forma [1]

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{K} c_k z^{-k}}, \;\; S(z) = U(z)H(z). \tag{3.5}$$

Sendo:

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

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

c_k - coeficientes LPC;

K - ordem da predição (número de coeficientes).

O sistema é excitado por um trem de pulsos para sons sonoros ou por uma seqüência de ruído aleatório para sons não sonoros. Assim, os parâmetros do modelo são: classificação sonoro/não sonoro, parâmetro de ganho (G) e os coeficientes do filtro digital (c_k). Esses parâmetros variam muito pouco em curtos intervalos de tempo [1].

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 3.4, as amostras de voz, $s(n)$, são relacionadas com a excitação, $u(n)$, pela equação diferença [1]

$$s(n) = \sum_{k=1}^{K} c_k s(n-k) + Gu(n). \tag{3.6}$$

Uma predição linear com coeficientes de predição, c_k, é definida como um sistema cuja saída é

$$\widetilde{s}(n) = \sum_{k=1}^{K} c_k s(n-k). \tag{3.7}$$

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

$$e(n) = s(n) - \widetilde{s}(n) = s(n) - \sum_{k=1}^{K} c_k s(n-k). \tag{3.8}$$

Para formular o problema, inicialmente é selecionado um segmento do sinal de voz por intermédio de uma janela de comprimento finito e igual a N_A (Figura 3.5). A melhor escolha do valor de N_A permite uma boa aproximação às hipóteses de ergodicidade e estacionariedade no sentido amplo [1]. Em virtude da inércia dos articuladores, é intuitivo que o sinal de voz possa ser considerado estacionário em intervalos apropriados, de curta duração.

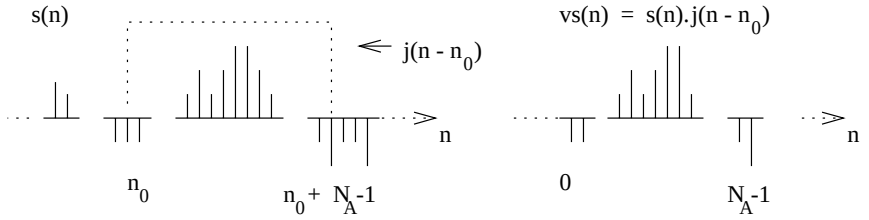


Figura 3.5: Exemplo de um segmento de voz selecionado a partir da seqüência $s(n)$ por meio de uma janela retangular, $j(n)$.

Nas próximas equações, $vs(n)$ 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 (3.6), modificada pela janela, será escrita no domínio do tempo como [2]

$$vs(n) = Gu(n) + \sum_{k=1}^{K} c_k vs(n-k), \tag{3.9}$$

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 K o número de amostras passadas utilizadas na combinação linear, pode-se formalizar a aproximação da amostra genérica $vs(n)$ pela relação [2]

$$\widetilde{v}s(n) = \sum_{k=1}^{K} c_k vs(n-k), \tag{3.10}$$

dado que $\widetilde{v}s(n)$ é a aproximação de $vs(n)$ e c_k é o k-ésimo coeficiente da combinação linear; $\widetilde{v}s(n)$ é normalmente denominada a estimativa ou predição de ordem K da amostra $vs(n)$.

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

$$e(n) = vs(n) - \widetilde{v}s(n) = vs(n) - \sum_{k=1}^{K} c_k vs(n-k), \tag{3.11}$$

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

$$Erro(n) = \sum_{n=-\infty}^{\infty} e(n)^2. \tag{3.12}$$

Como o segmento de voz é nulo para $n < 0$ e para $n > N_A$, o erro de predição (Equação (3.12)) é, portanto, nulo para $n < 0$ e $n > N_A + K - 1$. A partir dessa consideração, e substituindo a Equação (3.11) na Equação (3.12), obtém-se

$$Erro(n) = \sum_{n=0}^{N_A+K-1} [vs(n) - \sum_{k=1}^{K} c_k vs(n-k)]^2. \tag{3.13}$$

O conjunto de coeficientes c_k que minimiza $Erro(n)$ é obtido a partir de

$$\frac{\partial[Erro(n)]}{\partial[c_k]} = 0, \ 1 \leq k \leq K. \tag{3.14}$$

Com a substituição da Equação (3.13) em (3.14) e a realização das K derivadas parciais, chega-se ao seguinte sistema de equações lineares:

$$\sum_{k=1}^{K} c_k R_r(|i-k|) = R_r(i), \ 1 \leq i \leq K, \tag{3.15}$$

com

$$R_r(k) = \sum_{n=0}^{N_A-K-1} vs(n)vs(n+k), \tag{3.16}$$

denominada função de autocorrelação a curto prazo. As Equações (3.15) e (3.16), conhecidas como Equação de Wiener-Hopf, podem ser vistas mais facilmente se colocadas da seguinte forma (forma matricial) [2, 65]:

$$\begin{vmatrix} R_r(0) & R_r(1) & ... & R_r(K-1) \\ R_r(1) & R_r(0) & ... & R_r(K-2) \\ R_r(2) & R_r(1) & ... & R_r(K-3) \\ ... & ... & ... & ... \\ R_r(K-1) & R_r(K-2) & ... & R_r(0) \end{vmatrix} \begin{vmatrix} c_1 \\ c_2 \\ c_3 \\ ... \\ c_K \end{vmatrix} = \begin{vmatrix} R_r(1) \\ R_r(2) \\ R_r(3) \\ .. \\ R_r(K) \end{vmatrix} \tag{3.17}$$

Os coeficientes c_k do preditor são determinados a partir da solução das Equações (3.15) e (3.16) (ou Equação (3.17)) e são os coeficientes c_k do filtro $H(z)$ da Figura 3.4.

Utilizando a simetria da matriz de autocorrelação, pode-se utilizar algoritmos recursivos bastante eficientes para solução do sistema, a exemplo do algoritmo de Levinson-Durbin [2, 51] largamente utilizado (aplicado neste trabalho).

Após a estimação dos coeficientes do polinômio, falta determinar o ganho, G, expresso da seguinte forma [1]

$$G = [R_r(0) - \sum_{k=1}^{K} c_k R_r(k)]^{1/2}, \tag{3.18}$$

sendo $R_r(k)$ 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.

3.3.2 Coeficientes Cepestrais

Os coeficientes Cepestrais são usados para descrever a envoltória espectral do sinal de voz a curtos intervalos de tempo. O cepstrum é a transformada inversa de Fourier do logaritmo do espectro do sinal a curtos intervalos de tempo. Através da operação logaritmo, a função de transferência do trato vocal e da fonte de voz são separadas [57].

Uma das principais vantagens dos coeficientes Cepestrais reside no fato de que estes são geralmente descorrelacionados e isso gera covariâncias diagonais a serem utilizadas nos HMMs [66].

Existem duas formas de obtenção dos coeficientes Cepestrais [57]: coeficientes Cepestrais FFT e coeficientes Cepestrais LPC.

Na análise cepestral FFT é aplicada, diretamente ao sinal de voz, uma transformada inversa rápida de Fourier. O i-ésimo cepstrum, $ce_i(n)$, é calculado por[57]

$$ce_i(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log_{10} \left| X_i(e^{jw}) \right| e^{jwn} \, dw \quad , \tag{3.19}$$

em que $-\infty < n < \infty$ e X_i representa o i-ésimo bloco do espectro de potência do sinal de voz a curtos intervalos de tempo.

Na análise cepestral LPC, a transformada z é aplicada no sinal de voz modelado pela análise LPC. Os coeficientes Cepestrais, do espectro obtido da análise LPC, podem ser calculados recursivamente, a partir dos coeficientes LPC, c_i, por [19, 57, 67]

$$\begin{aligned} & ce_i(1) = c_i(1), \\ & ce_i(n) = c_i(n) + \sum_{j=1}^{n-1} (1 - \frac{j}{n}) c_i(j) ce_i(n-j) \quad , 1 < n \leq K, \end{aligned} \tag{3.20}$$

em que n é o índice do coeficiente e i o índice do bloco de amostras.

O uso desta relação recursiva leva a uma computação eficiente dos coeficientes Cepestrais, $ce_i(n)$, e evita a fatoração polinomial. Uma vez que $ce_i(n)$ tem duração infinita, o vetor de características, de dimensão K, é constituído das componentes $ce_i(1)$ a $ce_i(K)$, as quais são as mais significativas devido ao "decaimento" da seqüência com o aumento de n. Mesmo com esse truncamento, o erro médio quadrático entre dois vetores de coeficientes Cepestrais é aproximadamente igual ao erro médio quadrático entre os logaritmos do espectro dos correspondentes filtros LPC. Dessa forma, tem-se uma boa medida da diferença na envoltória espectral dos blocos de amostras a partir dos quais os coeficientes Cepestrais foram obtidos [19].

O método LPC vem sendo comumente utilizado para definir as características do locutor, pois modela o trato vocal, o qual é a peça chave para distinguir um locutor dos outros. Por outro lado, o FFT modela a forma de onda, necessitando de outras técnicas [55] para auxiliar a extração das características do locutor. Portanto, o método LPC foi utilizado neste trabalho.

3.3.3 Coeficientes Cepestrais Ponderados

A idéia básica dos coeficientes Cepestrais Ponderados refere-se à capacidade de minimizar a sensibilidade dos coeficientes Cepestrais de baixa ordem em relação à envoltória espectral e à sensibilidade dos coeficientes Cepestrais de alta ordem em relação ao ruído [19].

A ponderação é obtida multiplicando-se $ce_i(n)$ por uma janela $jp(n)$ (a escolha correta de $jp(n)$ melhora a robustez), obtendo-se assim, o cepstrum ponderado como um vetor de características. A operação de ponderação é também conhecida como filtragem ou suavização (*liftering*) [19].

Existem algumas técnicas de ponderação que diferem de acordo com o tipo de janela cepestral, $jp(n)$, usada. A janela mais simples é a janela retangular, dada por [19]

$$jp(n) = \begin{cases} 1 & , \quad n = 1, 2, ..., K \\ 0 & , \text{ caso contrário} \end{cases} \tag{3.21}$$

sendo K o tamanho da janela. As primeiras K amostras, que são as mais significativas, em virtude da propriedade do decaimento, são mantidas.

Outras janelas incluem a ponderação linear (*quefrency liftering*), na qual [19]

$$jp(n) = \begin{cases} n & , \quad n = 1, 2, ..., K \\ 0 & , \text{ caso contrário} \end{cases} \tag{3.22}$$

e a filtragem (ou suavização) passa-faixa (BPL - *bandpass liftering*), em que [19]

$$jp(n) = \begin{cases} 1 + \frac{K}{2} sen(\frac{n\pi}{K}) & , \quad n = 1, 2, ..., K \\ 0 & , \text{ caso contrário} \end{cases} \tag{3.23}$$

Ponderando os coeficientes Cepestrais por uma das janelas citadas, um conjunto de coeficientes Cepestrais Ponderados, $cp_i(n)$, é obtido a partir da expressão [19, 67]

$$cp_i(n) = ce_i(n) \cdot jp(n). \tag{3.24}$$

A ponderação linear ajusta cada componente cepestral individualmente pelo índice n, suavizando as componentes de ordem inferior. A BPL pondera uma seqüência de coeficientes Cepestrais por uma função senoidal deslocada, de forma que as componentes de baixa e de alta ordem são *de-enfatizadas*. Portanto, essa foi a janela utilizada

neste trabalho. Os esquemas de ponderação descritos são baseados na idéia de que os pesos são apenas função do índice do coeficiente cepestral e não tem nenhuma relação explícita com as variações instantâneas dos coeficientes Cepestrais, que são introduzidas pelas condições ambientais (*e.g.*, ruído e efeitos do canal) [19].

3.3.4 Coeficientes Delta Cepestrais

Os coeficientes Cepestrais representam as propriedades espectrais de um dado bloco de amostras de voz. Entretanto, estes não caracterizam a informação temporal ou de transição de uma seqüência de blocos de amostras de voz. Para aplicações relacionadas ao texto, como por exemplo, reconhecimento de voz dependente do texto, um aumento do desempenho tem sido obtido com a introdução da derivada cepestral no espaço de características, porque a derivada cepestral captura a informação de transição da voz. A primeira derivada do cepstrum (também conhecida como Delta Cepstrum), $\Delta ce_i(n)$, é definida como [19]

$$\frac{\Delta ce(n,t)}{\Delta t} = \Delta ce_i(n) \approx \phi \sum_{q=-Q}^{Q} qce(n, t+q), \tag{3.25}$$

em que $ce(n,t)$ é o n-ésimo coeficiente Cepestral no tempo t, ϕ é uma constante de normalização, $2Q+1$ é número de blocos de amostras sobre os quais o cálculo é realizado.

Os coeficientes Delta Cepestrais também podem ser obtidos a partir de uma versão simplificada da Equação (3.25), da forma [19, 67]

$$\Delta ce_i(n) = [\sum_{q=-Q}^{Q} qce_{i-q}(n)]G \quad , 1 \leq n \leq K, \tag{3.26}$$

sendo G o termo de ganho (=0,375), K o número de coeficientes Delta Cepestrais, $Q = 2$, n o índice do coeficiente e i o índice do bloco de amostras.

Neste trabalho foi utilizada a Equação (3.26) para obtenção dos coeficientes Delta Cepestrais.

3.3.5 Coeficientes Delta Cepestrais Ponderados

Substituindo os coeficientes Delta Cepestrais (Equação (3.26)) na Equação (3.24), obtém-se os coeficientes Delta Cepestrais Ponderados, que associam as características

dos coeficientes Cepestrais Ponderados e Delta Cepestrais, da seguinte forma [67]

$$\Delta cp_i(n) = \Delta ce_i(n) \cdot jp(n). \tag{3.27}$$

3.4 Discussão

O processo de reconhecimento (identificação) de identidade vocal consiste na extração de parâmetros de voz (características) de um locutor de forma a definir um modelo que preserve as suas características vocais que o diferenciam de outros indivíduos. 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. Dentre essas características, destacam-se: F_0 (freqüência fundamental) e os coeficientes obtidos a partir da análise por predição linear.

A freqüência fundamental, usualmente, funciona como uma característica para classificar locutores de forma preliminar dentre grupos gerais (quanto ao sexo e idade: homens, mulheres e crianças). Sendo útil, portanto, para a redução do número de locutores a ser analisado no processo de identificação.

O método da análise por predição linear (LPC) tem sido a técnica predominante para estimar os parâmetros básicos da voz. A importância desse método reside tanto na habilidade de fornecer estimativas extremamente corretas dos parâmetros da voz, capazes de diferenciar locutores em um grupo, quanto na relativa velocidade computacional. Neste trabalho, é levada a efeito uma análise comparativa de desempenho de alguns coeficientes obtidos através da análise LPC (coeficientes LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados), visando determinar qual(is) melhor representa(m) as características vocais dos locutores para tarefa de reconhecimento (identificação) automático da identidade vocal desses locutores.

Capítulo 4

Métodos para o Reconhecimento Automático de Locutor

4.1 Introdução

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 automático de locutor 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 locutores a serem testados. Uma memória de vetor de padrões é estabelecida durante o treinamento, quando cada locutor pronuncia um vocabulário, e os segmentos acústicos são convertidos em características representativas de cada locutor. 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 àquela referência. Contudo, se o "casamento" é relativamente pobre ou se outras referências fornecem casamento similar, outro procedimento de decisão pode ser adotado

(*e.g.*, uma decisão pendente pode ser adiada e ao locutor é solicitado que repita seu padrão) [3].

Todas as tarefas de reconhecimento de padrões, incluindo RAL, utilizam duas fases: TREINAMENTO e RECONHECIMENTO. Realizada *off-line*, a fase de treinamento estabelece uma memória de referência ou padrões de referência (de voz), aos quais são atribuídos rótulos. Na fase do reconhecimento automático 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 [68].

Nos métodos paramétricos, após a deteção 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 por meio de modelos estatísticos, tais como Modelos de Markov Escondidos (HMMs) [8, 23, 24, 69]. 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 tomada usando o 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 [16]. Esse método tem possibilitado bons resultados. Apesar do sucesso, métodos alternativos de reconhecimento têm sido estudados, devido principalmente aos seguintes fatores [23]: o alto custo computacional do método usando programação dinâmica e as dificuldades de estender o método para problemas mais difíceis, como por exemplo, o reconhecimento de locutor em 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 ou o uso da quantização vetorial para eliminar o processamento da própria programação dinâmica [19, 30, 33, 70]. 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, ainda existem problemas relacionados à redução das 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 (visto que baseia-se apenas no cálculo de uma medida de probabilidade) e por basear-se em modelos estocásticos do sinal de voz sendo capaz de modelar vários eventos, tais como fonemas, sílabas, etc. [70], o que o torna bastante flexível.

O sistema de reconhecimento (identificação) automático da identidade vocal, proposto neste trabalho, como dito anteriormente, se constitui em um sistema híbrido, que utiliza tanto o método paramétrico quanto o estatístico, para a realização das tarefas de treinamento e reconhecimento (identificação), visando a obtenção de um sistema eficiente. Para a tarefa de treinamento, após a extração e escolha dos parâmetros que melhor representam um dado locutor é realizada a quantização vetorial paramétrica, para obtenção dos símbolos representativos dos locutores (dicionários), um para cada locutor. Para o projeto do dicionário do quantizador vetorial são avaliados três métodos: o primeiro utilizando o algoritmo LBG [40], o segundo utilizando o algoritmo KMVVT (Kohonen Modificado com Vizinhança Centrada em Torno do Vetor de Treino) e o terceiro utilizando o algoritmo SSC (Competitivo no Espaço Sináptico), os dois últimos propostos por Vilar França et al [41, 42, 43]. Em seguida, são construídos os Modelos de Markov Escondidos (HMMs) de Densidades Discretas, sendo associado um HMM a cada locutor. Na tarefa de reconhecimento (identificação), são utilizados dois parâmetros para discriminação de locutores: a medida de distorção obtida a partir da quantização vetorial, seguida da probabilidade obtida do HMM. Esse último é utilizado como parâmetro de "refinamento" do processo de identificação.

A seguir, serão feitas as descrições das técnicas paramétricas e estatísticas utilizadas: Quantização Vetorial, Redes Neurais e Modelos de Markov Escondidos.

4.2 Quantização Vetorial

A maior parte das formas de comunicação atuais utiliza a transmissão digital como um meio dominante para comunicação de 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 necessários à transmissão dos sinais, de forma a manter a inteligibilidade e a qualidade em valores aceitáveis. Na engenharia elétrica, o campo que trata desse problema é chamado compressão ou codificação de dados, que aplicado à voz é conhecido por codificação ou compressão de voz. Uma técnica de compressão bastante utilizada na área de processamento digital de sinais de voz é a quantização. Nesta técnica 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 [71].

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.

A Quantização Vetorial é uma técnica de codificação usada, tipicamente, para transmissão a baixa taxa de bits. 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, sendo a sua principal vantagem a produção do dicionário para determinação da similaridade entre elocuções de um mesmo locutor e divergências entre locutores [72].

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

O conjunto W é referido como dicionário de reconstrução, M é o tamanho do dicionário, e $\vec{w}_i$ 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 [30].

A seqüência de vetores de reprodução $\vec{\hat{x}} = q(\vec{x})$ 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

$$\frac{log_2 M}{K}. \tag{4.1}$$

A desvantagem da QV está no aumento da complexidade na 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 (*e.g.*, Medida de Distorção do Erro Médio Quadrático) é usada como um critério de decisão para o projeto e operação do dicionário.

O ponto em questão na implementação do quantizador vetorial 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 necessários para que o dicionário 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, de forma 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 Euclideana Simples).

Aumentando o tamanho do dicionário cresce o tempo de computação, mas decresce a probabilidade de erro pela redução dos desvios padrões das distorções.

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 [3].

Obtendo-se, para um dado sinal de voz, o conjunto de vetores de coeficientes, $\vec{c}_t$ ($t = 1, 2, ..., T$), de dimensão K, a idéia principal do projeto do dicionário é determinar um conjunto ótimo de vetores que represente os vetores de coeficientes, $\vec{\hat{c}}_m$, $m = 1, 2, ..., M$, tal que para um dado M, a distorção obtida pela substituição do conjunto de vetores de treinamento, $\vec{c}_t$, pelos vetores do dicionário, seja mínima [24].

Dito de uma maneira mais formal, define-se $d(\vec{\hat{c}}_m, \vec{c}_t)$ como a distância entre dois vetores, $\vec{\hat{c}}_m$ e $\vec{c}_t$. Assim, o objetivo do projeto do dicionário é encontrar o conjunto, $\vec{\hat{c}}_m$, tal que [24]

$$||D_M|| = \{\frac{1}{T}\sum_{t=1}^{T} \min_{1\leq m\leq M}[d(\vec{\hat{c}}_m, \vec{c}_t)]\}, \tag{4.2}$$

seja satisfeita. O valor $||D_M||$ é a medida de distorção do quantizador vetorial [23].

O procedimento da quantização vetorial é descrito a seguir [24, 33]:

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.

Em RAL um dicionário é usualmente projetado para cada combinação de locutor e sentença, baseado em uma ou mais elocuções da sentença. 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 é comparada com um limiar).

O Reconhecimento de Locutor via QV pode produzir alta precisão para casos dependente e independente do texto, com elocuções de teste relativamente curtas [3].

4.2.1 Projeto do dicionário

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

$$q(x) = \vec{w}_i, \quad se \quad \vec{x} \in C_i. \tag{4.3}$$

A Figura 4.1 mostra um exemplo de um particionamento do espaço bi-dimensional ($K = 2$) para o propósito da quantização vetorial, a região limitada pelas linhas mais

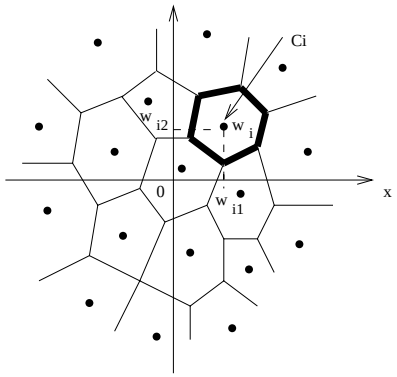


Figura 4.1: Partição do espaço bi-dimensional ($K = 2$).

fortes é a célula C_i. As posições dos vetores-códigos, correspondentes as outras células, são mostradas pelos pontos [30, 40].

Para $K = 1$ (uma dimensão), a quantização vetorial se reduz a quantização escalar. A Figura 4.2 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 4.2 é $M = 10$.

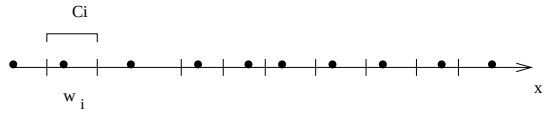


Figura 4.2: Particionamento da linha real em 10 células ou intervalos para quantização escalar ($K = 1$).

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 [30].

Existem várias formas de se escolher o alfabeto de reprodução inicial $\hat{A}_0$ exigido pelo algoritmo do quantizador vetorial. Um dos métodos utilizados nas distribuições amostrais é o método *K-means*, pela escolha dos primeiros M vetores da seqüência de treinamento [40, 71]. Esse foi o método utilizado neste trabalho.

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

Os métodos práticos conhecidos para projeto de quantizadores vetoriais para o caso mais geral, usam alguma técnica de aglomeração: uma seqüência de treino típica da fonte vetorial a ser quantizada é observada e algum algoritmo de aglomeração é usado para gerar o dicionário. Um desses algoritmos, conhecido na área de processamento de voz como Algoritmo de Linde-Buzo-Gray (LBG), também conhecido como GLA (*Generalized Lloyd Algorithm*) ou algoritmo das K-médias, tem sido bastante utilizado na produção de dicionários [30, 40, 51]. O algoritmo 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 [71].

O algoritmo LBG consiste da seguinte seqüência de passos [40]:

- *Passo 1*) condição inicial: inicialize com qualquer configuração inicial desejada, estabeleça um valor para o limiar de distorção e um valor elevado para a distorcão;
- *Passo 2*) particionamento: aloque cada dado (ou vetor de entrada) na respectiva classe segundo o critério do vetor-código mais próximo;
- *Passo 3*) calcule o valor da redução percentual da distorção. Se este valor é superior ao limiar estabelecido, vá para o *passo 4*); caso contrário finalize o processo;
- *Passo 4*) atualização do dicionário: compute os novos vetores-código como os centróides das classes de dados. Retorne para o *passo 2*).

Em essência, no algoritmo LBG a função distorção decresce monotonicamente, uma vez que o dicionário é iterativamente atualizado visando satisfazer as condições de centróide e de vizinho mais próximo. Infelizmente, como a função distorção é geralmente não convexa e pode conter múltiplos mínimos locais [74], o algoritmo LBG freqüentemente produz dicionários que não são ótimos.

Embora o algoritmo LBG seja o mais utilizado no projeto de dicionários para quantização vetorial, este apresenta alguns problemas e limitações comumente reportados [75, 76, 77, 78, 79, 80], como por exemplo: a velocidade de convergência e o desempenho do dicionário final dependem do dicionário inicial; alguns vetores-código podem ser subutilizados e, em casos extremos, até mesmo nunca serem acessados, ou seja, o algoritmo pode resultar em células de Voronoi vazias [73]. Outras abordagens têm sido utilizadas para projeto de quantizadores vetoriais, como por exemplo: rede neural de Kohonen [81, 82, 83] e outros algoritmos competitivos de redes neurais [28, 84, 85].

4.2.2 Medidas de Distorção

A medida de distorção é uma função de atribuição de um valor não negativo para o par entrada/saída de um sistema. A distorção entre o sinal original ou entrada e o sinal de reprodução ou saída indica o custo da representação do sinal original por um sinal quantizado [51].

Pesquisas indicam que uma redução na distorção de poucos decibéis é muito perceptível pelo ouvido humano em determinadas situações. Idealmente, para uma função de distorção ser utilizada em sistemas de processamento de voz, ela deve possuir algumas características fundamentais tais como: significância subjetiva, ou seja pequenos valores na distorção devem resultar numa boa qualidade de voz, assim como grandes valores devem indicar uma péssima qualidade do sinal; ser analiticamente tratável, de modo que possa ser analisada através de métodos matemáticos convencionais e não muito complexos; tratabilidade computacional, no sentido da função de distorção poder ser eficientemente calculada e aplicada em sistemas operando em tempo real [40, 51, 71].

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.

Assume-se que a distorção causada pela reprodução de um vetor de entrada $\vec{x}$ por um vetor de reprodução $\vec{\widehat{x}}$ é dada por uma medida de distorção não-negativa $d(\vec{x}, \vec{\widehat{x}})$. Muitas medidas são propostas, tais como: Medida de Distorção do Erro Médio Quadrático [71], Erro Médio Quadrático Ponderado, Medida de Distorção de Itakura-Saito [40], dentre outras. Neste trabalho foi utilizada a Medida de Distorção do Erro Médio Quadrático. É a medida mais simples e comum, por sua simplicidade e tratamento matemático. Os espaços de reprodução e de entrada são espaços Euclideanos K-dimensionais e $d(\vec{x}, \vec{\widehat{x}})$ é denominada distorção do erro médio quadrático, em que [71]:

$$d(\vec{x}, \vec{\widehat{x}}) = \frac{1}{K} \sum_{i=1}^{K} |x_i - \widehat{x}_i|^2. \tag{4.4}$$

4.3 Redes Neurais Artificiais

As células principais da estrutura do cérebro são os neurônios. Há diversos tipos, cada tipo cumprindo tarefas distintas. Pelo menos, duas características comuns são encontradas em todos os tipos de neurônios: são excitáveis e executam processos de comunicação. Essas duas características conferem aos neurônios a habilidade de processar informação [82].

A Figura 4.3 apresenta a estrutura básica de um neurônio. Muitas fibras nervosas chamadas de dendritos são conectadas ao corpo do neurônio ou soma, que funciona como elemento processador. Uma outra fibra nervosa divergente única, podendo ser ramificada, estende-se, a partir do corpo do neurônio. Essa fibra, denominada de axônio, conecta o neurônio a outros dendritos ou somas, através das junções sinápticas [82, 85, 86].

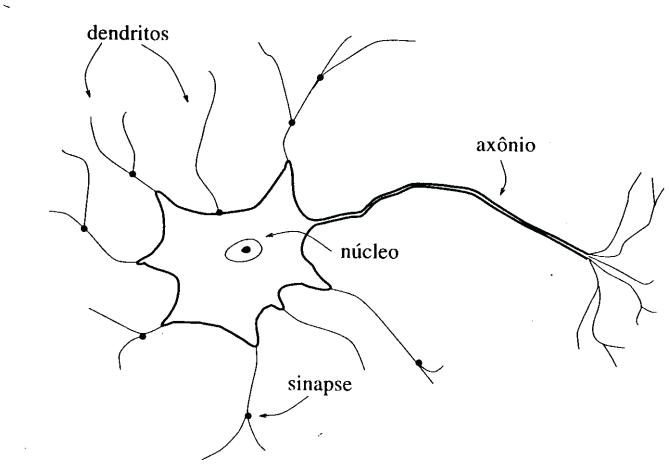


Figura 4.3: Estrutura básica de um neurônio.

Os métodos de processamento da informação usados pelo cérebro são efetuados por cerca de cem bilhões de neurônios, conectados, cada um, segundo estruturas complexas, a milhares de outros neurônios [87].

A partir do momento em que as máquinas começaram a evoluir, um grande desejo do homem tem sido a criação de uma máquina que possa operar independentemente do controle humano. Uma máquina cuja independência seja desenvolvida de acordo com seu próprio aprendizado e que tenha capacidade de interagir com ambientes incertos

(desconhecidos por ela), uma máquina que possa ser chamada de autônoma, inteligente ou cognitiva [82].

Para criar sistemas artificiais que sejam capazes de efetuar tarefas inteligentes, são necessários modelos simplificados de neurônios e de Redes Neurais, além de métodos de aprendizado. Redes Neurais naturais estão continuamente aprendendo por meio de adaptação com o meio ambiente. Acredita-se que esse aprendizado ocorra por meio de modificações adequadas nas junções sinápticas.

Uma rede neural artificial é um sistema de processamento de informação adaptativo, geralmente não-linear, cujo comportamento pode ser ajustado de acordo com o ambiente de uso. Ou seja, uma rede neural é capaz de aprender [82, 83, 85, 88].

Devido as suas características, a aplicação das Redes Neurais no projeto dos dicionários para a quantização vetorial vem se tornando cada vez mais popular nos sistemas de compressão de voz/imagem e reconhecimento de voz e locutor [55, 84]. A importância da utilização de redes neurais em quantização vetorial reside no fato de que o método tradicional, LBG, apresenta algumas limitações, citadas anteriormente. Os resultados obtidos com o uso de redes neurais em situações semelhantes às encontradas no projeto de dicionários do quantizador vetorial, sugerem que sua implementação merece ser investigada. Neste trabalho, essas investigações são direcionadas para o uso de rede neurais que utilizam algoritmos denominados KMVVT (Kohonen Modificado com Vizinhança Centrada em Torno do Vetor de Treino) [41, 42] e SSC (Competitivo no Espaço Sináptico) [43].

4.3.1 Topologia das Redes Neurais

Um dos objetivos da pesquisa sobre Redes Neurais artificiais é desenvolver morfologias neurais matemáticas, não necessariamente baseadas em Modelos Biológicos, que possam realizar funções diversas. Na maior parte dos casos, modelos neurais são compostos de muitos elementos não lineares que operam em paralelo e que são classificados de acordo com padrões de operação ligados a Modelos Biológicos.

Em virtude das diferenças entre algumas ou, as vezes, todas as entidades envolvidas, diferentes estruturas de redes neurais têm sido desenvolvidas por pesquisadores [82].

Em geral, as redes podem ser classificadas como redes de propagação direta e redes com realimentação. Em ambos os casos os neurônios são organizados em camadas e

alguns padrões básicos de ligação podem ser identificados. As camadas podem ou não apresentar ligações internas; o que normalmente as identifica é o padrão de ligação dos seus neurônios com os de outras camadas.

As camadas de uma rede que têm ligação direta com as entradas externas são chamadas de camadas de entrada e as que fornecem resposta são as camadas de saída. Camadas que não são de entrada nem de saída são ditas camadas ocultas. Qualquer camada que efetue transformações nos dados pode ser chamada de camada de processamento.

As ligações individuais podem ser excitatórias ou inibitórias, dependendo de contribuírem para o aumento ou diminuição da ativação do neurônio [82, 85]. Podem existir também outros tipos de entradas.

Uma arquiteura bastante comum, encontrada em várias aplicações, é a rede de propagação direta sem realimentação (Figura 4.4 [86]). São também bastante utilizadas as redes de camadas com conexões laterais (Figura 4.5 [86]) e redes com realimentação entre camadas ou com neurônios realimentados. Outra forma de arquitetura de grande importância é a rede interconectada (Figura 4.6 [86]). Nesse tipo de arquitetura não existem camadas ocultas e a interconexão pode ser total ou parcial [86]. Redes que apresentam algum tipo de realimentação são também chamadas de redes recorrentes.

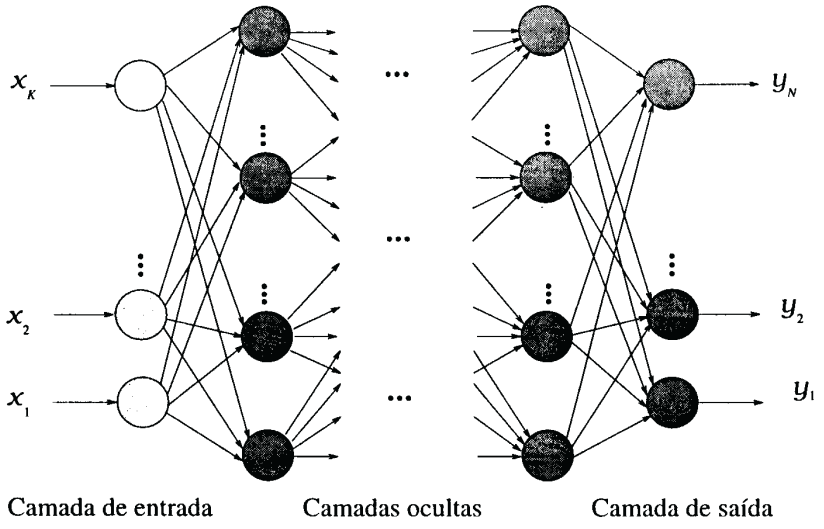


Figura 4.4: Rede de propagação direta sem realimentação.

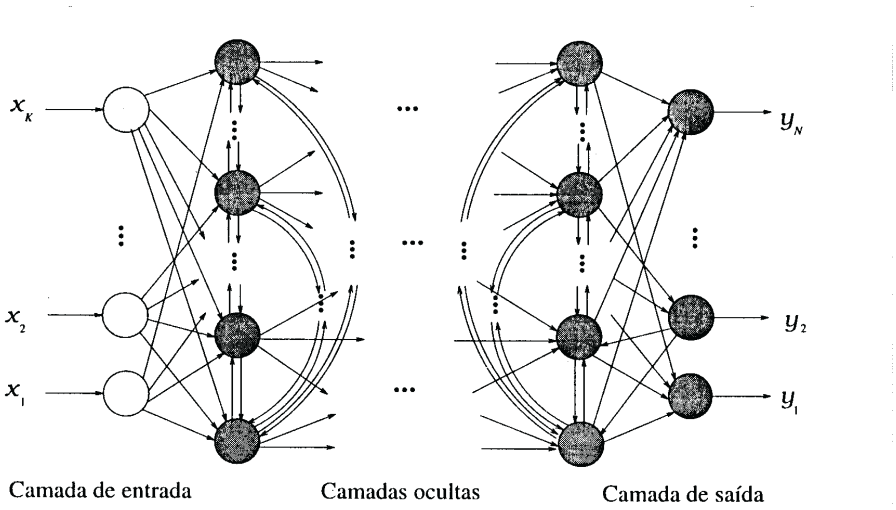


Figura 4.5: Rede de camadas com conexões laterais.

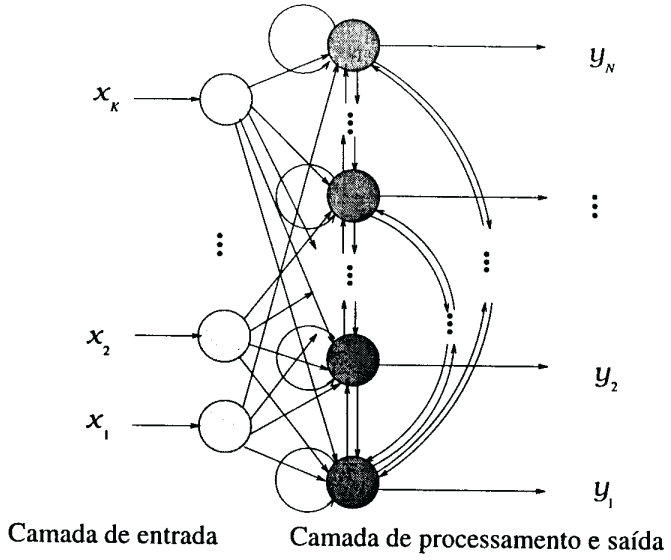


Figura 4.6: Rede interconectada.

Em algumas arquiteturas, os neurônios apresentam um comportamento competitivo: para cada entrada, apenas um neurônio tem o direito de permanecer ativo. Essas redes podem ser chamadas de redes competitivas. A rede competitiva mais simples consiste de uma única camada de neurônios totalmente interconectados, todos eles ligados com todas as entradas [88, 89]. As sinapses das entradas externas e de realimentação são excitatórias; as sinapses laterais são inibitórias (Figura 4.7 [86]). Os critérios usados pelo algoritmo para determinar o neurônio vencedor dependem da aplicação. As redes de Kohonen ou mapas de aspecto auto-organizativo são exemplos de redes competitivas [81, 82, 86].

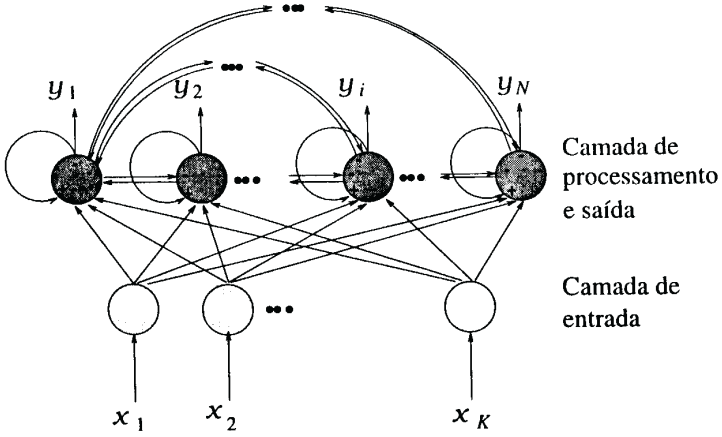


Figura 4.7: Rede competitiva simples.

4.3.2 Regras de Treinamento

A propriedade mais importante das redes neurais é a habilidade de aprender com seu ambiente e com isso melhorar seu desempenho. Isso é realizado por meio de um processo (treinamento) iterativo de ajustes aplicado a seus pesos. O aprendizado ocorre quando a rede neural atinge uma solução generalizada para uma classe de problemas.

Denomina-se algoritmo de aprendizado a um conjunto de regras bem definidas para a solução de um problema de aprendizado. Existem muitos tipos de algoritmos de aprendizado específicos para determinados modelos de redes neurais, os quais diferem entre si principalmente pelo modo como os pesos são modificados.

Qualquer mudança na memória da rede caracteriza um aprendizado e cada arquitetura necessita de um algoritmo adequado para efetuar essas mudanças, de acordo

com uma seqüência de treinamento. Os algoritmos de treinamento podem ser supervisionados ou não supervisionados, embora em algumas situações sejam também usados algoritmos híbridos [82, 86].

Algoritmos Supervisionados

A principal característica dos algoritmos supervisionados é a observação de algum tipo de medida de erro externo para orientar os ajustes dos parâmetros, buscando sempre minimizar ou pelo menos diminuir esse erro. Outra atribuição desses algoritmos é a de decidir quando parar de treinar [82, 86]. O algoritmo supervisionado não foi objeto deste trabalho.

Algoritmos Não Supervisonados

No aprendizado não supervisonado ou auto-organizativo, apenas erros internos são observados para efetuar as modificações nos parâmetros [82, 85, 86].

O algoritmo não supervisonado competitivo é usado para treinar redes neurais competitivas. Esse algoritmo foi inicialmente estudado para a área de reconhecimento de padrões [90]. A regra de aprendizado competitivo padrão é definida da forma [82, 83, 86]:

$$\Delta w_{ij} = \eta(n) \cdot \mathcal{O}_i \cdot (x_j - w_{ij}), \tag{4.5}$$

em que Δw_{ij} é a modificação introduzida na j-ésima componente (sinapse) do neurônio $\vec{w}_i$, $\eta(n)$ é a taxa de aprendizagem na n-ésima iteração, x_j é a j-ésima componente do vetor de treino $\vec{x}$ e w_{ij} é a j-ésima componente do neurônio $\vec{w}_i$ ($1 \leq i \leq M$ e $1 \leq j \leq K$) e $\mathcal{O}_i$ é a função que define a vizinhança em torno do neurônio vencedor $\vec{w}_{i^*}$.

A função $\mathcal{O}_i$ é definida como

$$\mathcal{O}_i = \begin{cases} 1 & \text{, se } d(x, w_i) < d(x, w_j), \quad \forall j \neq i \\ 0 & \text{, caso contrário} \end{cases} \tag{4.6}$$

Ou seja, se o neurônio $\vec{w}_i$ é o vencedor (ou seja, se $i = i^*$), então $\mathcal{O}_i = 1$. Para os demais neurônios, $i \neq i^*$, então $\mathcal{O}_i = 0$.

Convém salientar que, em se tratando de projeto de dicionários, M e K correspondem, respectivamente, ao número de vetores-código (ou número de níveis, em analogia com quantização escalar) e à dimensão do quantizador vetorial.

Depois da inicialização dos pesos, o algoritmo continua com a aquisição de um vetor de treinamento $\vec{x}$ da seqüência de treino. O neurônio vencedor (nesse contexto, vetor-código vencedor), $\vec{w}_{i^*}$, é determinado de acordo com a regra escolhida e seus pesos são ajustados de acordo com a regra competitiva padrão.

A taxa de aprendizado $\eta(n)$ é uma função empírica da iteração n e, para assegurar a convergência, deve ser uma função monoticamente decrescente.

O número total de iterações n_{max} é um parâmetro que deve ser estimado; depois de alguma experiência certas regras são aprendidas. Para evitar "sobretreinamento" [83, 86], esse valor não deve ser muito grande.

O Algoritmo de Kohonen

O algoritmo de Kohonen produz dicionários para quantização vetorial através de uma regra de aprendizagem não-supervisionada para atualização das sinapses da rede neural.

Ao se utilizar o mapa auto-organizativo de Kohonen (SOM, *self-organizing map*) [81] para projetos de dicionários, os vetores-código (neurônios) K-dimensionais são geralmente associados a nós em um arranjo unidimensional ou em uma grade bidimensional. Os neurônios $\vec{w}_i$ são aleatoriamente inicializados e iterativamente atualizados de acordo com a utilização de uma seqüência de treino.

Os pesos sinápticos são inicializados com pequenos valores aleatórios e atualizados através da generalização da regra de adaptação competitiva: uma seqüência de treinamento é apresentada e, para cada vetor de entrada (vetor de treino), um neurônio vencedor $\vec{w}_{i^*}$ é determinado e todos os neurônios pertencentes a uma vizinhança, $\mathcal{N}_{\vec{w}_{i^*}}$, definida em torno do nó do neurônio vencedor são atualizados.

O algoritmo de Kohonen consiste da seguinte seqüência de passos [81, 82, 83]:

1. Apresente o vetor de treino $\vec{x}$;
2. Encontre o neurônio vencedor $\vec{w}_{i^*}$ (de acordo com o critério de distância mínima $d(\vec{x}, \vec{w}_{i^*}) < d(\vec{x}, \vec{w}_i), \forall i \neq i^*$);
3. Atualize $\vec{w}_{i^*}$ e a vizinhança $\mathcal{N}_{\vec{w}_{i^*}} = \{\vec{w}_i | d_g(\vec{w}_i, \vec{w}_{i^*}) \leq r_g(n)\}$, na direção de $\vec{x}$, ou seja, $\Delta w_{ij} = \eta(n) \cdot \mathcal{O}_i \cdot (x_j - w_{ij})$.

No algoritmo $d(\cdot)$ é a medida de distorção definida no espaço de padrões (espaço de entrada, R^K), n é a iteração corrente ($1 \leq n \leq n_{max}$), $\eta(n)$ é a taxa de aprendizagem na n-ésima iteração ($0 < \eta(n) < 1$), $d_g(\cdot)$ é uma distância medida na grade de nós, $r_g(n)$ é o raio de vizinhança (definida na grade de nós) na n-ésima iteração, $\mathcal{O}_i$ é a função que define a vizinhança em torno do neurônio $\vec{w}_{i^*}$ ($\mathcal{O}_i = 1$ para $\vec{w}_i \in \mathcal{N}_{\vec{w}_{i^*}}$, $\mathcal{O}_i = 0$, caso contrário), x_j é a j-ésima componente de $\vec{x}$ e w_{ij} é a j-ésima componente de $\vec{w}_i$ ($1 \leq i \leq N, 1 \leq j \leq K$).

Os passos 1 a 3 são repetidos até que todos os vetores da seqüência de treino sejam apresentados. Tanto a taxa de aprendizagem como a função que define a vizinhança decrescem com a iteração. O procedimento completo é repetido iterativamente n_{max} vezes (n_{max} passagens da seqüência de treino). A taxa de aprendizagem $\eta(n)$ e o raio de vizinhança $r_g(n)$ decrescem a cada iteração n.

Resumindo:

- Encontre a unidade mais semelhante à unidade de treinamento;
- Aumente a similaridade dessa unidade, e das unidades pertencentes à vizinhança, com a entrada.

Podem ser identificadas duas fases no processo de aprendizagem que define o algoritmo não supervisionado de Kohonen: fase de auto-organização ou ordenamento e fase de convergência. Na primeira, é obtido um ordenamento topológico dos neurônios na grade de nós; para tanto, a função raio de vizinhança deve incluir, no início do processo de treinamento, quase todos os neurônios cujos nós (na grade) estejam centrados no nó que identifica o neurônio vencedor; a função raio de vizinhança é então gradualmente reduzida de modo a incluir poucos neurônios ou, eventualmente, apenas o neurônio vencedor. A segunda fase é necessária para proporcionar uma melhor "sintonia" dos neurônios com a distribuição estatística da seqüência de treinamento. Nessa fase, tanto o raio de vizinhança como a taxa de aprendizagem são mantidos pequenos.

O Algoritmo Modificado de Kohonen com Vizinhança Centrada em Torno do Vetor de Treino (KMVVT)

O algoritmo KMVVT, embora inspirado no processo de treinamento proposto por Kohonen, utiliza uma abordagem diferente para definição da vizinhança de atualização

dos neurônios. Em virtude dos pesos sinápticos poderem ser vistos como as coordenadas dos neurônios (vetores-código, nesse contexto) no espaço R^K, a vizinhança pode ser adequada e satisfatoriamente definida no próprio espaço de padrões R^K, como uma hiper-esfera centrada em torno do vetor de treino $\vec{x}$.

São definidas duas fases no algoritmo KMVVT. Na primeira, todos os neurônios pertencentes à vizinhança $\mathcal{N}_{\vec{x}}$ têm seus pesos atualizados. Na segunda, apenas o neurônio vencedor tem seus pesos ajustados.

Podem ser apontadas, portanto, diversas diferenças entre o algoritmo KMVVT e o algoritmo de Kohonen (SOM): no primeiro, a vizinhança é definida no próprio espaço R^K (precisamente, como uma hiper-esfera), enquanto que no segundo é definida em um arranjo topológico de nós (em geral, um mapa bidimensional, que pode ser retangular, hexagonal, etc.); no algoritmo KMVVT, a vizinhança é centrada no vetor de treino, enquanto que no algoritmo SOM é centrada em torno do nó (em uma grade) do neurônio vencedor. Por este motivo, em sua primeira fase, ao contrário do algoritmo SOM, o algoritmo KMVVT dispensa a necessidade de determinação do neurônio vencedor. Além disso, a segunda fase do algoritmo de Kohonen contempla a possibilidade de serem atualizados os neurônios pertencentes a uma pequena vizinhança; no algoritmo KMVVT, por sua vez, não se utiliza vizinhança na segunda fase de treinamento: apenas o neurônio vencedor é atualizado. A modificação produz uma influência mais efetiva do sinal de treinamento nos vetores peso [73].

As Figuras 4.8 e 4.9 ilustram as diferenças apresentadas entre o algoritmo de Kohonen e o algoritmo KMVVT.

É importante ressaltar que o algoritmo KMVVT é denotado por KMTAU (Kohonen Modificado com Taxa de Aprendizagem Uniforme) em [73] e por MKOH (*Modified Kohonen's Algorithm*) em [91].

Algoritmo Competitivo no Espaço Sináptico (SSC)

No algoritmo SSC, apenas o neurônio vencedor tem seus pesos sinápticos atualizados. Neste sentido, o algoritmo SSC corresponde à segunda fase de treinamento do algoritmo KMVVT. O algoritmo SSC pode ser visto como uma versão simplificada do algoritmo KMVVT, em virtude de não ser necessário definir uma estrutura de vizinhança [92]. Em termos de complexidade computacional, portanto, o algoritmo SSC apresenta-se como uma alternativa mais adequada que o KMVVT.

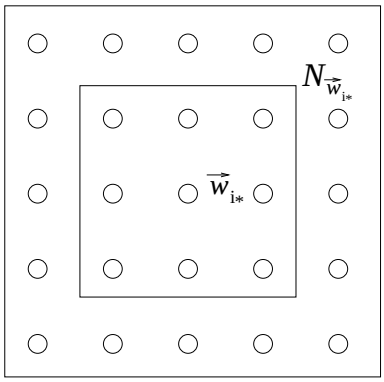


Figura 4.8: Uma vizinhança quadrada $\mathcal{N}_{\vec{w}_{i^*}}$ em torno do nó que identifica o neurônio vencedor $\vec{w}_{i^*}$. A vizinhança é definida em uma grade ou mapa bidimensional.

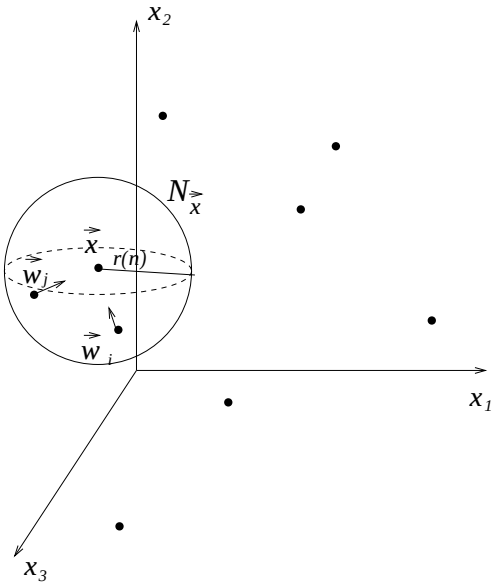


Figura 4.9: Uma vizinhança esférica $\mathcal{N}_{\vec{x}}$ em torno do vetor de treino $\vec{x}$. A vizinhança é definida no espaço sináptico.

Apesar da utilização do acrônimo SSC, o algoritmo ora apresentado difere do algoritmo SSC proposto por França e Aguiar Neto em [93], em virtude de possuir um parâmetro a mais: a *taxa de aprendizagem final*, introduzido (não explicitamente relatado) em [43].

Os algoritmos KMVVT e SSC se apresentaram como alternativas adequadas para projeto de dicionários aplicados à codificação de forma de onda de voz [73, 92, 94], à codificação de imagem [43, 73, 92] e à codificação de sinais com Distribuição Gaussiana e de Gauss-Markov [73, 92]. Torna-se interessante, portanto, a investigação do uso desses algoritmos no contexto do reconhecimento automático de identidade vocal, como alternativa no projeto de dicionários para a Quantização Vetorial Paramétrica.

4.4 Modelos de Markov Escondidos

Um problema de fundamental importância no estudo dos sinais presentes no mundo real é a caracterização desses sinais reais em termos de modelos. Uma classe de modelos de sinais de bastante interesse agrupa os modelos estatísticos, nos quais tenta-se caracterizar as propriedades estatísticas do sinal. A hipótese levantada pelos modelos estatísticos é que o sinal pode ser caracterizado como um processo aleatório paramétrico, e que os parâmetros do processo estocástico podem ser determinados (estimados) de uma maneira precisa e bem definida. Exemplos desses modelos incluem processos Gaussianos, processos de Poisson, processos de Markov, processos de Markov Escondidos, dentre outros [23, 61].

Para aplicações em processamento de sinais de voz, os modelos estatísticos têm proporcionado bons resultados. Neste trabalho, em particular, estuda-se a modelagem de sinais de voz a partir de funções probabilísticas de cadeias de Markov, ou ainda denominados, Modelos de Markov Escondidos (HMM - *Hidden Markov Models*).

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, com cada função aleatória associada a cada um dos estados. Para instantes de tempo discretos, assume-se que o processo está em algum estado e uma seqüência de observações 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 a cada estado e não pode observar diretamente os estados do canal de Markov básico; daí o termo Modelo de Markov Escondido [37]. Portanto, as principais características dos modelos de Markov Escondidos são: os estados não são diretamente observáveis, as observações são funções probabilísticas dos estados e as transições dos estados são probabilísticas [69].

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 a canais de Markov de ordem 1, i.e., aqueles nos quais a probabilidade de transição para algum estado depende somente desse estado e do estado predecessor [23, 24, 37].

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 conseqüentemente podem formar uma base teórica para uso em um grande grupo de aplicações; segundo, os modelos, quando aplicados apropriadamente, trabalham muito bem para várias aplicações práticas.

O uso de HMMs foi proposto, inicialmente, por Baker e, independentemente, por um grupo da IBM [23]. Até um certo tempo atrás pouco se ouvia falar em Modelos de Markov para reconhecimento de locutor, muito embora esses tenham sido estudados desde os anos 60. Com o passar do tempo, percebeu-se que as técnicas de reconhecimento mais usuais (*e.g.*: Análise por Predição Linear, Quantização Vetorial e Alinhamento Dinâmico no Tempo) apesar de apresentarem bons resultados, tinham alguns problemas. Por exemplo, o alto custo computacional do método utilizando programação dinâmica. Além disso, apresentavam sérias dificuldades quando da execução de tarefas mais complicadas (*e.g.*, reconhecimento de locutor independente do texto) [24, 37].

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 processamento) e por basear-se em modelos estocásticos do sinal de voz, sendo capazes de modelar

vários eventos, tais como fonemas, sílabas, etc. [70], o que os tornam bastante flexíveis. Algumas das vantagens do uso de HMM são [95]:

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 são modeladas inerentemente.
3. Considera as variações estatísticas do sinal de entrada por estarem implícitas na própria formulação probabilística.
4. Não é necessário, a priori, uma distribuição estatística das entradas para estimação dos parâmetros, o que não é o caso, usualmente, em outras técnicas estatísticas.

É 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 o caracterizam. Assim, a potência espectral para 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 [37].

Os sistemas de reconhecimento de locutor, utilizando HMM, baseiam-se na descrição das características da voz a ser reconhecida. A informação temporal é modelada pelo HMM e os blocos de dados consecutivos são forçados a se alinhar com a seqüência de sons da linguagem que está sendo falada pelo locutor [96].

Em um SRAL, o som produzido por cada locutor é produto das características do trato vocal, tamanho da garganta, posição da língua e uma série de outros fatores. Cada um desses fatores interage para produzir o som de uma elocução de acordo com quem fala, e os sons que o sistema detecta são as variações dos sons gerados dessas alterações físicas internas da pessoa que está falando. Algumas técnicas para reconhecimento de locutor consideram a produção interna da voz como sendo uma seqüência de estados escondidos, e o som resultante como uma seqüência de estados observáveis gerados por uma voz processada que mais se aproxima do estado verdadeiro (escondido) [8]. Esta é outra característica que permite a aplicabilidade dos Modelos de Markov Escondidos aos sistemas de reconhecimento de locutor.

4.4.1 Tipos de HMM

Para o reconhecimento de locutor utilizando HMM, a primeira questão envolvida na determinação dos HMMs ótimos [1], para cada locutor, é a estrutura do modelo.

Existem dois casos gerais do modelo que são de interesse: o caso "ergódico" (ou sem restrição), no qual a cadeia de Markov é ergódica e todos os estados são aperiódicos e recorrentes não nulos (Figura 4.10) e o caso "esquerda-direita" (ou serial restrito) (Figura 4.11), no qual uma transição do estado q_i para o estado q_j é possível se $j \geq i$ (i.e., existe uma progressão seqüencial através dos estados do modelo) [97].

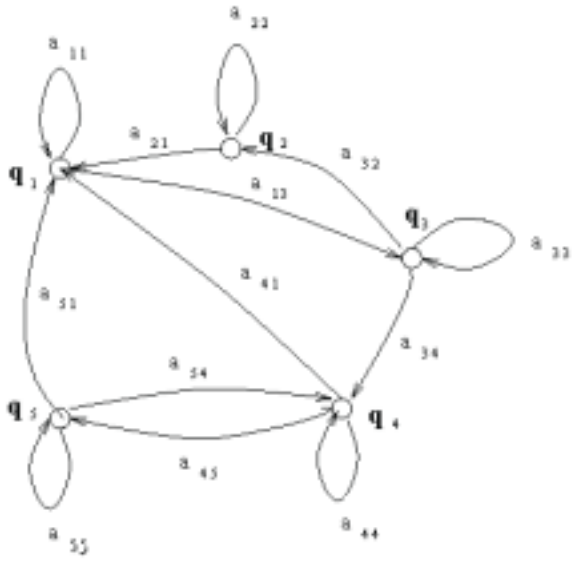


Figura 4.10: HMM - "ergódico" com 5 estados.

Os modelos "esquerda-direita" apresentam melhor desempenho do que os ergódicos no caso do reconhecimento de locutor. Esse resultado advém da própria estrutura da fala ser inerentemente seqüencial. Além disso, a liberdade adicional de transição de estados presente nos modelos ergódicos não reflete as variações dos parâmetros da fala caracterizados por um vetor de padrões. Foi verificado, portanto, que o uso de modelos ergódicos apresentam um desempenho inferior, quando comparados com modelos "esquerda-direita", para reconhecimento de fala e de locutor [24, 61].

[1] HMM ótimo é o conjunto de parâmetros de um HMM que melhor representa um determinado locutor.

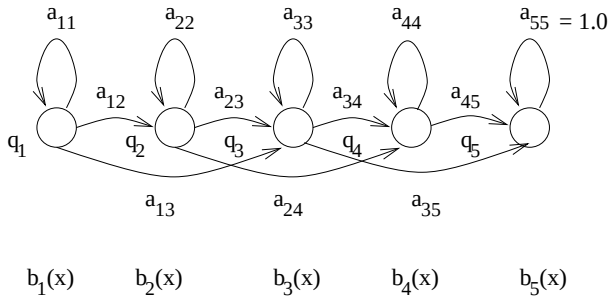


Figura 4.11: HMM - "esquerda-direita" com 5 estados.

Os modelos "esquerda-direita" têm as seguintes propriedades [37]:

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, 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 visitado num tempo posterior.

Para o reconhecimento de locutor, o sinal de voz é assumido como sendo uma função estocástica da seqüência de estados da cadeia de Markov. O objetivo é escolher os parâmetros do HMM que correspondam de maneira ótima às características observadas de um dado sinal, para um dado locutor [70].

Diante do exposto, e devido à maior simplicidade em relação ao HMM do tipo "ergódico", optou-se pelo uso do HMM "esquerda-direita" neste trabalho.

4.4.2 Parâmetros do Modelo

Os parâmetros que caracterizam o HMM, $\lambda = (\mathcal{A}, \mathcal{B}, \pi)$, da Figura 4.11, são:

1. N, número de estados do modelo. Estados individuais são denotados por $(q_1, q_2, ..., q_N)$.

2. $\mathcal{A} = [a_{ij}]$, $1 \leq i, j \leq N$, a matriz transição de estados. Cada a_{ij} corresponde à probabilidade de ocorrer uma 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.

   $a_{ij} = \text{prob}(q_j \text{ em t+1} | q_i \text{ em t})$. Para modelos "esquerda-direita" usa-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 um dado estado q_j, a saída do modelo através de um vetor aleatório com uma função densidade de probabilidade (f.d.p) b_j [97].

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

Visando o tratamento matemático e computacional, são realizadas as seguintes considerações acerca da teoria dos HMMs [69, 98, 99]:

1. A suposição de Markov – A partir da definição anterior, as probabilidades de transição são dadas por:

$$a_{ij} = p\{q_{t+1} = j | q_t = i\}. \tag{4.7}$$

   Em outras palavras, assume-se que o próximo estado depende apenas do estado corrente. Isto resulta em um HMM de primeira ordem. Entretanto, geralmente o próximo estado pode depender de n outros estados anteriores, gerando assim um modelo chamado HMM de ordem n. Porém, observa-se que um HMM de ordem mais elevada resulta no aumento da complexidade computacional. Embora o HMM de primeira ordem seja o mais comum, alguns estudos têm usado HMM de ordem superior [98].

2. A suposição da estacionariedade – Assume-se que as probabilidades de transição de estados são independentes do tempo atual, no qual as transições foram realizadas. Matematicamente,

$$p\{q_{t_1+1}\} = p\{q_{t_1+1} = j | q_{t_1} = i\} = p\{q_{t_2+1} = j | q_{t_2} = i\}. \tag{4.8}$$

3. A suposição da independência entre as saídas – Assume-se que a saída atual (observação) é estatisticamente independente das saídas anteriores (observações). Pode-se formular essa suposição matematicamente, considerando-se uma seqüência de observações, $\mathbf{O} = \{O_1, O_2, ..., O_T\}$ e um modelo λ, obtendo-se

$$P\{\mathbf{O}|q_1, q_2, ..., q_T, \lambda\} = \prod_{t=1}^{T} P(O_t|q_t, \lambda) \tag{4.9}$$

   Entretanto, diferentemente das suposições anteriores, esta tem limitado a aplicação dos HMMs em alguns casos [98].

Para reconhecimento de locutor, assume-se que o sinal de voz do locutor a ser representado pelo HMM consiste de uma seqüência de vetores de observações $\mathbf{O} = \{O_1, O_2, ..., O_T\}$. Cada vetor O_t é formado pelos parâmetros obtidos para cada bloco de amostras, que caracteriza o sinal de voz no t-ésimo intervalo de tempo. Assim, cada bloco de amostras do sinal de voz, corresponderá a um determinado intervalo de tempo. Dessa 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 o HMM de densidades contínuas [70]. Uma forma alternativa para o uso de HMMs é a combinação com a quantização vetorial [24], em que os parâmetros de interesse são transformados em um conjunto de observações discretas. Tem-se, então, os chamados HMMs de densidades discretas [23] (utilizado neste trabalho).

Para o HMM do tipo discreto, representa-se cada vetor O_t de uma seqüência de vetores de observações do l-ésimo locutor, $\mathbf{O^l} = \{O_1, O_2, ..., O_T\}$, $1 \leq l \leq L$ e $1 \leq t \leq T$, por um dos M possíveis símbolos $w_k \in W$, $1 \leq k \leq M$, em que W representa um alfabeto discreto obtido por meio da quantização vetorial dos vetores de observações. Nesse caso, a matriz $\mathcal{B} = [b_j(k)]$ indica a probabilidade de observar-se um símbolo w_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 w_k, no instante de tempo t (t-ésimo vetor, tempo discreto), no estado q_j [23].

Em resumo, tem-se:

1. O modelo para cada l-ésimo locutor é denotado por $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$.

2. Inicia-se no estado particular q_i (t=1), que depende da distribuição do estado inicial e produz um símbolo de saída $O_t = w_k$, de acordo com $b_i(k)$.

3. Este se move para o estado q_j ou permanece no estado q_i, de acordo com a_{ij}.

4. Esse processo de saída do símbolo e transição para o próximo estado se repete até que o objetivo seja atingido (*e.g.*, quando o número de iterações estabelecido é alcançado).

5. Em modelos "esquerda-direita", o processo se inicia no estado q_1 ($t = 1$) e termina quando se atinge T passos ($t = T$).

6. Assim, a partir da seqüência de observações $\mathbf{O^l} = \{O_1, O_2, ..., O_T\}$ e dos parâmetros necessários, obtém-se o HMM referente a cada l-ésimo locutor [23, 95].

A análise espectral é uma das formas mais usuais de se obter os vetores O_t de uma seqüência de observações $\mathbf{O^l} = \{O_1, O_2, ..., O_T\}$ para as amostras de voz de um l-ésimo locutor. O tipo de análise espectral aqui usada é a Análise por Predição Linear e os coeficientes obtidos a partir dessa análise (descritos no Capítulo 3) constituem cada vetor O_t da seqüência de observações $\mathbf{O^l} = \{O_1, O_2, ..., O_T\}$ [23].

Seja $W = \{w_1, w_2, ..., w_M\}$ o alfabeto discreto utilizado para representar a seqüência de observações $\mathbf{O^l} = \{O_1, O_2, ..., O_T\}$, define-se que a probabilidade de ocorrência de uma dada seqüência será [23]

$$P\{O_1, O_2, .., O_T\} = \pi_i \cdot \mathcal{A} \cdot \mathcal{B}. \tag{4.10}$$

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 observações $\{O_1, O_2, .., O_T\}$.

Como na maioria dos sistemas de identificação de locutor, 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 identificar um locutor, calcula-se a medida de probabilidade associada aos HMM's de referência já armazenados. O locutor aceito é aquele que apresentar o maior valor de probabilidade. Dessa forma, a modelagem por HMM necessita, para a as etapas de treinamento e identificação, da resolução de três problemas básicos, descritos a seguir.

4.4.3 Os três problemas básicos dos HMMs e suas soluções

Os aspectos teóricos dos Modelos de Markov Escondidos podem ser caracterizados em termos da solução de três problemas fundamentais, descritos a seguir [23, 61, 69, 99].

1. Gerar um HMM dada uma seqüência de observações (treinamento). Ou seja, ajustar os parâmetros do modelo de modo a representar com maior eficiência o sinal que está sendo modelado;

2. Encontrar a probabilidade de uma seqüência de observações dado um HMM (estimação ou reconhecimento). Ou seja, realizar o cálculo da probabilidade (ou verossimilhança) de uma seqüência de observações dado um HMM específico;

3. Encontrar a seqüência de estados escondidos que mais provavelmente gerou uma seqüência observada (decodificação). Ou seja, determinar a seqüência "ótima" de estados do modelo.

Esses problemas podem ser colocados de maneira mais explícita, como a seguir.

Problema 1. O problema do treinamento

Dado um modelo λ e uma seqüência de observações $\mathbf{O} = \{O_1, O_2, ..., O_T\}$, como ajustar os parâmetros do modelo $\{A, B, \pi\}$, visando maximizar $P\{\mathbf{O}|\lambda\}$. Ou seja, gerar um HMM de uma seqüência de observações. É comum utilizar na solução desse problema o algoritmo *forward-backward* e o método de reestimação de *Baum-Welch* [69, 98, 99].

Problema 2. O problema da estimação (reconhecimento)

Dado um modelo λ e uma seqüência de observações $\mathbf{O} = \{O_1, O_2, ..., O_T\}$, qual é a probabilidade de que as observações tenham sido geradas pelo modelo, $P\{\mathbf{O}|\lambda\}$? Ou seja, dado um conjunto de modelos (HMMs), determinar qual, mais provavelmente, gerou a seqüência de observações.

Em reconhecimento de locutor uma seqüência de observações é formada a partir da elocução da sentença por um dado locutor. O locutor é reconhecido pela identificação do HMM mais provável de ter gerado a seqüência de observações [69, 98, 99].

Problema 3. O problema da decodificação

Dado um modelo λ e uma seqüência de observações $\mathbf{O} = \{O_1, O_2, ..., O_T\}$, qual a seqüência de estados do modelo que mais provavelmente produziu a seqüência de

observações ? Para a solução desse problema utiliza-se o algoritmo de Viterbi [69, 98, 99].

Para o reconhecimento de locutor é necessário solucionar os problemas 1 e 2, que constituem as fases de treinamento e reconhecimento, respectivamente. A solução do problema 3 pode ser útil, como etapa de otimização do problema 2.

4.4.3.1 Solução do Problema 1

Na fase de treinamento é feita a estimação dos parâmetros dos modelos $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$, um modelo para cada l-ésimo locutor $(1 \leq l \leq L)$ [95]. 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 a reestimação determinar os valores ótimos (máxima verossimilhança), que corresponderão aos HMMs de referência, um para cada um dos L locutores.

No caso de Modelos de Markov, a estimação pode ser realizada usando o processo iterativo de *Baum-Welch*, que pode ser descrito por meio dos passos [23, 37]:

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* (as 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 regras simples, de forma a satisfazer as restrições do modelo "esquerda-direita". O vetor de probabilidade inicial é $\pi_i = \{1, ..., 0\}$, 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

visitado num instante de tempo posterior. Esta restrição deverá se manter até o final do processo de reestimação. Para a matriz $\mathcal{B} = [b_j(k)]$, assume-se que todos os símbolos nos estados são "igualmente prováveis" e $b_j(k)$ inicia-se com $1/M$ para todo j, k, para simplificar.

As equações do método de reestimação de *Baum-Welch* são [23, 37]:

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 w_k)/(número esperado de vezes no estado j).

ou seja,

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_t(i)}, \qquad 1 \leq i \leq N,\ 1 \leq j \leq N \tag{4.11}$$

$$\overline{b_j(k)} = \frac{\sum_{t=1, O_t = w_k}^{T} \alpha_t(j) \beta_t(j)}{\sum_{t=1}^{T} \alpha_t(j) \beta_t(j)}, \qquad 1 \leq j \leq N,\ 1 \leq k \leq M \tag{4.12}$$

com:

$$\sum_{j=1}^{N} a_{ij} = 1;\ \sum_{k=1}^{M} b_j(k) = 1;\ \sum_{i=1}^{N} \pi_i = 1,\ a_{ij} \geq 0;\ b_j(k) \geq 0;\ \pi_i \geq 0. \tag{4.13}$$

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 w_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 observações $\mathbf{O^1} = \{O_1, O_2, ..., O_T\}$, segundo o tempo crescente (iniciando em $t = 1$ indo até $t = T$), sendo formulada em [23] como:

1. Inicialização:

$$\alpha_1(i) = \pi_i b_i(O_1), \qquad 1 \leq i \leq N; \tag{4.14}$$

2. Indução:

$$\alpha_{t+1}(j) = \left\{ \sum_{i=1}^{N} \alpha_t(i) a_{ij} \right\} b_j(O_{t+1}), \qquad 1 \leq t \leq T-1, \ 1 \leq j \leq N. \tag{4.15}$$

A probabilidade P_l, associada ao modelo $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$, é determinada por [23]

$$P_l = \text{Prob}(\mathbf{O^l}|\lambda_l) = \sum_{\text{i=1}}^{\text{N}} \sum_{\text{j=1}}^{\text{N}} \alpha_\text{t}(\text{i}) \text{a}_\text{ij} \text{b}_\text{j}(\text{O}_\text{t+1}) \beta_\text{t+1}(\text{j}), \tag{4.16}$$

para algum t, $1 \leq t \leq T$.

Fazendo-se $t = T - 1$, obtém-se

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

sendo

$$\alpha_T(i) = P_l(O_1, ..., O_T, q_T = i|\lambda_l). \tag{4.18}$$

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 4.12.

A Figura 4.12 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 $\mathbf{O^l}$. Assim $\alpha_t(i)$ é a probabilidade de que os vetores $\{O_1, O_2, ..., O_T\}$ 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 $\{O_1, O_2, ..., O_T\}$ 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, obtém-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 (4.17) é realizada para todos os j-ésimos estados, $1 \leq j \leq N$, para um dado t; a computação é, então, iterativa para $t = 1, 2, ..., T-1$.

O cálculo da probabilidade de avanço é baseado na estrutura de treliça mostrada na Figura 4.13. Desde que há somente N estados (nós para cada instante de tempo na

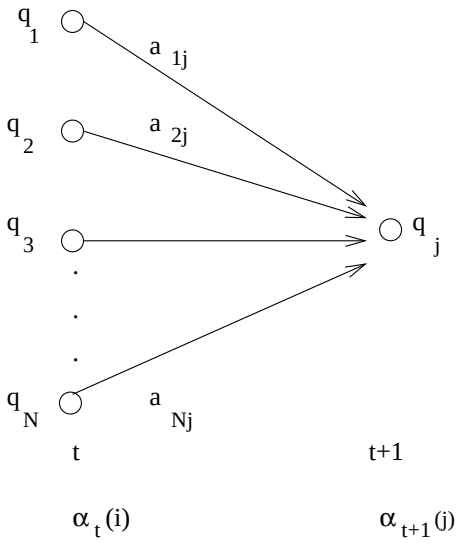


Figura 4.12: Ilustração da seqüência de operações necessárias à computação da variável *forward* $\alpha_{t+1}(j)$.

treliça), todas as possíveis seqüências de estado serão agrupadas dentro desses N nós, não importando o tamanho da seqüência de observações. No instante $t = 1$, o primeiro instante de tempo na treliça, referente ao primeiro vetor O_1 de uma dada seqüência de observações $\mathbf{O^l}$, 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 $\alpha_t(j)$, $1 \leq j \leq N$, em que 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 partir dos mesmos N pontos da grade no instante de tempo anterior [23].

De forma similar, $\beta_t(i)$ é denominada probabilidade de retrocesso (*backward probability*), pois está associada à ocorrência da seqüência de observações $\mathbf{O^l} = \{O_1, O_2, ..., O_T\}$ segundo o tempo decrescente, sendo definida como [23]

$$\beta_t(i) = P_l(O_{t+1}, O_{t+2}, ..., O_T | q_t = i, \lambda_l). \tag{4.19}$$

ou, seja, a probabilidade da seqüência de observações 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 [23]:

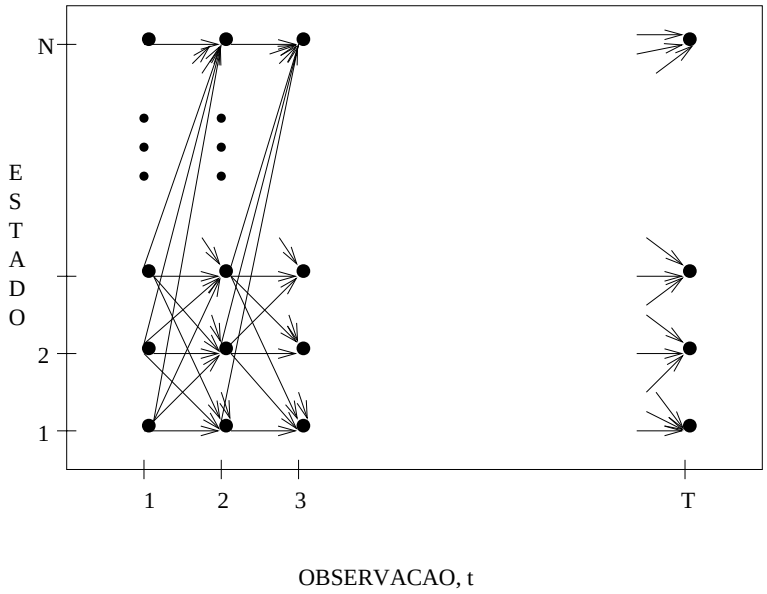


Figura 4.13: Implementação da computação de $\alpha_t(i)$ em termos de uma treliça de observações t e estados i.

1. Inicialização:

$$\beta_T(i) = 1, \quad 1 \leq i \leq N. \tag{4.20}$$

2. Indução:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, ..., 1, \ 1 \leq i \leq N. \tag{4.21}$$

O passo 1, inicialização, define arbitrariamente $\beta_T(i) = 1$ para todo i. O passo 2, ilustrado na Figura 4.14, mostra que para ter ocorrido o estado q_i no instante de tempo t, levando-se em conta a seqüência de observações 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 $b_j(O_{t+1})$).

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 observações finita como dado de treinamento. Entretanto, no método iterativo proposto por Baum e Welch

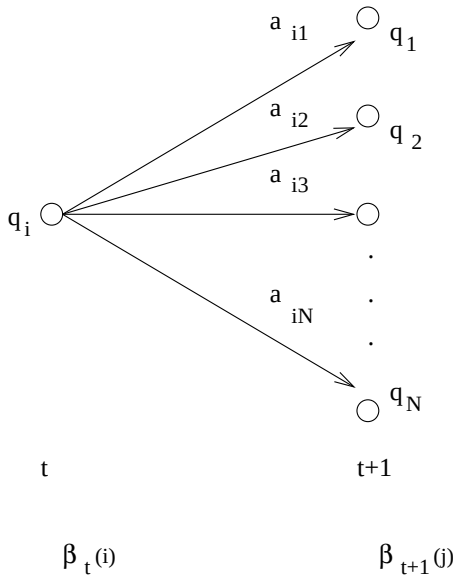


Figura 4.14: Ilustração da seqüência de operações necessárias à computação da variável *backward* $\beta_t(i)$.

em [100] escolhe-se λ_l tal que $P_l(\mathbf{O^l}|\lambda_l)$ seja localmente máxima. 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 $P_l(\mathbf{O^l}|\overline{\lambda_l}) \geq P_l(\mathbf{O^l}|\lambda_l)$. Assim, utiliza-se $\overline{\lambda_l}$ no lugar de λ_l repetindo o processo de reestimação para uma dada seqüência observada, $\mathbf{O^l}$, 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 [23], obtendo-se assim os HMMs de referência, um para cada um dos L locutores.

Uso de Múltiplas Seqüências de Observações

O maior problema associado ao HMM do tipo "esquerda-direita", reside no fato de que não se pode usar uma única seqüência de observações para treinar o modelo (isto é, para a reestimação dos parâmetros do modelo) [23, 61]. Isso se deve à natureza transitória dos estados dentro do modelo, permitindo apenas um pequeno número de observações para qualquer estado (até que uma transição seja feita para um estado sucessor). Assim, a fim de se obter dados suficientes para se fazer estimativas confiáveis

de todos os parâmetros do modelo, deve-se usar múltiplas seqüências de observações.

A modificação do método de reestimação é direta e apresentada a seguir [23].

Seja o conjunto de U seqüências de observações representado por:

$$\mathbf{O} = [\mathbf{O}^{(1)}, \mathbf{O}^{(2)}, ..., \mathbf{O}^{(u)}], \tag{4.22}$$

em que $\mathbf{O}^{(u)} = [O_1^{(u)} O_2^{(u)} ... O_{T_u}^{(u)}]$, $1 \leq u \leq U$, é a u-ésima seqüência de observações.

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

$$P(\mathbf{O}|\lambda) = \prod_{u=1}^{U} P(\mathbf{O}^{(u)}/\lambda) = \prod_{u=1}^{U} P_u. \tag{4.23}$$

Uma vez que as fórmulas de reestimação são baseadas em freqüências de ocorrências de eventos, para as múltiplas seqüências de observações essas fórmulas são modificadas adicionando-se as freqüências de ocorrências individuais de cada seqüência. Assim, as fórmulas de reestimação modificadas são [23]:

$$\overline{a_{ij}} = \frac{\sum_{u=1}^{U} \frac{1}{P_u} \sum_{t=1}^{T_u - 1} \alpha_t^u(i) a_{ij} b_j(O_{t+1}^{(u)}) \beta_{t+1}^u(j)}{\sum_{u=1}^{U} \frac{1}{P_u} \sum_{t=1}^{T_u - 1} \alpha_t^u(i) \beta_t^u(i)}, \qquad 1 \leq i \leq N,\ 1 \leq j \leq N \tag{4.24}$$

$$\overline{b_j(k)} = \frac{\sum_{u=1}^{U} \frac{1}{P_u} \sum_{t=1, s.t. O_t = w_k}^{T_u} \alpha_t^u(j) \beta_t^u(j)}{\sum_{u=1}^{U} \frac{1}{P_u} \sum_{t=1}^{T_u} \alpha_t^u(j) \beta_t^u(j)}, \qquad 1 \leq j \leq N,\ 1 \leq k \leq M. \tag{4.25}$$

4.4.3.2 Solução do Problema 2

Na fase de reconhecimento (identificação) é realizada a estimação da probabilidade de ocorrência de uma dada seqüência de observações $\mathbf{O^l} = \{O_1, O_2, ..., O_T\}$, associada a cada modelo $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$, obtido durante a fase de treinamento ($1 \leq l \leq L$).

Uma vez que os HMMs tenham sido treinados para cada locutor, a estratégia de identificação é direta [101, 102]. Para o locutor a ser identificado é obtida a seqüência de observações $\mathbf{O^l} = \{O_1, O_2, ..., O_T\}$ e gerada a tabela de códigos associadas à seqüência, através da quantização vetorial. Em seguida, é calculada a probabilidade associada a cada modelo de referência $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$ (obtidos durante a fase de treinamento).

Após o cálculo da probabilidade, através de uma regra de decisão, o locutor é aceito ou rejeitado pelo sistema.

O procedimento para o cálculo da probabilidade $P(\mathbf{O^l}|\lambda_\mathbf{l})$ é o mesmo já apresentado anteriormente na Equação (4.26), descrito a seguir.

Fazendo-se $t = T - 1$, obtém-se [101]

$$P_l(\mathbf{O^l}|\lambda_l) = \sum_{i=1}^{N} \alpha_T(i), \tag{4.26}$$

sendo:

$$\alpha_1(i) = \pi_i b_i(O_1), \qquad 1 \leq i \leq N, \tag{4.27}$$

$$\alpha_{t+1}(j) = \left\{ \sum_{i=1}^{N} \alpha_t(i) a_{ij} \right\} b_j(O_{t+1}), \qquad 1 \leq t \leq T-1, \;\; 1 \leq j \leq N. \tag{4.28}$$

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 índice do quantizador vetorial(símbolo w_k). Cada coeficiente $b_j(k)$ representa a probabilidade de ocorrência de um dado símbolo w_k, no estado j. Assim, cada coeficiente $b_j(O_t)$ corresponde ao valor da probabilidade do símbolo associado àquele estado j.

O locutor que apresentar o maior valor de probabilidade é o locutor identificado pelo sistema (ou aceito), desde que esta seja maior que um dado limiar (se é desejado evitar o acesso de locutores não cadastrados), caso contrário o locutor é rejeitado.

4.4.3.3 Solução do Problema 3

Diferentemente do que acontece no Problema 2, para o qual existe uma solução exata, existem várias maneiras possíveis de se resolver o Problema 3. A dificuldade de se encontrar a seqüência de estados ótima, associada com uma dada seqüência de observações, está exatamente na definição do que seja essa seqüência ótima, isto é, existem vários critérios de otimização. Existe uma técnica formal para se encontrar a seqüência de estados ótima única, baseada em métodos de programação dinâmica,

chamada de Algoritmo de Viterbi [23, 69, 99, 103]. Esse algoritmo é uma solução ótima recursiva ao problema de estimar a seqüência de estados de um processo de Markov discreto no tempo [61, 104].

Considerando a estrutura de treliça apresentada na Figura 4.13, a propriedade mais importante, inerente a essa estrutura, é que para cada seqüência de estados possível, Q, corresponde um único caminho através da treliça e vice-versa [61, 104].

Observando a Figura 4.15, pode-se notar que para vários instantes de tempo diferentes, existe mais de um caminho parcial chegando em cada nó (estado), cada um com determinado comprimento (valor de probabilidade). O segmento de caminho mais curto ou seja, aquele que apresenta maior valor de probabilidade, é chamado de "sobrevivente" correspondente a cada nó. Em outras palavras, para cada instante de tempo existe um número de sobreviventes igual ao número de nós na treliça.

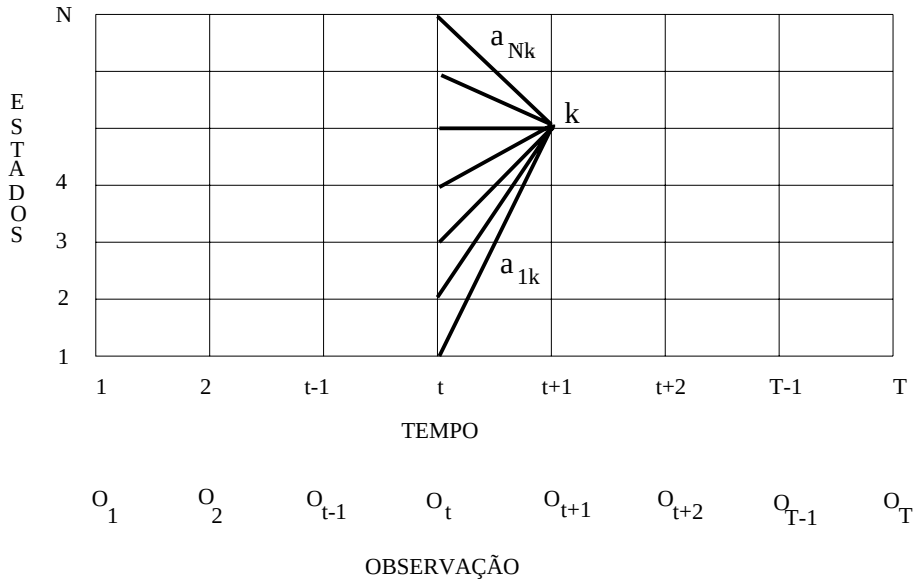


Figura 4.15: Algoritmo de Viterbi.

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

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

$$\delta_{t+1}(j) = [\max_i \delta_t(i)a_{ij}]b_j(O_{t+1}) \quad 1 \leq i \leq N. \tag{4.29}$$

Para se obter a seqüência de estados, é necessário reter a trilha do argumento que maximiza a Equação (4.29), para cada t e j. Para tanto, define-se a variável $\Psi_t(j)$. O método para se encontrar a seqüência de estados ótima é dado por [23, 61]:

1. Inicialização:

$$\delta_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N \tag{4.30}$$

$$\Psi_1(i) = 0. \tag{4.31}$$

2. Recursividade:

$$\delta_t(j) = \max_{1 \leq i \leq N}[\delta_{t-1}(i)a_{ij}]b_j(O_t), \tag{4.32}$$

$$\Psi_t(j) = \arg\max_{1 \leq i \leq N}[\delta_{t-1}(i)a_{ij}], \quad 2 \leq t \leq T, 1 \leq j \leq N. \tag{4.33}$$

3. Término:

$$P^* = \max_{1 \leq i \leq N}[\delta_T(i)], \tag{4.34}$$

$$q_T^* = \arg\max_{1 \leq i \leq N}[\delta_T(i)]. \tag{4.35}$$

4. Seqüência de estados ótima

$$q_t^* = \Psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, ..., 1. \tag{4.36}$$

O algoritmo descrito tem, portanto, a propriedade de determinar a seqüência de estados, que maximiza a probabilidade $P(\mathbf{O^l}|\lambda_\mathbf{l})$, para o l-ésimo locutor [23, 69, 99]. Assim, esse algoritmo pode ser usado para o "ajuste" da etapa de reconhecimento (identificação) e determinação da seqüência de estados ótima do modelo.

A solução desses 3 problemas, permite a elaboração de um sistema de reconhecimento automático da identidade vocal de locutores, utilizando HMM.

4.5 Discussão

O ponto principal do processo de reconhecimento de locutor é uma comparação entre padrões obtidos de um sinal de voz de um locutor a ser reconhecido (ou de teste) com padrões de referência previamente armazenados, associados aos locutores cadastrados. 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. 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 àquela referência.

Com o objetivo de se obter sistemas de reconhecimento automático de locutor eficientes, diversas técnicas estatísticas e paramétricas têm sido utilizadas, dentre as quais destacam-se: Modelos de Markov Escondidos (HMMs - *Hidden Markov Models*) [23, 24, 25, 26], Redes Neurais Artificiais [27, 28, 29], Quantização Vetorial (VQ - *Vector Quantization*) [30, 31, 32, 33, 34], Análise por Predição Linear [35, 36] e Alinhamento Dinâmico no Tempo (DTW - *Dynamic Time Warping*) [16].

A principal vantagem da QV em reconhecimento de locutor está na produção do dicionário para determinação da similaridade entre elocuções de um mesmo locutor. Para projeto dos dicionários do quantizador vetorial existem vários métodos. O método tradicional, LBG, apresenta algumas limitações. Os resultados obtidos com o uso de redes neurais em situações semelhantes às encontradas no projeto de dicionários do quantizador vetorial, sugerem que sua implementação merece ser investigada. Neste trabalho, essas investigações são direcionadas para o uso de rede neurais que utilizam algoritmos não supervisionados: algoritmo KMVVT e o SSC, visando determinar qual desses algoritmos produz dicionários que melhor representam as características vocais dos locutores a serem identificados pelo sistema.

O uso de HMM, em reconhecimento de locutor, se torna 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. Considerando que um sistema pode ser descrito como um HMM, três problemas devem ser resolvidos. O primeiro problema consiste em gerar um HMM dada uma seqüência de observações (treinamento) [69, 98, 99]. Os dois últimos problemas são: encontrar a probabilidade

de uma seqüência de observações dado um HMM (estimação ou identificação) e encontrar a seqüência de estados escondidos que mais provavelmente gerou uma seqüência observada (decodificação).

O sistema de reconhecimento (identificação) automático da identidade vocal dos locutores pode ser implementado utilizando uma das técnicas citadas ou até mesmo a combinação dessas. Neste trabalho optou-se pelo uso combinado das técnicas, em detrimento ao uso de uma única técnica.

Na tarefa de reconhecimento (identificação) são utilizados dois parâmetros para discriminação de locutores: a medida de distorção obtida a partir da quantização vetorial, seguida da probabilidade obtida do HMM. Esse último é utilizado como parâmetro de "refinamento" do processo de identificação.

A identificação é levada a efeito, portanto, através de um método híbrido, que utiliza técnicas paramétrica e estatística, para modelagem das características vocais dos locutores.

Capítulo 5

Descrição do Sistema de Identificação Automática de Locutor

5.1 Introdução

O sistema de reconhecimento (identificação) automático da identidade vocal de locutores desenvolvido, apresentado na Figura 5.1, é constituído das seguintes etapas:

1. Processamento do sinal;
2. Extração de características;
3. Quantização vetorial;
4. Regra de decisão 1;
5. Modelagem utilizando HMM;
6. Regra de decisão 2.

A seguir será feita a descrição de cada uma dessas etapas.

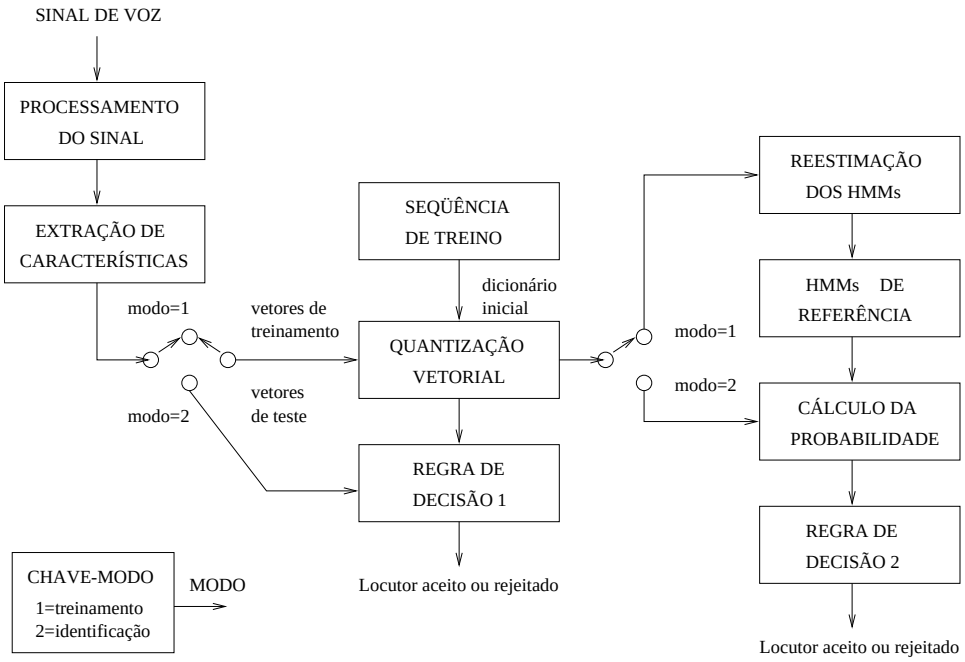


Figura 5.1: Diagrama de blocos do sistema de identificação automática de locutor.

5.2 Processamento do sinal de voz

Esta etapa inclui a aquisição do sinal, pré-ênfase e janelamento em T blocos, cada bloco contendo N_A amostras.

A aquisição e o tratamento do sinal de voz foram realizados da seguinte forma:

1. Aquisição: Placa Sound Blaster [105] disponível no LAPS/DEE/UFPB, em ambiente relativamente silencioso, com um microfone comum (modelo *Leadership*);
2. Formato: WAV (sem cabeçalho);
3. Taxa de amostragem: 11 kHz (uma vez que sinais de voz raramente possuem energia significativa acima de 5,5 kHz, metade da freqüência de amostragem [20]);
4. Resolução: 16 bits (mono);
5. Pré-ênfase do sinal [106];

6. Tamanho dos segmentos de voz: 220 amostras (20 ms), garantindo as condições de estacionariedade do sinal [1].

7. Janelamento do sinal: Janela de Hamming com superposição de 50% [106].

5.2.1 Pré-ênfase

Resultados experimentais mostram que as características do aparelho fonador mudam lentamente na geração do sinal de voz. As mudanças ocorrem para períodos em torno de 10 a 30 ms [107]. Pode-se, então, modelar o aparelho fonador humano por um sistema linear, lentamente variante com o tempo, que pode ser excitado por um trem de pulsos quase-periódicos (sons sonoros) ou por ruído branco (sons surdos). Filtrando o sinal de voz proveniente do microfone com um filtro $L(z)$ do tipo

$$L(z) = 1 - a_p z^{-1}. \tag{5.1}$$

obtém-se o modelo do sistema glotal, com os efeitos da radiação dos lábios e da variação da área da glote reduzidos [108]. A esse processo de tratamento do sinal de voz, dá-se o nome de *pré-ênfase*. O parâmetro a_p é denominado *fator de pré-ênfase*. Valores típicos de a_p são próximos de 1,0. Neste trabalho foi utilizado $a_p = 0,95$. Assim, a pré-ênfase é realizada por meio da fórmula usual [1, 27]

$$s_p(n) = s(n) - 0,95 \cdot s(n-1). \tag{5.2}$$

Além disso, as componentes de alta freqüência do sinal de voz são caracterizadas por apresentarem baixas amplitudes e por isso são facilmente afetadas pelo ruído. Apesar do sinal de voz ter a energia mais concentrada nas baixas freqüências, as freqüências mais altas são responsáveis pela geração dos sons surdos (fricativos). Assim sendo, após a aquisição do sinal, também é realizada uma pré-ênfase a fim de tornar mais plano o espectro desse sinal.

5.2.2 Segmentação para análise a curtos intervalos

A segmentação consiste em particionar o sinal de voz em segmentos, selecionados por janelas ou quadros de duração perfeitamente definida, como mostra a Figura 5.2.

Os segmentos são escolhidos dentro dos limites de estacionariedade do sinal. A segmentação é levada a efeito com superposição de 50%, visando reduzir os efeitos da

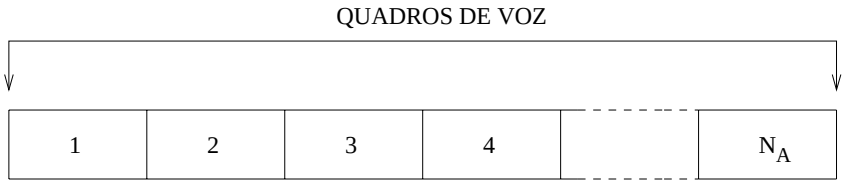


Figura 5.2: Sinal de voz segmentado.

descontinuidade entre segmentos. Em cada bloco de amostras é feito um janelamento, para minimizar os efeitos adversos resultantes da segmentação abrupta que causa descontinuidades no espectro do sinal de voz [109].

Os tipos de janela utilizados são: Janela Retangular, Janela de Hamming e Janela de Hanning, cujas características são mostradas a seguir [109].

- Janela Retangular

$$J(n) = \begin{cases} 1 & , 0 \leq n \leq N_A - 1 \\ 0 & , \text{caso contrário} \end{cases} \tag{5.3}$$

- Janela de Hamming

$$J(n) = \begin{cases} 0,54 - 0,46\cos[2\pi n/(N_A - 1)] & , 0 \leq n \leq N_A - 1 \\ 0 & , \text{caso contrário} \end{cases} \tag{5.4}$$

- Janela de Hanning

$$J(n) = \begin{cases} 2a\cos[\pi n/N_A] + b & , 0 \leq n \leq N_A - 1 \\ 0 & , \text{caso contrário} \end{cases} \tag{5.5}$$

em que $2a + b = 1$ (com $0 \leq a \leq 0,25$ e $0,5 \leq b \leq 1$).

Se o sinal é, simplesmente, particionado em blocos consecutivos, então está sendo aplicada, implicitamente, uma janela retangular de comprimento igual ao comprimento do bloco. Entretanto, como resultado no domínio da freqüência, surgem fugas espectrais alterando o espectro do sinal. Para evitar este efeito indesejável no domínio da freqüência, utilizam-se janelas de Hamming ou Hanning, definidas no domínio do tempo, que proporcionam, no domínio da freqüência, um lóbulo principal de amplitude bastante superior a dos lóbulos secundários, diminuindo, portanto, o efeito destes. O efeito do janelamento de Hamming é a manutenção das características espectrais do centro do quadro e a eliminação das transições abruptas das extremidades.

A janela de Hanning assemelha-se a de Hamming porém, proporciona um reforço menor nas amostras do centro e uma suavização maior nas amostras da extremidade.

A janela de Hamming porém apresenta uma característica nem sempre desejável, que corresponde à atribuição de um peso muito baixo às amostras da extremidade. Entretanto, estas amostras podem representar eventos importantes de curta duração do sinal de voz e multiplicá-los por um peso baixo representa pouca atenção no processamento subseqüente realizado a nível de blocos. Para assegurar que a tais eventos seja dado o peso necessário, blocos adjacentes são sobrepostos de modo que um evento seja "coberto" por outros blocos.

Para o contexto da produção da voz, a característica do janelamento de Hamming se mostra, portanto, mais eficiente, quando comparada aos outros tipos de janela (Retangular e Hanning), com uma boa aproximação da janela ideal. Assim sendo, essa foi a janela utilizada neste trabalho.

5.3 Extração de características

Para cada segmento de fala janelado, é realizada, inicialmente, a estimação (deteção) da freqüência fundamental, visando separar os locutores em grupos gerais, de acordo com o sexo (pré-identificação).

Após a deteção da freqüência fundamental, é extraído um conjunto de K coeficientes, a partir da Análise por Predição Linear, obtendo-se um conjunto para cada tipo de coeficiente: LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados (descritos no Capítulo 3). Em seguida, é realizada uma análise comparativa do desempenho desses coeficientes, de forma a determinar qual(is) irá(ao) compor o vetor de características de cada locutor, tanto para a fase de treinamento quanto de identificação.

5.3.1 Deteção da Freqüência Fundamental

Para deteção da freqüência fundamental, como descrito no Capítulo 3, faz-se necessário a separação dos sons sonoros da voz, pois a mesma só existirá nos intervalos da fala que contêm esses sons. Portanto, antes da estimação de F_0 é necessário a implementação do detetor surdo-sonoro.

Após ter sido calculado F_0 (ou P_0) em cada bloco de amostras, calcula-se então a Freqüência Fundamental média (F_0 *média*), correspondente à elocução completa. Esta medida corresponde à média aritmética das F_0 obtidas ao longo dos segmentos sonoros.

Objetivando melhorar o desempenho do sistema, calculou-se uma medida de dispersão relativa (Coeficiente de Variação - C.V.), que corresponde à razão entre o desvio padrão e a média de uma distribuição, expressa em percentual. O C.V. mede, portanto, o grau da dispersão entre a média e os valores de uma distribuição [110]. Neste trabalho, esta medida foi utilizada para fornecer a dispersão entre a F_0 *média* de cada locutor (para cada elocução da sentença) e os valores de F_0 em cada bloco de amostras do sinal de voz. Para as elocuções em que o Coeficiente de Variação é maior ou igual a 40%, indicando a existência de uma grande variabilidade dos valores de F_0 ao longo dos quadros, é solicitado ao locutor que repita a sua sentença.

A regra de decisão utilizada para definir o locutor como masculino ou feminino baseia-se em dois limiares (definidos de forma empírica) expressos da seguinte forma:

- Se $F_0 \leq 175$ Hz, o locutor é classificado como Masculino;
- Se $F_0 > 175$ Hz, o locutor é classificado como Feminino.

5.3.2 Obtenção do vetor de características

A análise LPC de uma elocução de voz de treinamento, irá fornecer o vetor de características a ser utilizado, na fase de treinamento, para obtenção dos dicionários do quantizador vetorial e da seqüência de observações $\mathbf{O^l} = \{O_1, ..., O_T\}$ que irá gerar o HMM de referência associado ao l-ésimo locutor ($1 \leq l \leq L$) e na fase de identificação para determinação da distorção mínima associada ao dicionário e da probabilidade máxima associada ao HMM.

Como uma limitação prática e computacional geralmente é desejável usar, principalmente em quantização vetorial, um número mínimo de parâmetros para modelar precisamente as características significativas do sinal de voz. A ordem do filtro preditor para o modelo LPC está diretamente relacionada à precisão desejada em relação ao trato vocal e depende da freqüência de amostragem escolhida para representar o sinal de voz. A ordem do filtro deve ser escolhida de maneira a se obter uma boa representação de todos os formantes presentes no sinal. De [1, 65] tem-se que

$$K \cong \frac{f_s}{1000}, \tag{5.6}$$

sendo K a ordem do filtro preditor e f_s a freqüência de amostragem. A Equação (5.6) indica que deve existir pelo menos uma seção cilíndrica sem perdas no modelo do tubo acústico que representa o trato vocal, para cada kHz da freqüência de amostragem [1, 51]. Assim, sendo K uma aproximação, com $f_s = 11$ kHz, foi escolhido $K = 12$.

Após a obtenção do conjunto de 12 coeficientes (LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados), um conjunto para cada tipo de coeficiente, é realizada a análise comparativa do desempenho desses coeficientes na tarefa de identificação, de forma a definir qual(is) o(s) tipo(s) de coeficiente(s) que irá(ão) compor o vetor de características.

5.4 Quantização Vetorial

A quantização vetorial é utilizada para a geração dos dicionários e da tabela de códigos (símbolos) associada à seqüência de observações $\mathbf{O^l} = \{O_1, ..., O_T\}$, para cada l-ésimo locutor.

5.4.1 Projeto do dicionário

Para o projeto do dicionário do quantizador vetorial (um para cada l-ésimo locutor, $1 \leq l \leq L$), são avaliados os algoritmos LBG, KMVVT e SSC (descritos no Capítulo 4). Em seguida, é realizada uma análise comparativa do desempenho desses algoritmos no projeto do dicionário, de forma a determinar qual o que melhor reproduz as características vocais de cada locutor.

O dicionário inicial é constituído a partir de um conjunto de amostras iniciais da seqüência de treino, tomadas de forma aleatória.

A seqüência de voz (seqüência de treino) utilizada para geração do dicionário do quantizador vetorial é formada a partir de cinco elocuções da sentença.

A sentença de treino foi gravada em uma sessão diferente da utilizada para a gravação das sentenças de teste.

5.4.2 Medida de distorção

A medida de distorção utilizada no quantizador vetorial é a medida de Distância do Erro Médio Quadrático, cujas características foram descritas no Capítulo 4 [51].

5.4.3 Escolha da dimensão do quantizador

Os parâmetros escolhidos para representar o sinal de voz, como dito no Capítulo 3, são os coeficientes obtidos a partir da Análise por Predição Linear [2, 35, 51]. A ordem do preditor ($K = 12$), neste trabalho, corresponde à dimensão do quantizador.

5.4.4 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), em trabalho anterior [48], visando a obtenção de um resultado que melhor se adaptasse às características do sistema. Para tanto, foi escolhido $M = 64$.

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 $M = 128$, 256, 1024 ou 2048. Estes valores de M poderiam fornecer melhores resultados que os obtidos para $M = 64$, entretanto isso levaria a um aumento considerável do volume de dados, acarretando um custo computacional elevado, sem proporcionar um aumento significativo na eficiência do sistema.

5.5 Modelagem utilizando HMM

A modelagem, utilizando HMM, é levada a efeito nas etapas de treinamento e identificação, descritas a seguir.

Treinamento: obtenção do HMM de referência, $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$, para cada l-ésimo locutor ($1 \leq l \leq L$), utilizando o processo iterativo de Baum-Welch (descrito no Capítulo 4) [23, 37].

O critério de parada utilizado no processo de reestimação é o seguinte: se a diferença entre a probabilidade associada ao modelo atual e o modelo anteriormente estimado, for menor que um dado limiar (10^{-3}) ou se o número de iterações ultrapassar um determinado valor (> 500), o processo de reestimação é finalizado. Esses valores foram determinados de forma empírica, como também pela observação dos limiares utilizados em outros trabalhos na área [23, 24].

Identificação: determinação da probabilidade de ocorrência de uma dada seqüência de observações $\mathbf{O}^l = \{O_1, O_2, ..., O_T\}$, associada a cada um dos modelos $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$ já obtidos durante a fase de treinamento. Para o cálculo da probabilidade foi utilizado também o algoritmo de Viterbi (Capítulo 4) [23, 61, 69, 99, 103].

Para a modelagem utilizando HMM faz-se necessária a descrição da forma de determinação dos parâmetros do modelo, bem como das considerações de implementação, apresentadas a seguir.

5.5.1 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 [111]. Dessa forma, cada estado corresponderia a um intervalo de observação, ou seja, em torno de 20 ms para o sistema ora apresentado. Essa opção também implicaria na utilização de um número de estados bastante elevado, em torno de 60. Como o propósito deste trabalho não é o reconhecimento de palavras isoladas e sim a identificação de locutor dependente do texto, não é necessário verificar a variação dos sons de cada palavra, mas a forma como cada locutor a pronuncia. Portanto, a partir da bibliografia disponível, [4, 23, 24, 70, 95, 97, 101] e para evitar um número de cálculos elevado, optou-se pela utilização de $N = 5$, uma vez que a redução do erro para valores de $N > 5$ não é significativa, para o propósito deste trabalho.

5.5.2 Inicialização de $\mathbf{a}_{ij}$

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} [23, 37]. A maioria das estimativas é obtida através do método de "tentativas", verificando-se qual delas produz o melhor resultado. A partir de um conjunto de estimativas apresentadas em [24, 37, 95, 97] e

fazendo-se vários testes, utilizou-se, neste trabalho, a seguinte matriz de transição de estados:

$$A = \begin{vmatrix} 0.8 & 0.1 & 0.1 & 0.0 & 0.0 \\ 0.0 & 0.9 & 0.1 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.1 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.9 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{vmatrix} \tag{5.7}$$

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

5.5.3 Inicialização de $\mathbf{b}_j(\mathbf{k})$

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 [23, 24]. Todos esses 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 [24, 95].

$$B = \begin{vmatrix} 1/M & 1/M & 1/M & ... & 1/M \\ 1/M & 1/M & 1/M & ... & 1/M \\ 1/M & 1/M & 1/M & ... & 1/M \\ 1/M & 1/M & 1/M & ... & 1/M \\ 1/M & 1/M & 1/M & ... & 1/M \end{vmatrix} \tag{5.8}$$

5.5.4 Uso de múltiplas seqüências de observações

O maior problema associado ao HMM do tipo "esquerda-direita", como dito anteriormente (Capítulo 4), reside no fato de que não se pode usar uma única seqüência de observações para treinar o modelo (isto é, para a reestimação dos parâmetros do modelo) [23, 61]. Assim, a fim de se obter dados suficientes para se fazer estimativas confiáveis de todos os parâmetros do modelo, deve-se utilizar múltiplas seqüências de observações.

Neste trabalho, foram feitos alguns testes para a determinação do valor ótimo de U (número de seqüências de observações) e, a partir do critério "custo-benefício", ou seja, desempenho e custo computacional, optou-se pelo uso de $U = 10$.

5.5.5 Considerações de implementação

A implementação dos algoritmos de HMM (descritos no Capítulo 4), requer alguns cuidados especiais de forma a se obter 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 (4.15) e (4.21) (Capítulo 4), é fácil verificar que $T \rightarrow \infty$ (ou T muito grande, i.e., 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 a probabilidade irá resultar em *underflow* se as Equações (4.15) e (4.21) 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 [23].

O segundo problema é mais grave e mostra que sérias dificuldades de identificação irão ocorrer se algum elemento de $b_j(k)$ assume um valor zero durante a fase de treinamento [23, 24]. Isso ocorre porque a fase de identificação envolve a computação de $P_l(\mathbf{O}^l/\lambda_\mathbf{l})$ e 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 $\mathbf{O}_t = w_k$, então a probabilidade da seqüência associada ao modelo com $b_j(k) = 0$ é $P_l = 0$; assim um erro de identificação deverá ocorrer.

O último 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 $\sum_{k=1}^{M} b_j(k) = 1$. Dessa forma, todos os $b_j(k)$ são comparados com o limiar ϵ e aquele que fica abaixo de ϵ é substituído por ϵ para cada j. Após essa substituição, cada $b_j(k)$ que não foi modificado pelo valor ϵ é reescalonado pela quantidade $1 - R_{bj}\epsilon$ (R_{bj} é o número de $b_j(k)$ 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 locutor, fornecendo baixas taxas de erro [23, 24, 49]. Neste trabalho foram realizados alguns testes, optando-se pelo uso de $\epsilon = 10^{-6}$.

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 a_{ij} é realizado o mesmo procedimento de reescalonamento feito para os $b_j(k)$.

Escalonamento

Para entender porque o escalonamento é requerido para a técnica de reestimação de HMMs, deve-se considerar a definição de $\alpha_t(i)$ (Equação (4.15), Capítulo 4). Pode ser visto que $\alpha_t(i)$ consiste de uma soma de um grande número de termos, cada um da forma

$$(\prod_{s=1}^{t-1} a_{q_s q_{s-1}} \prod_{s=1}^{t} b_{q_s}(O_s)) \tag{5.9}$$

com $q_t = q_1$. Sendo cada a e b muito menor que 1 (geralmente significativamente menor que 1), pode-se ver que à medida que t torna-se grande (i.e., 10 ou mais), cada termo de $\alpha_t(i)$ tenderá exponencialmente para zero. Para t suficientemente grande (i.e., 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 esse 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.e., que 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

$$esc_t = (\sum_{i=1}^{N} \alpha_t(i))^{-1}. \tag{5.10}$$

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 é possível meramente somar os termos $\alpha_T(i)$ desde que estes tenham sido escalonados anteriormente. Entretanto, pode-se usar a propriedade

$$\prod_{t=1}^{T} esc_t \sum_{i=1}^{N} \alpha_T(i) = C_T \sum_{i=1}^{N} \alpha_T(i) = 1. \tag{5.11}$$

Assim, tem-se

$$\prod_{t=1}^{T} esc_t P_l(\mathbf{O^l}/\lambda_\mathbf{l}) = 1 \tag{5.12}$$

ou

$$P_l(\mathbf{O^l}/\lambda_\mathbf{l}) = \frac{1}{\prod_{t=1}^{T} esc_t} \tag{5.13}$$

ou

$$log[P_l(\mathbf{O^l}/\lambda_\mathbf{l})] = -\sum_{t=1}^{T} log(esc_t). \tag{5.14}$$

O *log* de $P_l(\mathbf{O^l}/\lambda_\mathbf{l})$ pode ser computado, mas $P_l(\mathbf{O^l}/\lambda_\mathbf{l})$ não, visto que o cálculo de $P_l(\mathbf{O^l}/\lambda_\mathbf{l})$ poderia levar a um resultado fora da faixa dinâmica da máquina [23].

5.6 Padrões de Referência e de Teste

Após o processamento do sinal, extração dos parâmetros representativos dos locutores (coeficientes obtidos a partir da análise por predição linear), a quantização vetorial e construção do HMM, obtém-se os padrões de referência, um dicionário e um $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$, para cada l-ésimo locutor ($1 \leq l \leq L$).

Durante a etapa de reconhecimento (identificação), é realizada, inicialmente, a pré-identificação dos locutores (pré-classificação em grupos gerais de acordo com o sexo, utilizando a freqüência fundamental). Em seguida, é obtido o padrão de teste do locutor a ser identificado pelo sistema, que corresponde ao vetor de características.

5.7 Regra de Decisão

A regra de decisão utilizada é dividida em duas, da seguinte forma:

Regra de decisão 1 – utiliza o método paramétrico, sendo feita a comparação do padrão de teste (vetor de características) com todos os padrões de referência (dicionários do quantizador vetorial) e o padrão de referência que proporcionar a menor distorção (desde que menor que um dado limiar) corresponde ao padrão do locutor identificado. O limiar é utilizando visando evitar a falsa aceitação de locutores não cadastrados, quando o sistema for implementado para um conjunto aberto.

Não é utilizado um único limiar para todos os tipos de coeficientes. Foi escolhido um limiar para os coeficientes LPC, outro para os coeficientes Cepestrais e Delta Cepestrais

e um terceiro limiar para os coeficientes Cepestrais Ponderados e Delta Cepestrais Ponderados. Essa decisão, bem como a escolha dos limiares, foi tomada a partir de resultados experimentais.

Regra de decisão 2 – utiliza o método estatístico. O padrão de teste (conjunto de símbolos obtido a partir do vetor de características) irá compor a seqüência de observações. Em seguida, será realizado o cálculo da probabilidade desta seqüência ter sido gerada por cada um dos $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$, $1 \leq l \leq L$, obtidos durante a fase de treinamento. O modelo que proporcionar o maior valor de probabilidade, corresponderá ao modelo do locutor identificado pelo sistema. Entretanto, essa regra só é realizada se a comparação do padrão de teste com os padrões de referência, da regra de decisão anterior, indicar "similaridade" entre os padrões vocais dos locutores. Ou seja, se a diferença entre os valores de distorção dos locutores for menor que um dado limiar. A escolha do limiar foi determinada de forma empírica. Sendo assim, a modelagem utilizando HMM se caracteriza como uma etapa de "refinamento" do processo de identificação da identidade vocal dos locutores.

5.8 Discussão

O sistema de reconhecimento (identificação) automático da identidade vocal desenvolvido neste trabalho é constituído das etapas: processamento do sinal, extração de características, quantização vetorial, modelagem utilizando HMM e regra de decisão.

O processamento do sinal inclui a aquisição, pré-ênfase e janelamento (janela de Hamming) em blocos de amostras de voz, de forma a garantir a extração de características acústicas representativas dos locutores.

A extração de características corresponde à obtenção de dois tipos de parâmetros. Inicialmente é feita a estimação da freqüência fundamental, que visa separar os locutores em grupos gerais, de acordo com o sexo (pré-identificação), diminuindo o volume de dados a ser avaliado no processo de identificação e minimizando a taxa de erro do sistema, pois impede que locutores masculinos sejam identificados como femininos e vice-versa. Após a estimação da freqüência fundamental, é extraído um conjunto de 12 coeficientes (LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados). Em seguida, é realizada a análise comparativa do desempenho desses coeficientes na tarefa de identificação, de forma a definir qual(is) irá(ão) compor o vetor de características, para as fases de treinamento e identificação.

O vetor de características será utilizado, na fase de treinamento, para obtenção dos dicionários do quantizador vetorial e, conseqüentemente, da seqüência de observações $\mathbf{O^l} = \{O_1, ..., O_T\}$ que irá gerar o HMM, associado a cada locutor (HMM de referência) e na fase de identificação para determinação da distorção mínima, em relação aos dicionários e da probabilidade máxima associada ao HMM.

Para o projeto do dicionário, do quantizador vetorial, são utilizados os algoritmos LBG, KMVVT e SSC, sendo realizada uma análise comparativa de forma a determinar o mais eficiente para a modelagem das características vocais dos locutores.

Para obtenção do HMM de referência, um para cada locutor, utiliza-se o processo iterativo de *Baum-Welch*. Sendo importante o uso de múltiplas seqüências de observações para garantir a obtenção de estimativas confiáveis de todos os parâmetros do modelo. Algumas considerações acerca da implementação dos HMMs devem ser tomadas afim de se evitar problemas de indeterminação na resolução das equações do modelo.

Os padrões de referência obtidos correspondem, portanto, a um dicionário (obtido a partir da QV) e a um $\lambda = (\mathcal{A}, \mathcal{B}, \pi)$ para cada locutor. Durante a etapa de identificação, é realizada, inicialmente, a pré-identificação dos locutores (utilizando a freqüência fundamental). Em seguida, é obtido o padrão de teste do locutor a ser identificado pelo sistema, que corresponde ao vetor de características.

A regra de decisão é dividida em duas. Na primeira, é feita a comparação do padrão de teste com todos os padrões de referência (dicionários do quantizador vetorial), o que proporcionar a menor distorção, desde que menor que um dado limiar, correponde ao padrão do locutor identificado. O limiar foi utilizando visando evitar a falsa aceitação de locutores não cadastrados quando o sistema for implementado para um conjunto aberto. Na segunda regra de decisão, o padrão de teste (conjunto de símbolos do vetor de características) irá compor a seqüência de observações. Em seguida, será realizado o cálculo da probabilidade desta seqüência ter sido gerada por cada um dos $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$ obtidos durante a fase de treinamento. O modelo que proporcionar o maior valor de probabilidade, corresponderá ao modelo do locutor identificado pelo sistema. Essa regra de decisão só é realizada se a comparação do padrão de teste com os padrões de referência, da regra anterior, indicar "similaridade" entre os locutores. Assim, a modelagem utilizando HMM corresponde a uma etapa de "refinamento" do sistema de identificação da identidade vocal de locutores.

Capítulo 6

Apresentação e Análise dos Resultados

6.1 Introdução

O sistema de reconhecimento, desenvolvido neste trabalho, descrito na Figura 5.1 (Capítulo 5), se caracteriza por ser um sistema híbrido (utiliza métodos paramétrico e estatístico) para o reconhecimento (identificação) automático da identidade vocal de locutores, composto de duas etapas: pré-identificação e identificação. A primeira etapa utiliza a freqüência fundamental como parâmetro de separação prévia dos locutores em grupos gerais, de acordo como sexo. A etapa de identificação utiliza a Análise por Predição Linear, a Quantização Vetorial Paramétrica e a modelagem por Modelos de Markov Escondidos, para a obtenção dos padrões representativos dos locutores, que os distinguem entre si.

6.2 Apresentação e Análise dos Resultados

O sistema de identificação automática de locutor (implementado em linguagem C), consiste de duas fases, descritas a seguir: fase de treinamento e fase de identificação.

- Fase de Treinamento (Figura 6.1):

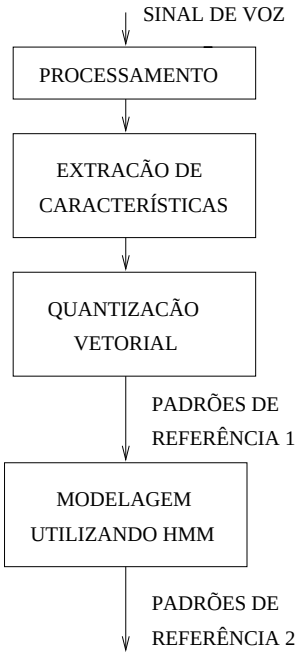


Figura 6.1: Fase de treinamento do Sistema de Identificação Automática de locutor.

1. Elocução da sentença de treinamento.

2. Processamento do Sinal: aquisição do sinal, pré-ênfase e janelamento em T blocos, cada bloco contendo N_A amostras.

3. Extração dos parâmetros da fala, a partir da elocução da sentença de treinamento, de forma a obter o vetor de características, utilizando a análise por predição linear (sendo realizada a anaálise comparativa do desempenho dos coeficientes LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados, na representação das características vocais dos locutores), que irá compor o vetor de coeficientes, a ser utilizado na Quantização Vetorial Paramétrica. Para cada tipo de coeficiente é obtido um vetor de características.

4. Quantização Vetorial: Quantização Vetorial Paramétrica que permite a obtenção o primeiro padrão de referência (dicionário) para cada l-ésimo locutor.

5. Modelagem utilizando HMM: estimação do HMM, obtendo-se o modelo $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi)$, que corresponde ao segundo padrão de referência para cada l-ésimo locutor.

- Fase de Identificação (Figura 6.2):

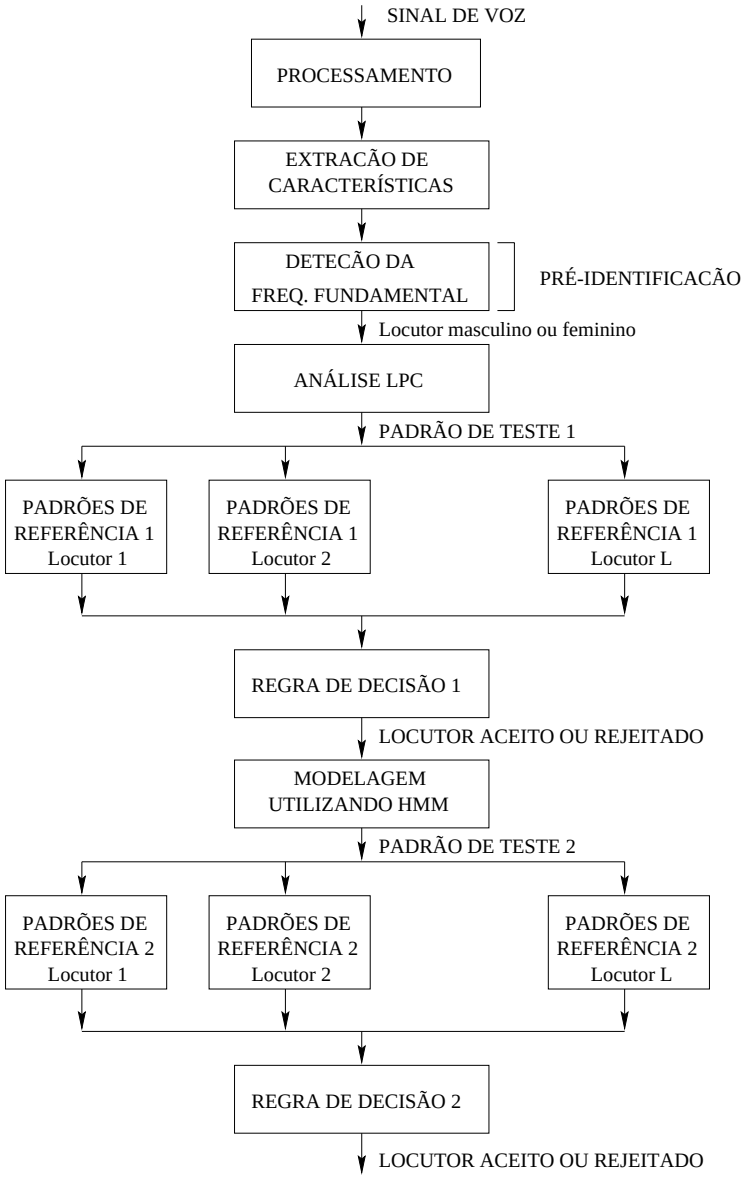


Figura 6.2: Fase de identificação do Sistema de Identificação Automática de locutor.

1. Elocução da sentença de teste.

2. Processamento do Sinal: aquisição do sinal, pré-ênfase e janelamento em T blocos, cada bloco contendo N_A amostras.

3. Extração de características: extrai parâmetros da fala, obtendo-se dois tipos de características, descritas a seguir.

3.1. Freqüência Fundamental: realiza a pré-identificação dos locutores, separando-os em dois grupos de acordo com o sexo (masculino e feminino);

3.2. Análise por Predição Linear: permite a obtenção do vetor de características (análise comparativa do desempenho dos coeficientes LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados), que irá compor o padrão de teste 1 do locutor.

4. Regra de decisão 1: compara o padrão de teste com os padrões de referência associados aos L locutores do sistema (dicionários obtidos durante a fase de treinamento). Essa regra irá determinar se o locutor será aceito ou rejeitado pelo sistema ou se será necessário prosseguir para a próxima etapa.

5. Modelagem utilizando HMM: se o valor de distorção obtido na regra de decisão 1 indicar similaridade entre as características vocais dos locutores (diferença entre as distorções menor que um limiar), cálcula-se a probabilidade de ocorrência de uma dada seqüência de observações $\mathbf{O}^l = \{O_1, O_2, ..., O_T\}$ (vetor de teste 2), associada a cada um dos modelos $\lambda_l = (\mathcal{A}, \mathcal{B}, \pi), 1 \leq l \leq L$, já obtidos durante a fase de treinamento.

6. Regra de decisão 2: verifica qual o padrão de referência que proporciona o maior valor de probabilidade, este corresponderá ao padrão do locutor identificado pelo sistema.

6.2.1 Parâmetros para Avaliação do Desempenho

Para avaliação do desempenho, em sistemas de reconhecimento de padrões, quatro parâmetros são comumente utilizados [112, 113, 114]: reconhecimento, erro, rejeição e confiabilidade. Como o reconhecimento de locutor se constitui em uma tarefa de reconhecimento de padrões, esses parâmetros também serão avaliados neste trabalho, conforme descrição a seguir.

- Reconhecimento (Identificação): a taxa de reconhecimento ou a taxa de sucesso é a porcentagem de locutores classificados (identificados) corretamente.

- Erro (Falsa aceitação): porcentagem de locutores classificados (identificados) erroneamente pelo sistema.

- Rejeição (Falsa rejeição): porcentagem de locutores cadastrados que não são identificados pelo sistema.

- Confiabilidade: corresponde à razão percentual entre o número de locutores classificados corretamente e o total de locutores classificados pelo sistema (seja a classificação correta ou não). Essa taxa pode ser obtida da seguinte forma:

$$\text{Confiabilidade} = \frac{\text{Reconhecimento}}{\text{Reconhecimento} + \text{Erro}} \tag{6.1}$$

Em virtude dos parâmetros utilizados para análise de desempenho corresponderem à taxas médias, foi introduzido um parâmetro adicional para essa análise, o coeficiente de variação, C.V. (razão percentual entre o desvio padrão e a média de uma distribuição), visto que o C.V. mede o grau de variabilidade dos dados em torno do valor médio de uma distribuição. Altos valores de C.V. indicam, portanto, que o valor médio obtido não é representativo (dispersão acentuada) [110]. Neste trabalho foi calculado o C.V. apenas para a taxa média de identificação, não sendo necessário calculá-lo para as demais taxas (pois essas são obtidas em relação à taxa média de identificação).

A seguir são apresentados os resultados, bem como os parâmetros utilizados para avaliação de desempenho, correspondentes às etapas de pré-identificação e identificação, respectivamente.

6.2.2 Pré-identificação dos locutores

Para classificação dos locutores em grupos gerais, de acordo com o sexo (pré-identificação), torna-se necessário a utilização do detetor surdo-sonoro, seguido da estimação da freqüência fundamental. A seguir, são apresentados os resultados associados ao detetor surdo-sonoro e ao estimador da freqüência fundamental, respectivamente.

6.2.2.1 Detetor Surdo-Sonoro

Para implementação do sistema foi utilizada uma *amostra* formada por locutores adultos dos sexos masculino e feminino.

O sistema foi testado, inicialmente, apenas com palavras isoladas, utilizando 5 locutores femininos e 5 locutores masculinos. Cada locutor pronunciou as palavras *aplausos* e *bola*, em sessões diferentes, 5 vezes.

A palavra *aplausos* foi escolhida por apresentar uma combinação de sons sonoros, surdos e explosivos e *bola* por apresentar a seguinte característica: iniciar com um som explosivo, diferentemente de *aplausos* que se inicia por um som sonoro.

Os resultados do detetor surdo-sonoro, para a palavra *aplausos*, estão na Tabela A.1 (Anexo A), na qual são apresentados os parâmetros temporais utilizados e a conseqüente decisão tomada pelo detetor.

Verifica-se através da Tabela A.1, como era de se esperar, que o sinal apresenta valores de energia mais altos nos intervalos sonoros e valores inferiores para os sons surdos e silêncio. Os valores do coeficiente de correlação normalizado para os sons sonoros, são próximos da unidade enquanto que, para os sons surdos e silêncio assumem valores muito pequenos. A taxa de cruzamento por zero assume valores muito elevados para os sons surdos e valores menores para os sons sonoros. O número total de picos se comporta de forma semelhante ao coeficiente de correlação, para os diversos sons, facilitando assim, uma melhor deteção desses sons.

Pode-se concluir, portanto, que o detetor surdo-sonoro se mostra eficiente na determinação dos sons da voz, visto que é capaz de detetar, de forma eficaz, os diversos sons que compôem as palavras, bem como as sentenças utilizadas neste trabalho. É importante destacar que o bom desempenho do detetor está diretamente associado ao ajuste dos limiares de decisão (energia, número total de picos, diferença de picos, taxa de cruzamento por zero e o coeficiente de correlação normalizado), que depende da forma de aquisição do sinal. Portanto, não existe um valor "ótimo" para os limiares que possa ser utilizado em qualquer ambiente, existem sim, valores indicados de forma bastante genérica. Torna-se necessário, portanto, quando da elaboração do sistema, a realização de testes empíricos (dentro dos valores indicados), que permitam determinar os valores para os limiares que melhor se adaptam à aplicação. Essa foi a técnica utilizada neste trabalho.

6.2.2.2 Estimação da Freqüência Fundamental

Para análise de desempenho do detetor da freqüência fundamental, cada locutor pronunciou, inicialmente, as palavras *aplausos* e *bola*, 5 vezes cada uma. Dessa forma, cada locutor pronunciou um total de 10 elocuções.

A Figura 6.3 e a Tabela A.2 (Anexo A) apresentam os valores médios de F_0 dos locutores masculinos e femininos, para as cinco elocuções da palavra *aplausos*. De uma

forma geral, existe pouca variação da F_0 média para cada locutor, com uma taxa de erro de 0%, exceto para o locutor LF2, que é classificado como masculino em uma das elocuções da palavra (erro de 20%). Para todos os locutores os valores do C.V. estão abaixo de 15%, indicando portanto, que a freqüência fundamental média é representativa para cada locutor [110].

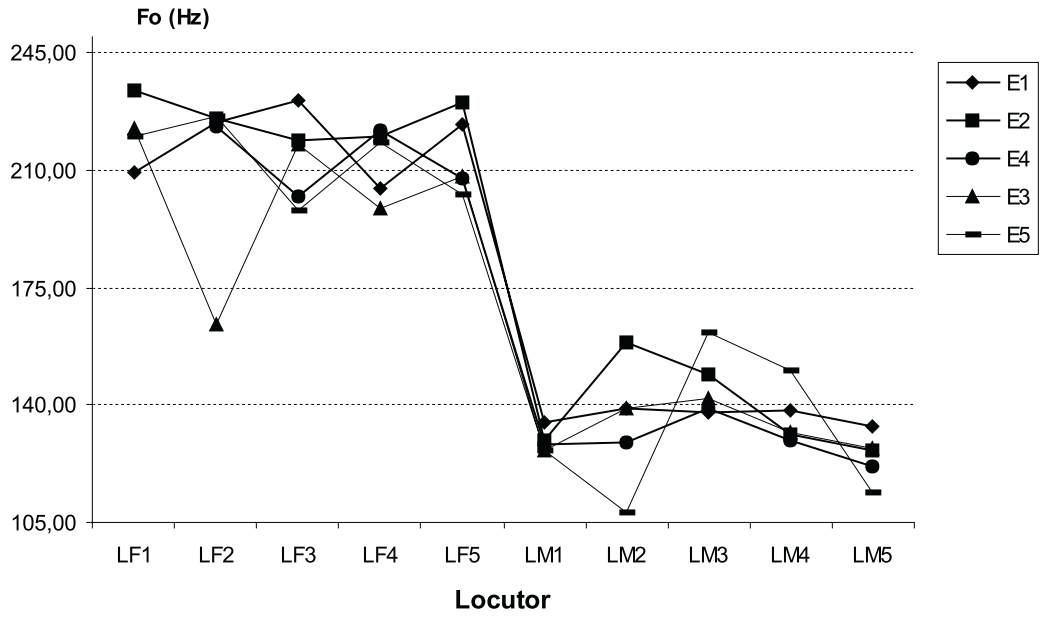


Figura 6.3: Freqüência Fundamental dos locutores femininos (LF1 a LF5) e masculinos (LM1 a LM5), para as cinco elocuções da palavra *aplausos* (E1 a E5).

A Figura 6.4 e a Tabela A.3 (Anexo A) apresentam os valores médios de F_0 dos locutores masculinos e femininos, para as cinco elocuções da palavra *bola*. Existe, de uma forma geral, pouca variação da F_0 média para alguns locutores. Entretanto, os locutores LF2 e LF5, LM2 e LM4, são classificados erroneamente como masculinos e femininos, respectivamente, em algumas elocuções da palavra (erros de 20%, 20%, 40% e 20%, respectivamente). Para a maioria dos locutores, os valores do coeficiente de variação estão abaixo de 15% (exceto para os locutores LF5, LM2 e LM4), indicando portanto, que a freqüência fundamental média desses locutores é representativa.

As taxas de erro obtidas, que variam entre 20% e 40% ocorrem, principalmente, devido às variações na forma de elocução dos locutores durante a pronúncia de cada palavra [45], ao tamanho da palavra que pode não ser suficiente para a completa

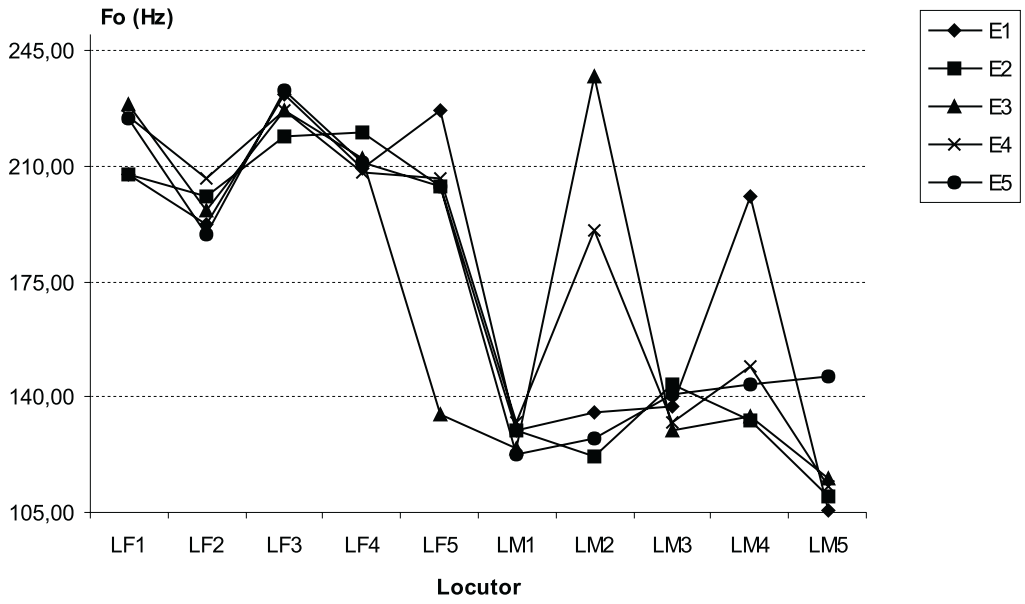


Figura 6.4: Freqüência Fundamental dos locutores femininos (LF1 a LF5) e masculinos (LM1 a LM5), para as cinco elocuções da palavra *bola* (E1 a E5).

identificação da F_0, ao número de elocuções utilizadas, ao ruído no ambiente de gravação, ao erro associado aos algoritmos de deteção surdo-sonoro e estimação da F_0, à sensibilidade do algoritmo no ajuste de alguns parâmetros, etc.

Em uma segunda etapa, o algoritmo de estimação da F_0 foi analisado utilizando palavras conectadas (sentenças). Foram utilizados 4 locutores masculinos e 4 femininos. A quantidade de locutores diminuiu em virtude da disponibilidade dos mesmos para as sessões de gravação. Para cada locutor foi atribuída uma sentença, com duração em torno de 2 segundos. As sentenças são: *Quero usar a máquina* (locutor feminino 1 - LF1), *Vou ganhar o prêmio* (locutor feminino 2 - LF2), *Eu terei a vitória* (locutor feminino 3 - LF3), *Minha senha é secreta* (locutor feminino 4 - LF4), *O dia será bonito* (locutor masculino 1 - LM1), *Farei o máximo hoje* (locutor masculino 2 - LM2), *A luta foi rápida* (locutor masculino 3 - LM3) e *Meu login está correto* (locutor masculino 4 - LM4). Foram utilizadas essas sentenças porque apresentam uma combinação dos sons sonoros, surdos, explosivos e silêncio. Além disso, são fáceis de pronunciar, facilitando assim as elocuções.

Para verificar as taxas de erro (locutor feminino é considerado masculino ou vice-versa), cada locutor pronunciou sua sentença dez vezes e as sentenças dos demais locutores cinco vezes. Dessa forma, cada locutor pronunciou um total de quarenta e cinco elocuções.

As Figuras 6.5 e 6.6 e a Tabela A.4 (Anexo A) apresentam os valores médios da F_0 dos locutores masculinos e femininos, para as 45 elocuções de todas as sentenças.

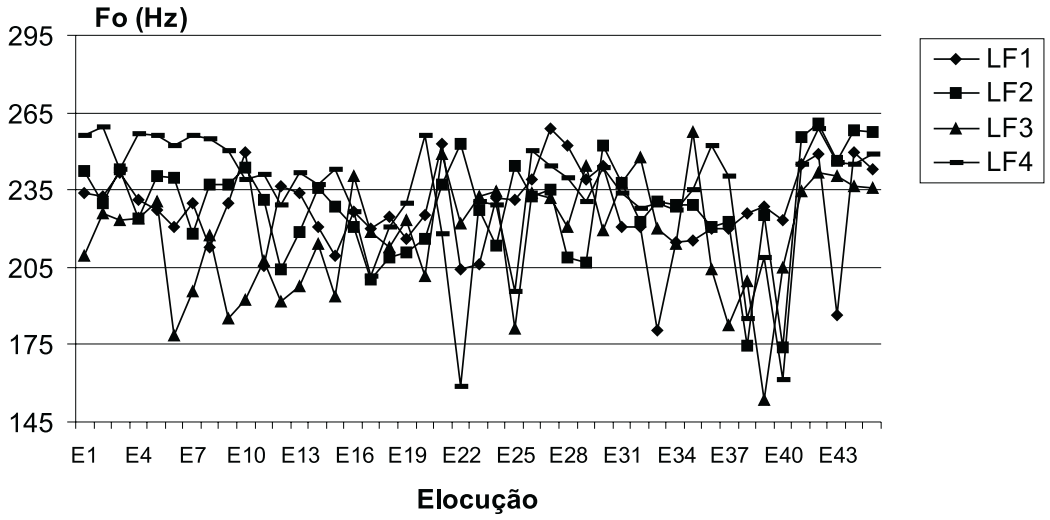


Figura 6.5: Freqüência Fundamental dos locutores femininos (LF1 a LF4), para as 45 elocuções de todas as sentenças (E1 a E45).

Os resultados mostram que, de uma forma geral, existe pouca variação da F_0 média para um mesmo locutor (exceto para os locutores LF3 e LM4). A maior parte dos locutores apresenta valores do C.V. abaixo de 15%, indicando portanto, que a freqüência fundamental média é representativa para cada locutor, exceto para o locutor LM4.

As variações ocorrem com os locutores cujas sentenças apresentam uma quantidade considerável de sons surdos, demonstrando assim uma certa dificuldade do algoritmo, de estimação da F_0, na separação desses sons. Além disso, esses locutores apresentam também entonações diferenciadas na elocução das várias sentenças, indicando a sensibilidade da F_0 às variações de entonação, fato já observado em [2]. É importante destacar que essas variações não foram tão elevadas.

Para a maior parte dos locutores a taxa de erro está abaixo de 10%, exceto para os

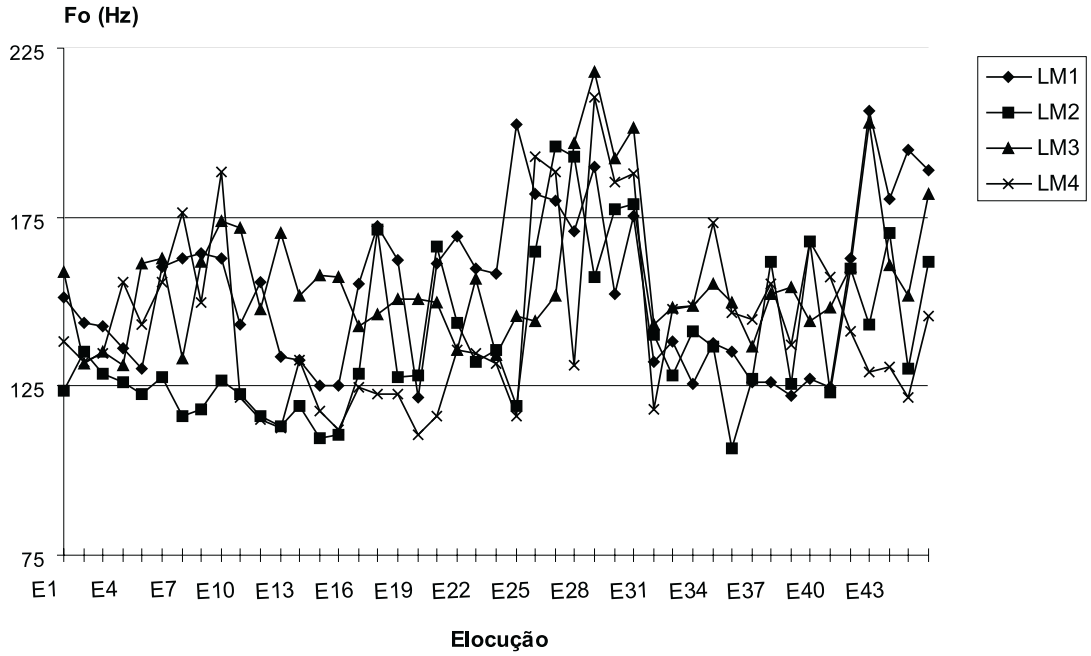


Figura 6.6: Freqüência Fundamental dos locutores masculinos (LM1 a LM4), para as 45 elocuções de todas as sentenças (E1 a E45).

locutores LM1, LM2 e LM4. Para alguns locutores existe variabilidade considerável entre os valores médios de F_0, para cada elocução (C.V. $> 15\%$). Porém, o resultado final mostra que o valor médio da F_0 de cada locutor está dentro dos limites determinados para locutores femininos e masculinos.

Em algumas elocuções é solicitado ao locutor que repita a sua sentença pois o C.V. está acima do limiar permitido ($> 40\%$), indicando a existência de uma grande variabilidade dos valores de F_0 ao longo dos quadros da sentença. Esse método diminuiu a taxa de erro do sistema.

Em seguida, visando projetar um sistema mais robusto, de forma a tornar os resultados mais representativos do locutor e não da sentença, decidiu-se utilizar uma única senha falada para todos os locutores. Para tanto, foi escolhida a sentença *Quero usar a máquina*, por apresentar uma boa combinação de sons sonoros, surdos e explosivos.

Para avaliação do desempenho do sistema na tarefa de estimação da freqüência fundamental, para uma única sentença, foram utilizados, inicalmente, 20 locutores (10 do sexo feminino e 10 do sexo masculino), com 20 elocuções da sentença para cada locutor. Com o objetivo de se obter uma base de dados mais representativa, essa foi ampliada,

passando a ser composta por 40 locutores (20 do sexo masculino e 20 do sexo feminino). Os resultados são apresentados nas Figuras 6.7 e 6.8 e Tabelas A.5 e A.6 (Anexo A).

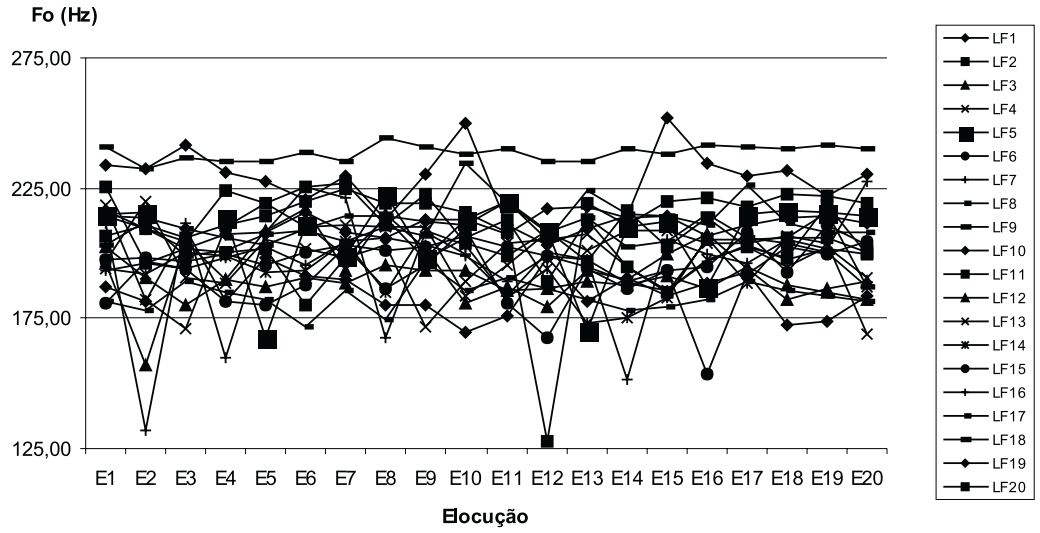


Figura 6.7: Freqüência Fundamental dos locutores femininos (LF1 a LF20), para as 20 elocuções da sentença: *Quero usar a Máquina* (E1 a E20).

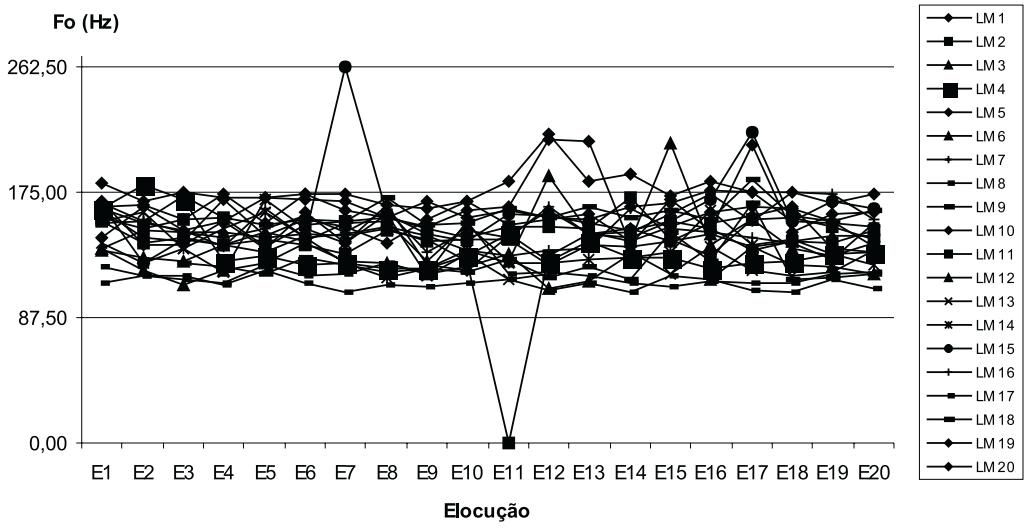


Figura 6.8: Freqüência Fundamental dos locutores masculinos (LM1 a LM20), para as 20 elocuções da sentença: *Quero usar a Máquina* (E1 a E20).

As taxas de erro são um pouco elevadas, para alguns locutores, principalmente para os locutores femininos. Verifica-se uma viariabilidade acentuada dos valores de F_0 ao longo dos segmentos. Observa-se também, a ineficiência, para vários locutores, da utilização de um único limiar de decisão na tarefa de identificação do sexo do locutor. Dessa forma, algumas modificações foram introduzidas no algoritmo de estimação da freqüência fundamental (método AMDF), apresentadas a seguir (Figura 6.9).

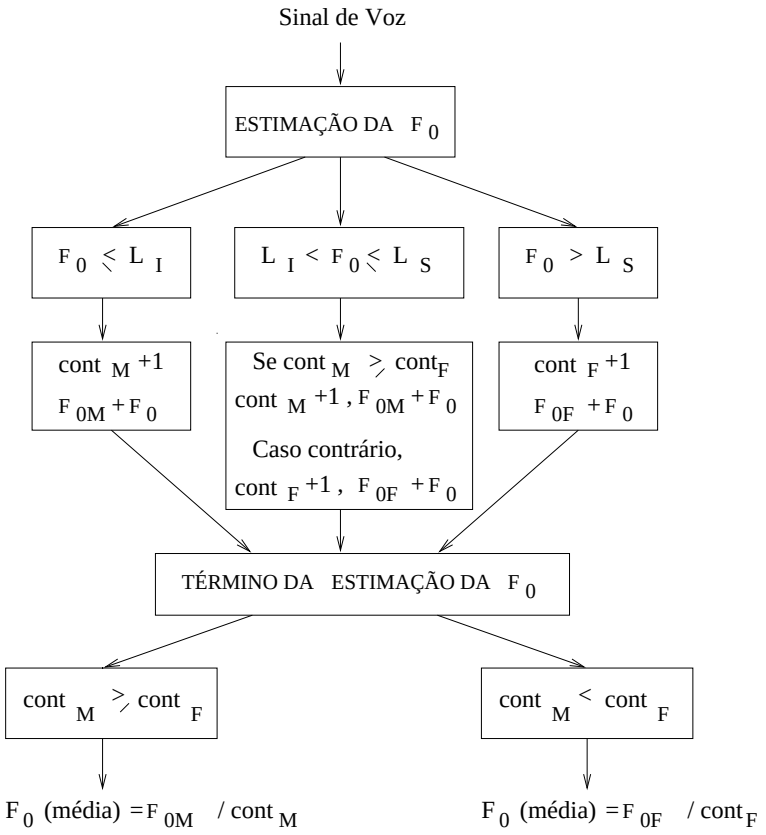


Figura 6.9: Descrição da modificação introduzida no algoritmo de estimação da Freqüência Fundamental.

As modificações podem ser descritas da seguinte forma:

1. Ajuste dos valores do *pitch* que se encontram fora da faixa de aceitação (muito baixos ou muito elevados, que não correspondem aos valores possíveis, tanto para locutores femininos quanto masculinos). O ajuste corresponde à multiplicação

(ou a divisão) dos valores do *pitch* por um fator, se estes estiverem abaixo (ou acima) do menor (ou maior) valor possível.

2. Utilização de dois limiares de decisão para determinação do sexo do locutor, em cada bloco de amostras (apenas os blocos que permitem a estimação da F_0).

3. Se F_0 for menor ou igual ao menor limiar (L_I), essa caracteriza um locutor masculino, adiciona-se F_0 à freqüência fundamental masculina (F_{0M}) e incrementa-se o contador masculino ($cont_M$).

4. Se a freqüência fundamental estimada (F_0) for maior que o maior limiar (L_S), essa caracteriza um locutor feminino, adiciona-se F_0 à freqüência fundamental feminina (F_{0F}) e incrementa-se o contador feminino ($cont_F$).

5. Se a freqüência fundamental está compreendida entre os dois limiares, observa-se o contador. Se $cont_M \geq cont_F$, adiciona-se F_0 a F_{0M} e incrementa-se $cont_M$. Caso contrário, adiciona-se F_0 a F_{0F} e incrementa-se $cont_F$.

6. Ao término da análise de todos os blocos de amostras de voz, observa-se, novamente, o contador e determina-se a freqüência fundamental média da sentença em análise, da seguinte forma:

$$F_0(\text{média}) = \frac{F_{0M}}{cont_M}, \quad \text{se } cont_M \geq cont_F \quad \text{ou} \tag{6.2}$$

$$F_0(\text{média}) = \frac{F_{0F}}{cont_F}, \quad \text{se } cont_M < cont_F. \tag{6.3}$$

Os valores dos limiares L_I e L_S foram determinados de forma empírica, sendo $L_S = 175$ Hz e $L_I = 166$ Hz (o que corresponde a, aproximadamente, 95% de L_S).

Com essas modificações no processo de estimação da freqüência fundamental tem-se, portanto, uma escolha mais representativa das F_0(média) do locutor, em virtude da utilização de dois limiares de decisão, que proporcionam uma decisão mais "suave", bem como a seleção de valores de F_0, para um dado sexo, como menor variabilidade.

Os resultados obtidos, com a modificação introduzida no processo de estimação da F_0, são apresentados nas Figuras 6.10 e 6.11 e Tabelas A.7 e A.8 (Anexo A).

A Tabela 6.1 apresenta uma análise comparativa dos dois métodos utilizados para a estimação da freqüência fundamental.

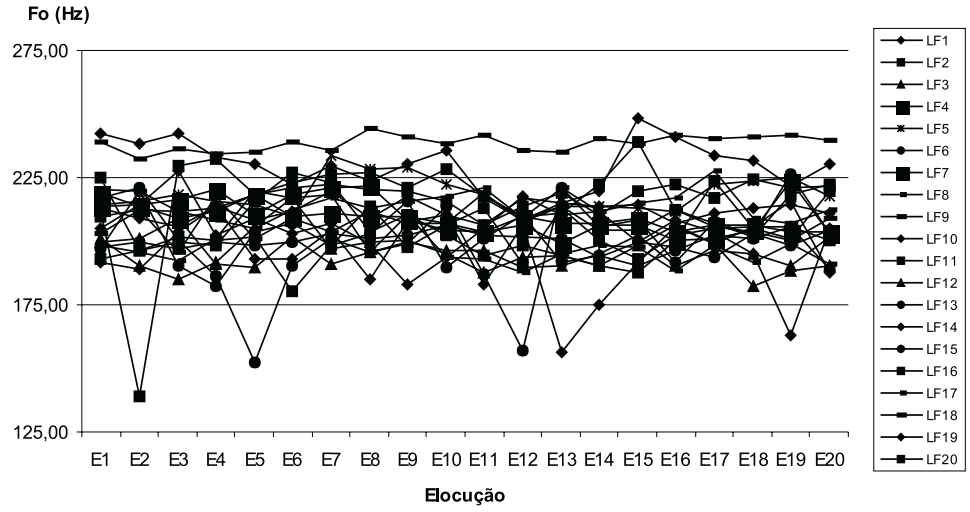


Figura 6.10: Freqüência Fundamental dos locutores femininos (LF1 a LF20), para as 20 elocuções da sentença: *Quero usar a Máquina* (E1 a E20), algoritmo AMDF modificado (AMDF-2).

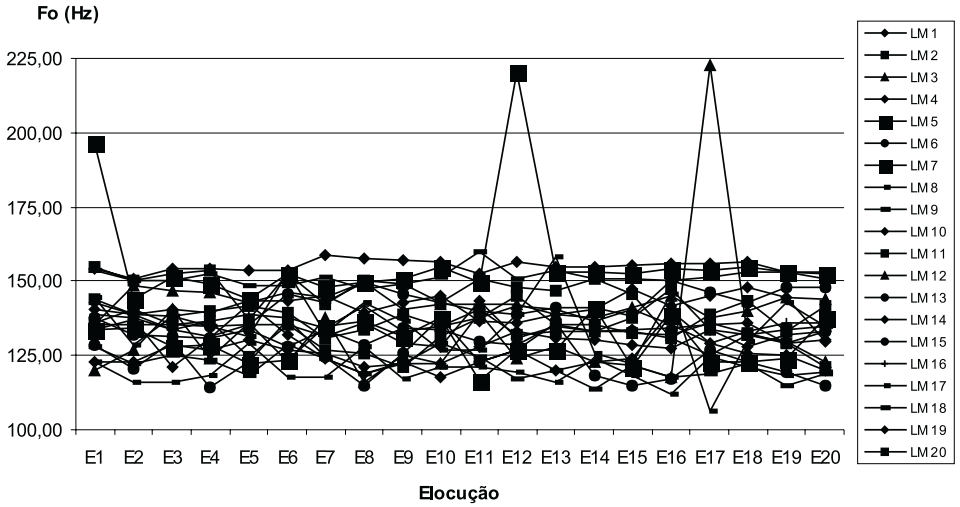


Figura 6.11: Freqüência Fundamental dos locutores masculinos (LM1 a LM20), para as 20 elocuções da sentença: *Quero usar a Máquina* (E1 a E20), algoritmo AMDF modificado (AMDF-2).

Tabela 6.1: Análise comparativa do desempenho (taxas médias de classificação correta) dos métodos utilizados para estimação da freqüência fundamental: AMDF(AMDF-1) e AMDF modificado (AMDF-2), para os locutores femininos (LF) e masculinos (LM), para a amostra composta de 40 locutores.

AMDF-1		AMDF-2	
LF	LM	LF	LM
95,3%	96,0%	98,8%	99,3%
95,6%		99,0%	

Observa-se, portanto, que a introdução da modificação no método de estimação da freqüência fundamental proporciona um aumento significativo no desempenho do sistema, tanto para os locutores femininos (95,3% para 98,8%) quanto para os masculinos (96,0% para 99,3%). Nos primeiros, o aumento de 3,5% na taxa média de classificação corresponde à redução de 19 classificações incorretas (locutor feminino é classificado como masculino) para apenas 5 (Tabelas A.5 e A.7). Em se tratando dos locutores masculinos, o aumento de 3,3% na taxa média de classificação, equivale a redução de 16 para apenas 3 classificações incorretas (Tabelas A.6 e A.8). A taxa média de classificação correta para os locutores masculinos é, para os dois métodos, um pouco superior à obtida para os locutores femininos.

Outra característica relevante obtida com a modificação, refere-se à obtenção de valores de F_0 mais representativos. Esse fato pode ser avaliado a partir das Tabelas A.7 e A.8, nas quais não existe nenhuma solicitação ao locutor para que repita a sentença, diferentemente das Tabelas A.5 e A.6.

Visando proporcionar uma avaliação mais geral dos resultados obtidos, a Figura 6.12 apresenta os valores da F_0 média de cada locutor. Os valores da F_0 média de todos os locutores estão dentro dos limites estabelecidos para os locutores masculinos e femininos, como também são representativos dos mesmos, pois a dispersão relativa (C.V.) é baixa.

As taxas de erro são baixas. Os locutores que possuem taxas de erro mais elevadas são, de uma forma geral, os que apresentam também dispersão mais elevada, indicando assim que esses locutores apresentam variações das suas características vocais para as diversas elocuções de uma mesma sentença (Tabelas A.7 e A.8).

Os valores de F_0, para um mesmo locutor, nas várias elocuções das sentenças, podem

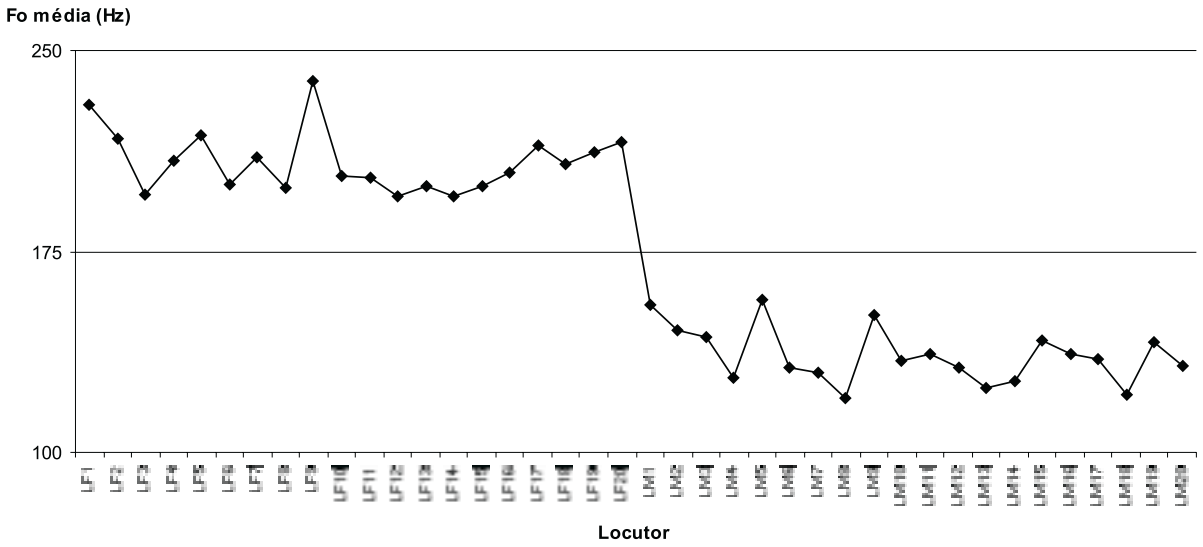


Figura 6.12: Freqüência Fundamental média dos locutores masculinos (LM1 a LM20), para as 20 elocuções da sentença: *Quero usar a Máquina* (E1 a E20), algoritmo AMDF modificado (AMDF-2).

sofrer algumas alterações em virtude da entonação, dos tipos de sons que compõem a sentença, dos ruídos presentes no ambiente de gravação, das dificuldades de escolha dos parâmetros ótimos utilizados no método de estimação de F_0, dentre outras.

Outro aspecto a ser destacado, em relação à estimação da F_0, assim como no detetor surdo-sonoro, refere-se à escolha dos limiares utilizados para os parâmetros temporais do sinal de voz. Sendo necessário, portanto, quando da eleboração do sistema, a realização de testes empíricos (dentro dos valores indicados), de forma a determinar os valores para os limiares que melhor se adaptam à aplicação.

Diante dos resultados obtidos, pode-se concluir que a Freqüência Fundamental (estimada da forma apresentada) pode ser utilizada como um parâmetro de classificação de locutores de acordo com o sexo, visto que a taxa de erro é pequena (classificação correta de 99,0%), tanto para os locutores femininos quanto masculinos. Dessa forma, a etapa de pré-identificação proporcionará um redução no volume de dados a ser analisado na etapa posterior (identificacão), visto que os locutores só serão avaliados, com confiabilidade elevada, em seu subgrupo (locutores masculinos ou femininos).

6.2.3 Identificação dos locutores

A tarefa de identificação dos locutores, como descrita anteriormente, é dividida em duas etapas. A primeira utiliza a medida de distorção, obtida a partir de dicionários projetados com a QV Paramétrica, como parâmetro de decisão para identificação do locutor. Na segunda etapa o parâmetro de decisão utilizado para identificação do locutor corresponde à medida de probabilidade obtida através da modelagem por HMM. Essa última etapa corresponde a etapa de "refinamento", sendo usada quando as medidas de distorção indicarem similaridade entre as características vocais dos locutores.

6.2.3.1 Identificação: primeira etapa

Para a primeira etapa de identificação dos locutores foram analisados três métodos. O primeiro utilizando quantização vetorial com dicionários de padrões acústicos projetados a partir do algoritmo LBG (QV-LBG), o segundo e o terceiro a quantização vetorial com dicionários de padrões acústicos projetados a partir de Redes Neurais (QV-RN), utilizando os algoritmos KMVVT [41, 42] e SSC [43], respectivamente. Nessa etapa, o sistema foi avaliado, inicialmente, para a base de dados composta de 20 locutores (10 do sexo feminino e 10 do sexo masculino).

É importante ressaltar que os resultados referentes a essa etapa, não levam a efeito a pré-identificação, pois o objetivo é determinar a eficiência dos parâmetros, de uma forma ampla, sem a redução do grupo de locutores.

Método QV-LBG

Os resultados obtidos para o SRAL, para cada um dos cinco parâmetros acústicos utilizados: LPC, Cepestrais (CEP), Cepestrais Ponderados (CEP-P), Delta Cepestrais (DCEP) e Delta Cepestrais Ponderados (DCEP-P), utilizando o método QV-LBG, são apresentados na Tabela 6.2.

Pode-se, a partir da Tabela 6.2, realizar as seguintes análises acerca do desempenho do sistema:

- Identificação - os coeficientes Cepestrais proporcionam o melhor desempenho, seguido dos Delta Cepestrais, Cepestrais Ponderados, Delta Cepestrais Ponderados e LPC, respectivamente. A taxa média de identificação do primeiro é bastante superior aos três últimos. Os locutores femininos (LF) apresentam taxas médias de identificação mais elevadas que os locutores masculinos (LM).

Tabela 6.2: Parâmetros para avaliação de desempenho do SRAL, método QV-LBG (parâmetros acústicos: LPC, CEP, CEP-P, DCEP e DCEP-P), para a amostra composta de 20 locutores.

Taxas Médias	LPC		CEP		CEP-P		DCEP		DCEP-P	
	LF	LM	LF	LM	LF	LM	LF	LM	LF	LM
identificação (%)	82,0	78,5	93,0	91,0	91,5	78,0	90,5	85,0	84,5	83,0
	80,3		92,0		84,8		87,8		83,8	
falsa aceitação (%)	10,0	7,0	3,0	0,0	0,5	2,5	5,5	6,5	4,0	2,0
	8,5		1,5		1,5		6,0		3,0	
falsa rejeição (%)	8,0	14,5	4,0	9,0	8,0	15,5	4,0	8,5	11,5	15,0
	11,3		6,5		11,8		6,3		13,3	
confiabilidade (%)	89,1	91,8	96,9	100,0	99,5	96,9	94,3	92,9	95,5	97,6
	90,5		98,4		98,2		93,6		96,6	
C.V. (%)	21,2	34,6	11,9	17,1	13,6	37,2	8,0	27,9	24,1	35,7
	27,8		14,4		26,9		19,7		29,6	

- Falsa aceitação - os coeficientes Cepestrais e Cepestrais Ponderados geram as menores taxas médias de falsa aceitação, sendo bastante inferiores às obtidas com os coeficientes LPC (essa é a mais elevada). Os locutores femininos apresentam, para a maior parte dos coeficientes, taxas de erro mais elevadas que os locutores masculinos (ou menores, porém próximas), exceto para os coeficientes Cepestrais Ponderados.

- Falsa rejeição - as menores taxas são obtidas para os coeficientes Cepestrais e Delta Cepestrais, sendo bastante inferiores às obtidas para os demais coeficientes (sendo bastante similares para os demais). As taxas médias de falsa rejeição para os locutores femininos são inferiores às obtidas para os locutores masculinos, sendo essa diferença mais acentuada do que a verificada para as taxas médias de falsa aceitação.

- Confiabilidade - esse parâmetro é bastante elevado, para a maior parte dos coeficientes (superior a 90%), principalmente para os coeficientes Cepestrais e Cepestrais Ponderados. Apesar das taxas de identificação serem mais elevadas para os locutores femininos, observa-se que a confiabilidade é maior para os locutores masculinos (para a maioria dos coeficientes). Esse fato só não é verificado para

os coeficientes Cepestrais Ponderados e Delta Cepestrais. Entretanto, no último a diferença é pequena.

- Coeficiente de Variação (C.V.) - os valores do C.V. para os coeficientes Cepestrais e Delta Cepestrais são os menores (e inferiores a 20%), o que indica que esses coeficientes apresentam as taxas médias de identificação, assim como a confiabilidade, mais representativas, indicando uma menor variabilidade dos dados em torno desses valores médios. O C.V. dos locutores masculinos é bastante superior ao obtido para os locutores femininos, principalmente para os coeficientes Cepestrais Ponderados. Portanto, as taxas médias de identificação, bem como a confiabilidade dos resultados obtidos para os locutores femininos são mais representativas que as dos locutores masculinos, principalmente para os coeficientes Cepestrais e Delta Cepestrais.

As taxas médias de falsa aceitação são, de uma forma geral, menores que as taxas médias de falsa rejeição. Esse fato mostra que o sistema é robusto no que se refere à identificação incorreta, o que o torna menos susceptível a impostores. Além disso, uma identificação incorreta é mais grave do que uma rejeição incorreta.

Com o objetivo de proporcionar uma melhor visualização dos resultados para cada um dos locutores, são apresentadas as Tabelas A.9, A.10 e A.11 (Anexo A). Alguns locutores apresentam taxas de identificação de 100%, para todos os coeficientes (p.ex.: o LM2). O locutor LM5, apresenta, para a maioria dos coeficientes, baixas taxas de identificação e elevadas taxas de falsa rejeição, porém baixas taxas de falsa aceitação, principalmente para os coeficientes Delta Cepestrais Ponderados.

Outra forma de avaliação das taxas de identificação, falsa aceitação e falsa rejeição pode ser obtida utilizando-se matrizes de confusão (apresenta a quantidade de identificações corretas e incorretas). A matriz de confusão é uma matriz quadrada $L \times L$ (L - número de locutores). Cada linha e coluna da matriz representam um locutor. O objetivo dessa matriz é permitir uma rápida visualização das taxas de identificação e falsa aceitação.

A diagonal principal da matriz de confusão representa as taxas de identificação, ou seja, a quantidade de elocuções, para as quais o locutor é identificado corretamente. As demais linhas e colunas da matriz indicam as taxas médias de falsa aceitação. O valor referente à linha i e coluna j, corresponde ao número de elocuções para as quais o locutor i é identificado de forma incorreta, como o locutor j.

Para a visualização das taxas médias de falsa rejeição, foi introduzida, ao lado da matriz de confusão, uma coluna adicional, denominada *frej*. Para cada locutor da linha i da matriz, o valor indicado pela linha i da coluna adicional, representa o número de vezes que esse locutor é rejeitado pelo sistema (considerado locutor não cadastrado).

Neste trabalho foi obtida uma matriz de confusão (Anexo A) para cada um dos cinco tipos de parâmetros acústicos: LPC, CEP, CEP-P, DCEP e DCEP-P, apresentadas nas Tabelas A.19, A.20, A.21, A.22 e A.23, respectivamente.

Para os coeficientes LPC (Tabela A.19), por exemplo, observando-se o locutor LF9, verifica-se que dentre as 20 elocuções da sentença, para 17 elocuções o locutor é identificado de forma correta, para 2 elocuções é identificado de forma incorreta (para 1 elocução é identificado como LF4 e para outra é identificado como LF10). Em apenas 1 elocução LF9 é considerado locutor não cadastrado. A maior confusão ocorre do locutor LM4 com LM8 (ou seja, LM4 é aceito como LM8), seguido de LF8 com LF10 e por fim, LF2 com LF9.

Em se tratando de Coeficientes Cepestrais (Tabela A.20), existe confusão do locutor LF8 com LF3 e LF10 (LF8 é aceito como LF3 e LF10), sendo a última mais elevada.

A matriz de confusão obtida para os coeficientes Cepestrais Ponderados (Tabela A.21) mostra que existe uma "maior confusão" do locutor LM8 com LM4.

A partir da Tabela A.22 observa-se que, para os coeficientes Delta Cepestrais, a confusão mais elevada ocorre do locutor LF2 com LF9 e LM8 com LM4, respectivamente.

A matriz de confusão obtida para os coeficientes Delta Cepestrais Ponderados (Tabela A.23) mostra que a maior confusão ocorre do locutor LF3 com LF6.

Pode-se concluir portanto, observando-se as matrizes de confusão, que os coeficientes Cepestrais e LPC, geram as menores e maiores taxas de identificação incorreta (falsa aceitação), respectivamente.

No que se refere às taxas de falsa rejeição, os coeficientes Cepestrais proporcionam, de uma forma geral, as menores taxas. Uma redução bastante significativa, em relação aos demais coeficientes, ocorre, principalmente, com o locutor LM5.

Outra característica observada através das matrizes de confusão, mostra que os coeficientes utilizados são, em sua maioria, robustos no que se refere à identificação de locutores de acordo com o sexo (ou seja, não identificar um locutor masculino como

locutor feminino, ou vice-versa). Este fato ocorreu apenas três vezes com os coeficientes Delta Cepestrais (Tabela A.22). Por exemplo, em uma elocução LF3 foi identificado como LM9. Esse problema poderá ser solucionado com a utilização da freqüência fundamental.

Os resultados obtidos para o método QV-LBG mostram que os coeficientes Cepestrais e Delta Cepestrais apresentam, de uma forma geral, desempenho superior aos demais coeficientes. Pode-se observar também, que o desempenho do sistema é, de uma forma geral, um pouco superior para os locutores femininos em detrimento aos locutores masculinos.

Método QV-KMVVT

Para a implementação do sistema, utilizando o algoritmo KMVVT no projeto dos dicionários de padrões acústicos dos locutores, foram avaliados quatro parâmetros acústicos (LPC, Cepestrais, Delta Cepestrais e Delta Cepestrais Ponderados). Os coeficientes Cepestrais Ponderados não foram utilizados nesse algoritmo, por apresentar desempenho próximo ao dos coeficientes Delta Cepestrais Ponderados, variabilidade elevada para as taxas médias de identificação (tanto para locutores masculinos quanto femininos) bem como, diferença acentuada de desempenho em relação ao sexo do locutor, conforme mostrado anteriormente (Tabela 6.2).

A Tabela 6.3 apresenta os parâmetros para avaliação do desempenho do SRAL, utilizando o método QV-KMVVT.

Pode-se, a partir da Tabela 6.3, realizar as seguintes análises acerca do desempenho do sistema:

- Identificação - as taxas médias de identificação mais elevadas são obtidas para os coeficientes Cepestrais, Delta Cepestrais, LPC e Delta Cepestrais Ponderados, respectivamente, sendo a primeira muito superior às demais. A diferença entre as taxas de identificação dos locutores femininos e masculinos não é elevada (exceto para os coeficientes LPC).

- Falsa aceitação - as menores taxas médias de falsa aceitação são obtidas para os coeficientes Cepestrais e Delta Cepestrais Ponderados. Sendo a primeira bem inferior às demais. Os locutores femininos apresentam, para a maior parte dos coeficientes, taxas de erro um pouco superiores às obtidas para os locutores masculinos (ou menores, porém próximas, exceto para os coeficientes LPC, nos quais

Tabela 6.3: Parâmetros para avaliação de desempenho do SRAL, método QV-KMVVT (parâmetros acústicos: LPC, CEP, DCEP e DCEP-P), para a amostra composta de 20 locutores.

Taxas Médias	LPC		CEP		DCEP		DCEP-P	
	LF	LM	LF	LM	LF	LM	LF	LM
identificação (%)	85,0	91,5	94,0	99,0	90,0	86,0	88,0	85,0
	88,3		96,5		88,0		86,5	
falsa aceitação (%)	12,5	3,0	2,5	0,0	6,0	8,5	6,5	2,0
	7,8		1,3		7,3		4,3	
falsa rejeição (%)	2,5	5,5	3,5	1,0	4,0	5,5	5,5	13,0
	4,0		2,3		4,8		9,3	
confiabilidade (%)	87,2	96,8	97,4	100,0	93,8	91,0	93,1	97,7
	92,0		98,7		92,4		95,4	
C.V. (%)	23,4	10,0	9,3	2,1	12,3	27,9	19,4	19,8
	17,5		7,0		20,8		19,1	

a diferença é elevada).

- Falsa rejeição - os coeficientes Cepestrais fornecem a menor taxa média de falsa rejeição (superior à taxa média de falsa aceitação), seguido dos coeficientes LPC e Delta Cepestrais, sendo as duas últimas bastante próximas. Os coeficientes Delta Cepestrais Ponderados geram a maior taxa média de falsa rejeição. As taxas médias de falsa rejeição para os locutores femininos são um pouco inferiores às obtidas para os locutores masculinos, exceto para os coeficientes Cepestrais (para esse coeficiente a taxa média de falsa rejeição dos locutores masculinos é nula). A diferença mais acentuada entre as taxas médias de falsa rejeição, para locutores masculinos e femininos, é observada com os coeficientes Delta Cepestrais Ponderados.

- Confiabilidade - para todos os coeficientes a confiabilidade é elevada e com resultados bastante próximos. O maior valor é obtido para os coeficientes Cepestrais. A diferença entre a confiabilidade dos locutores femininos e masculinos não é acentuada (exceto para os coeficientes LPC), sendo, em sua maioria, mais elevada para os locutores masculinos. Para os coeficientes Cepestrais, a confiabilidade dos locutores masculinos foi 100%, devido a não ocorrência de falsa aceitação.

- Coeficiente de Variação (C.V.) - as taxas médias de identificação e, conseqüentemente, confiabilidade são mais representativas para os coeficientes Cepestrais. Os demais coeficientes apresentam variabilidade mais acentuada, com valores bastante próximos entre si. Para os coeficientes Delta Cepestrais, o C.V. dos locutores masculinos é bastante superior ao obtido para os locutores femininos.

A partir do exposto, pode-se observar que a substituição do algoritmo LBG pelo KMVVT, para os coeficientes LPC, fez com que a taxa média de identificação aumentasse de 80,3% para 88,3%. Em se tratando de coeficientes Cepestrais, o algoritmo KMVVT resultou em taxa média de identificação de 96,5%. Em relação aos coeficientes Delta Cepestrais e Delta Cepestrais Ponderados, verifica-se também um melhor desempenho do algoritmo KMVVT, em relação ao algoritmo LBG. O método QV-KMVVT proporciona, de uma forma geral, uma diminuição das taxas médias de falsa rejeição e falsa aceitação, quando comparado ao QV-LBG. As taxas médias de identificação obtidas para o QV-KMVVT são mais representativas do que às obtidas para o QV-LBG.

Os resultados obtidos, para cada um dos locutores, são apresentados nas Tabelas A.12, A.13 e A.14 (Anexo A). Alguns locutores apresentam taxas de identificação de 100% (principalmente para os coeficientes Cepestrais). O locutor LM5, apresenta, para a maioria dos coeficientes, baixas taxas de identificação (exceto para os coeficientes Cepestrais, que proporcionam uma taxa de identificação de 100%) e elevadas taxas de falsa rejeição, porém menores do que as obtidas com o método QV-LBG. Essas características também podem ser observadas através das matrizes de confusão (Tabelas A.24, A.25, A.26 e A.27 - Anexo A).

Verifica-se, portanto, que o método QV-KMVVT apresenta desempenho superior ao apresentado pelo tradicional algoritmo LBG, uma vez que os dicionários de padrões acústicos produzidos pelo algoritmo KMVVT representam de forma mais eficiente os locutores cadastrados no sistema, quando comparados aos correspondentes dicionários projetados com o algoritmo LBG.

Dentre os tipos de coeficientes utilizados, os coeficientes Cepestrais e Delta Cepestrais apresentam maior eficiência no que se refere à identificação de pessoas a partir de suas vozes, pois proporcionam, de uma forma geral, as maiores taxas de identificação, bem como as menores taxas médias de falsa rejeição e de falsa aceitação. Os coeficientes Cepestrais e Delta Cepestrais estabelecem, portanto, um bom compromisso entre as taxas médias de falsa rejeição e de falsa aceitação.

Outra característica importante observada com a substituição do método QV-LBG pelo QV-KMVVT, refere-se ao fato da diminuição da diferença entre os valores das medidas de desempenho obtidas para os locutores masculinos e femininos. O que caracteriza um fato relevante, visto que o sistema deve ser robusto, sendo capaz de identificar de forma correta os locutores, independente do sexo.

Método QV-SSC

No último método, QV-SSC, são analisados apenas os coeficientes Cepestrais e Delta Cepestrais, visto que já tinho sido observado, para os demais métodos (QV-LBG e QV-KMVVT), que esses apresentam desempenho superior aos demais coeficientes avaliados (Tabelas 6.2 e 6.3). Os resultados obtidos são apresentados na Tabela 6.4.

Tabela 6.4: Parâmetros para avaliação do desempenho do SRAL, método QV-SSC (parâmetros acústicos: CEP e DCEP), para a amostra composta de 20 locutores.

Taxas Médias	CEP		DCEP	
	LF	LM	LF	LM
identificação (%)	97,5	98,0	90,5	86,5
	97,8		88,5	
falsa aceitação (%)	1,5	0,0	3,0	3,5
	0,8		3,3	
falsa rejeição (%)	1,0	2,0	6,5	10,0
	1,5		8,3	
confiabilidade (%)	98,5	100,0	96,8	96,1
	99,2		96,5	
C.V.(%)	5,5	3,6	14,1	27,7
	4,5		21,2	

Tem-se, a partir da Tabela 6.4, as seguintes análises acerca do desempenho do sistema:

- Identificação - os coeficientes Cepestrais proporcionam a taxa média de identificação mais elevada (bem superior à obtida com os coeficientes Delta Cepestrais). A diferença entre as taxas médias de identificação dos locutores femininos e masculinos é pequena, principalmente para os coeficientes Cepestrais.

- Falsa aceitação - a menor taxa média de falsa aceitação é obtida para os os

coeficientes Cepestrais (menor que 1%). A diferença entre as taxas médias de falsa aceitação para os locutores femininos e masculinos é pequena, sendo para os últimos 0%, quando da utilização dos coeficientes Cepestrais.

- Falsa rejeição - os coeficientes Cepestrais proporcionam a menor taxa média de falsa rejeição (superior à taxa média de falsa aceitação), sendo bastante inferior à obtida para os coeficientes Delta Cepestrais. As taxas médias de falsa rejeição para os locutores femininos são inferiores às obtidas para os locutores masculinos (porém a diferença não é elevada).

- Confiabilidade - para os dois coeficientes a confiabilidade é bastante elevada e com resultados bastante próximos, sendo um pouco superior para os coeficientes Cepestrais. A diferença obtida entre a confiabilidade dos locutores femininos e masculinos é bastante pequena. Para os coeficientes Cepestrais, a confiabilidade dos locutores masculinos foi de 100%, devido a não ocorrência de falsa aceitação.

- Coeficiente de variação (C.V.) - as taxas médias de identificação e, conseqüentemente, confiabilidade são mais representativas para os coeficientes Cepestrais. O C.V. para os coeficientes Delta Cepestrais, dos locutores masculinos é bastante superior ao obtido com os locutores femininos, para os coeficientes Cepestrais ocorre o inverso (com pequenos valores de C.V.).

Os resultados obtidos, para cada um dos locutores, são apresentados na Tabela A.15 (Anexo A). A maioria dos locutores apresenta taxas de identificação de 100% e baixas taxas de falsa rejeição e falsa aceitação (principalmente para os coeficientes Cepestrais). Exceção é feita ao LM5 que apresenta, para os coeficientes Delta Cepestrais, baixa taxa de identificação. Verifica-se, portanto, que o desempenho, para cada locutor, como uso do método QV-SSC é na maioria, superior ao obtido com os métodos QV-LBG e QV-KMVVT. Essas características também podem ser observadas a partir das matrizes de confusão (Tabela A.28 e A.29 - Anexo A).

Avaliando os resultados referentes aos métodos QV-LBG, QV-KMVVT e QV-SSC, observa-se que o sistema apresenta um melhor desempenho com o método QV-SSC (utilizando-se os coeficientes Cepestrais), visto que este proporciona a maior taxa média de identificação (97,8%) e as menores taxas médias de falsa aceitação (0,8%) e falsa rejeição (1,5%), bem como a maior confiabilidade (99,2%). É importante destacar que o método QV-KMVVT, mesmo com medidas de desempenho menores que o QV-SSC, apresenta desempenho superior ao método QV-LBG.

Para os três métodos avaliados, os coeficientes Cepestrais proporcionam o melhor desempenho, sendo mais elevado para o método QV-SSC. No que se refere aos coeficientes Delta Cepestrais, os três métodos apresentam desempenho bastante similar, sendo um pouco mais elevado para o QV-SSC.

Pode-se verificar também, a partir das Tabelas 6.2, 6.3 e 6.4, que para quaisquer dos métodos e coeficientes, as taxas de falsa rejeição são, de uma forma geral, maiores que as de falsa aceitação. Este fato representa uma característica importante do sistema, pois é melhor rejeitar um locutor erroneamente do que identificá-lo de forma incorreta.

Para os coeficientes Cepestrais, as taxas médias de identificação obtidas para o método QV-SSC são as mais representativas, seguidas das obtidas pelo QV-KMVVT e por fim pelo QV-LBG. Para coeficientes Delta Cepestrais, observa-se que a representatividade dos valores médios obtidos é bem menor quando comparada às obtidas com os coeficientes Cepestrais, existindo pouca diferença em relação aos três métodos.

No método QV-SSC, assim como no QV-KMVVT, verifica-se uma pequena diferença entre os valores das medidas de desempenho obtidas para os locutores masculinos e femininos, indicando a robustez do sistema no que se refere ao sexo do locutor.

Após a análise comparativa de desempenho do sistema, para os três métodos, e a verificação da superioridade do método QV-SSC sobre os demais, ampliou-se a base de dados, passando a ser composta por quarenta locutores (20 do sexo feminino e 20 do sexo masculino). Os resultados obtidos são apresentados na Tabela 6.5.

As medidas de desempenho, apresentadas na Tabela 6.5, mostram que, de forma similiar aos resultados obtidos para os 20 locutores, os coeficientes Cepestrais apresentam desempenho superior ao obtido com os coeficientes Delta Cepestrais mantendo-se, de uma forma geral, as mesmas características descritas anteriormente (apresentadas na Tabela 6.4). Há um aumento das taxas médias de identificação (principalmente para os coeficientes Delta Cepestrais), bem como uma diminuição da taxa média de falsa rejeição, porém um pequeno aumento da taxa média de falsa aceitação, para os coeficientes Delta Cepestrais. Esses resultados são relevantes pois mostram que o sistema mantém suas características mesmo com a ampliação da base de dados.

A Tabela A.16 (Anexo A) apresenta os resultados, para cada um dos locutores. A maioria apresenta taxas de identificação de 100% e baixas taxas médias de falsa rejeição e falsa aceitação. Essas características também podem ser observadas a partir das matrizes de confusão (Tabelas A.30 e A.31 - Anexo A).

Tabela 6.5: Parâmetros para avaliação de desempenho do SRAL, método QV-SSC (parâmetros acústicos: CEP e DCEP), para a amostra composta de 40 locutores.

Taxas Médias	CEP		DCEP	
	LF	LM	LF	LM
identificação (%)	98,3	98,3	92,3	91,0
	98,3		91,6	
falsa aceitação (%)	1,0	0,8	4,5	5,5
	0,9		5,0	
falsa rejeição (%)	0,8	1,0	3,3	5,0
	0,9		4,1	
confiabilidade (%)	99,0	99,2	95,3	94,3
	99,1		94,8	
C.V. (%)	4,1	3,4	12,9	19,8
	3,7		16,5	

Diante do exposto, pode-se verificar que o desempenho superior dos coeficientes Cepestrais (no que se refere à modelagem das características vocais dos locutores) era esperado, visto que o método utilizado para o cálculo destes coeficientes separa, através da operação logaritmo, a função de transferência do trato vocal e a fonte de voz, modelando o trato vocal, o qual é peça chave para distinguir locutores entre si [55], representando com mais eficiência as características vocais dos locutores.

Os coeficientes Delta Cepestrais, Cepestrais Ponderados e Delta Cepestrais Ponderados, apresentam desempenho inferior ao obtido com os Cepestrais. Esse fato é observado, visto que a operação de derivação é útil na captura das informações de transição da voz [19], o que não é um fator de interesse nesse sistema pois os locutores pronunciam a mesma sentença. A operação de ponderação não proporcionou também melhoria no desempenho, visto que é útil na minimização da sensibilidade dos coeficientes de baixa ordem em relação à envoltória espectral e a sensibilidade dos coeficientes Cepestrais de alta ordem em relação ao ruído [19], o que não se caracteriza em um problema a ser minimizado nesse sistema pois a influência da envoltória espectral, bem como do ruído não é crítica.

6.2.3.2 Identificação dos locutores: segunda etapa

Tomando-se como base os resultados obtidos na etapa anterior,utilizou-se, inicialmente, para implementação da última etapa do método, apenas os coeficientes Cepestrais para obtenção do vetor de observações (conjunto de símbolos, utilizando o método QV-SSC) a ser utilizado na construção do HMM de referência (treinamento) e para o cálculo da medida de probabilidade (identificação), para cada locutor do sistema.

A modelagem por HMM se constitui em uma etapa de "refinamento" do processo de identificação, utilizada quando as medidas de distorção indicam similaridade entre os padrões vocais dos locutores (Tabela A.32). Entretanto, observou-se que o uso dos coeficientes Cepestrais, unicamente, na construção do HMM de referência não foi suficiente para aumentar a eficiência do sistema. Assim sendo, adotou-se a seguinte metodologia:

- Fase de treinamento: construção de dois HMMs de referência para cada locutor, o primeiro utilizando os coeficientes Cepestrais e o segundo os coeficientes Delta Cepestrais, na obtenção dos símbolos do quantizador vetorial.

- Fase de teste (identificação): uso de uma medida de probabilidade, que corresponde a média aritmética ponderada das probabilidades obtidas a partir dos HMMs de referência que utilizam coeficientes Cepestrais (peso - 70%) e dos HMMs que utilizam coeficientes Delta Cepestrais (peso - 30%), respectivamente. A ponderação foi determinada de forma empírica, tomando-se como base os resultados obtidos na primeira etapa do processo de identificação.

Os resultados obtidos são apresentados na Tabela 6.6. Tem-se, a partir da Tabela 6.6, as seguintes análises acerca do desempenho do sistema:

- Identificação - a introdução da segunda etapa de identificação proporciona um aumento da taxa de identificação (98,3% para 98,8%). Esse aumento pode parecer, inicialmente, pequeno mas é bastante significativo, visto que a taxa média de identificação já era bastante elevada. Os locutores masculinos e femininos mantêm a característica do método QV-SSC, ou seja, apresentam taxas médias de identificação bastante similares.

- Falsa aceitação - ocorre um decréscimo da taxa média de falsa aceitação, tanto para os locutores femininos quanto masculinos. Para os últimos a taxa média

Tabela 6.6: Parâmetros para avaliação de desempenho do SRAL, método QV-SSC-HMM, para a amostra composta de 40 locutores.

Taxas Médias	LF	LM
identificação (%)	98,5	99,0
	98,8	
falsa aceitação (%)	0,8	0,0
	0,4	
falsa rejeição (%)	0,8	1,0
	0,9	
confiabilidade (%)	99,2	100,0
	99,6	
C.V. (%)	3,3	2,6
	3,0	

de falsa aceitação foi de 0%. Esse resultado comprova a eficiência do algoritmo, visto que o objetivo é diminuir o erro quando da ocorrência de similaridade entre as características vocais dos locutores.

- Falsa rejeição - não há alteração da taxa média de falsa rejeição em relação ao método QV-SSC, o que era esperado, pois a segunda etapa do processo de identificação não tem influência sobre essa taxa.

- Confiabilidade - em decorrência do aumento da taxa média de identificação e diminuição da taxa média de falsa aceitação, ocorre o aumento da confiabilidade do sistema. É importante destacar que a confiabilidade é de 100% para os locutores masculinos.

- Coeficiente de Variação (C.V.) - a taxa média de identificação, bem como a confiabilidade, além de elevadas são bastante representativas (C.V.=3,0%). Esse fato é observado tanto para os locutores femininos (C.V.=3,3%) quanto para os locutores masculinos (C.V.=2,6%).

Os resultados obtidos, para cada locutor, podem ser observados a partir da Tabela A.17 (ou da matriz de confusão apresentada na Tabela A.34 - Anexo A). A maior parte dos locutores apresenta taxas de identificação de 100% e baixas taxas de falsa rejeição e falsa aceitação.

Os resultados mostram, portanto, a eficiência do método apresentado, ou seja, que a modelagem estatística se mostra eficiente na discriminação de locutores que apresentam características vocais similares.

6.2.3.3 Pré-identificação + identificação

Após a avaliação, em separado, das etapas de pré-identificação e identificação do sistema, realizou-se a implementação do sistema completo, ou seja, implementação conjunta das duas etapas. Os resultados obtidos são apresentados na Tabela 6.7.

Tabela 6.7: Parâmetros para avaliação de desempenho do SRAL, método QV-SSC-HMM, adicionada a etapa de pré-identificação, para a amostra composta de 40 locutores.

Taxas Médias	LF	LM
identificação (%)	97,3	98,3
	97,8	
falsa aceitação (%)	1,3	0,3
	0,8	
falsa rejeição (%)	1,5	1,5
	1,5	
confiabilidade (%)	98,7	99,7
	99,2	
C.V. (%)	3,9	4,7
	4,3	

A introdução da etapa de pré-identificação proporciona uma pequena diminuição da taxa média de identificação e da confiabilidade e o aumento da taxa média de falsa aceitação do sistema. Um aspecto importante, observado com a utilização da etapa de pré-identificação, refere-se ao fato de que dentre as 8 classificações incorretas, em 5 destas o locutor foi considerado não cadastrado e em apenas 3 o locutor foi identificado erroneamente como um locutor do sexo oposto. Há portanto, uma aumento mais significativo na taxa média de falsa rejeição em detrimento à taxa média de falsa aceitação, o que representa um problema menos crítico.

Apesar da pequena diminuição na eficiência do sistema, o uso da freqüência fundamental, como parâmetro de separação prévia dos locutores em subgrupos de acordo

com sexo, diminui o volume de dados a ser analisado na etapa de identificação, reduzindo o tempo computacional dessa etapa. Além disso, a pré-identificação poderá diminuir a probabilidade de falsa aceitação a ser gerada pela identificação de um locutor feminino como masculino (ou vice-versa), apesar dos coeficientes Cepestrais serem bastante eficientes em relação a esse aspecto.

6.3 Análise Estatística de Desempenho

6.3.1 Conceitos Básicos

Na realização de uma pesquisa o ideal seria proceder-se a determinação das medidas sobre toda a "população" em estudo. Sendo esse procedimento impraticável (ou difícil), em alguns levantamentos, recorre-se ao processo de amostragem e, a partir da Estatística Indutiva (ou Inferência Estatística), pode-se tirar conclusões sobre a população com base nos resultados observados na amostra. O que é necessário garantir, em suma, é que a amostra seja *representativa* da população. Isso significa que, a menos de certas discrepâncias inerentes à aleatoriedade sempre presente, em maior ou menor grau, no processo de amostragem, a amostra deve possuir as mesmas características básicas da população, no que diz respeito à(s) variável(is) de interesse da pesquisa [110].

Considera-se "amostra", em estatística, um número limitado de observações retirado de um conjunto da mesma natureza chamado "população" ou "universo". Distingue-se dois tipos de amostragem: a *probabilística* e a *não probabilística*. A amostragem será probabilística se todos os elementos da população tiverem probabilidade conhecida, e diferente de zero, de pertencer à amostra. Caso contrário, a amostragem será não probabilística [110].

As técnicas da Estatística Indutiva pressupõem que as amostras utilizadas sejam probabilísticas, o que muitas vezes não se pode conseguir. No entanto, o bom senso irá indicar quando o processo de amostragem, embora não sendo probabilístico, pode ser, para efeitos práticos, considerado como tal. Isso amplia consideravelmente as possibilidades de utilização do método estatístico em geral.

Outro aspecto de relevância que deve ser observado é o tamanho da amostra. É importante a fixação do tamanho da amostra, pois a *lei empírica do acaso* ou *lei dos grandes números* consagra o princípio de que a aproximação relativa aumenta à medida

que cresce o número de determinações. A amostra deve incluir um número suficiente de casos, escolhidos aleatoriamente, para oferecer certa segurança estatística em relação à representatividade dos dados. Assim, o tamanho de uma amostra deve alcançar determinadas proporções mínimas, estabelecidas estatisticamente. Além disso, as necessidades práticas de tempo, custos, etc. recomendam não ultrapassar o tamanho mínimo determinado pela estatística.

É necessário conhecer o tamanho da amostra, não só para garantir a possibilidade de generalizar os resultados, mas também pelos aspectos práticos mencionados. Como ponto de referência, pode ser adotada a recomendação de Pearson limitando em 20 o número mínimo de observações ou o *tamanho da amostra*. Alguns autores recomendam um mínimo de 30 observações para esse limite. Como norma geral considera-se "pequenas amostras" as que contêm menos de 30 unidades amostrais ($n_A < 30$) e como "grandes amostras" as que contêm mais de 30 unidades amostrais ($n_A > 30$).

Em uma pesquisa, cujos resultados são obtidos através de uma amostra, é de fundametal importância comprovar se esses resultados apresentam um erro desprezível e são significativos, para tanto utiliza-se a teoria das probabilidades. Quando se calculam parâmetros (por exemplo médias) à base de amostras, as medidas encontradas estão sujeitas a erros. O erro cometido por esse processo deve ser *aceitável*. Uma medida utilizada para avaliar esse erro denomina-se *Erro Padrão*. Assim, com base nas médias obtidas nas amostras, pode-se inferir a *verdadeira* média da população considerada e verificar se o erro cometido nesse processo é desprezível, ou seja, se a média obtida é significativa [110].

6.3.2 Erro Padrão da Média

O *Erro Padrão* da média depende do número de observações da amostra tomada e da variabilidade das medidas. É representado pela notação: EP_m. É facilmente obtido, para amostras probabilísticas ou praticamente probabilísticas, pela seguinte fórmula [110]:

$$EP_m = \frac{s}{\sqrt{n_A}}. \tag{6.4}$$

Sendo s o desvio padrão da amostra e n_A o tamanho da amostra. Quanto maior o tamanho da amostra e menor o desvio padrão, tanto menor será o erro padrão da *média*. Se o erro padrão obtido da amostra é desprezível, ou seja, muito menor que a *média*

encontrada, a *média* observada pode ser tomada como representativa da "população" estudada (ou significativa).

Através da determinação do *Erro Padrão* pode-se estimar a média populacional, utilizando-se dois métodos:

1. *estimação por ponto* - procede-se a estimativa da média populacional através de um único valor estimado.

2. *estimação por intervalo* - constroi-se um intervalo, o qual deverá, com probabilidade conhecida, conter a média populacional. O intervalo denomina-se *intervalo de confiança* para a média e a probabilidade de que esse intervalo contenha o verdadeiro valor da média, denomina-se *nível de confiança* ou *grau de confiança*, sendo representado por (1-α). Logo, α representa a probabilidade de erro ao se afirmar que o intervalo contém o verdadeiro valor do parâmetro.

As estimativas por ponto são, em geral, utilizadas quando necessita-se, ao menos aproximadamente, conhecer o valor do parâmetro para utilizá-lo em uma expressão analítica qualquer. Entretanto, se a determinação de um dado parâmetro é a meta final do estudo estatístico em questão, a estimação por ponto é, em geral, insuficiente, pois a probabilidade de que a estimativa adotada venha a coincidir com o verdadeiro valor do parâmetro é, em geral, nula ou quase nula. Ou seja, é quase certo que se esteja cometendo um erro de estimação, quando procede-se a estimação por ponto de um parâmetro populacional. Portanto, deve-se preterir a estimação por intervalo em detrimento à estimação por ponto.

6.3.3 Estimativa do intervalo de confiança da média aritmética de uma população

Quando deseja-se estabelecer um intervalo de confiança para μ, a média da população, considera-se duas situações: a variância populacional (σ^2) é conhecida (ou, equivalentemente, o desvio padrão, σ) ou σ^2 não é conhecida.

Assim como a média aritmética da população μ é geralmente desconhecida, a variância da população (σ^2) tem pouca probabilidade de ser conhecida. Portanto, precisa-se obter uma estimativa do intervalo de confiança de μ utilizando somente as estatísticas de amostras de $\overline{x}$ (média da amostra) e s [115]. A construção do intervalo de confiança

pode ser realizada considerando duas situações: "grandes amostras" ou "pequenas amostras". Devido ao tamanho da base de dados utilizada neste trabalho, a análise estatística será realizada para "pequenas amostras".

Para "pequenas amostras", torna-se necessário uma correção na distribuição padronizada, que consiste em substituir a distribuição Normal (Z) (utilizada para "grandes amostras") pela distribuição *t*-Student (Tabela A.35 - Anexo A).

Assim, para "pequenas amostras", o intervalo de confiança para a média μ da população é da forma [110, 115]:

$$\begin{gathered} \overline{x} \pm t_{(n_A-1;\frac{\alpha}{2})}\frac{s}{\sqrt{n_A}} \qquad \text{ou} \\ \overline{x} - t_{(n_A-1;\frac{\alpha}{2})}\frac{s}{\sqrt{n_A}} \leq \mu \leq \overline{x} + t_{(n_A-1,\frac{\alpha}{2})}\frac{s}{\sqrt{n_A}} \end{gathered} \tag{6.5}$$

em que t_{n_A-1} é o valor crítico da distribuição *t*-Student, com $n_A - 1$ graus de liberdade, para uma área de $\frac{\alpha}{2}$. O erro padrão da amostra será multiplicado por um fator de correção dado por $t_{(n_A-1;\frac{\alpha}{2})}$.

O intervalo de confiança, ao nível de confiança 1-α, afirma que se tem $1-\alpha$ de certeza de que a amostra selecionada é uma amostra em que a média aritmética da população, μ, está localizada dentro do intervalo.

Outro aspecto a ser analisado em pesquisas por amostragem, que são formadas por subgrupos, é a observação das diferenças existentes entre os subgrupos, em relação os parâmetros estimados. Este tipo de análise pode ser levado a efeito através da realização de Teste de Hipóteses [115].

Como no âmbito deste trabalho os parâmetros obtidos correspondem à médias aritméticas, torna-se necessário avaliar as diferenças existentes entre essas médias. Por exemplo, é importante inferir se existe realmente, para uma população de locutores, uma diferença significativa entre a F_0 média feminina e masculina, observando se a primeira é maior do que a segunda.

Outro parâmetro que deve ser observado, através do teste de hipóteses, refere-se às taxas médias de pré-identificação e de identificação, de forma a ser possível avaliar, com um nível de confiança elevado, se existe diferença significativa entre essas taxas, para os locutores femininos e masculinos, o que permitirá inferir se o sistema realmente apresenta desempenho similar para os dois sexos. Essas análises já foram realizadas anteriormente, neste Capítulo (considerando apenas a amostra), porém é importante

inferir acerca da generalização dos resultados. Para tanto, será utilizado o *Teste t de Variância Combinada para Diferenças Entre Duas Médias Aritméticas* [115].

6.3.4 Aplicação do Teste t de Variância Combinada para Diferenças Entre Duas Médias Aritméticas

A estatística do teste t de variância combinada segue uma distribuição t com $n_{A1} + n_{A2}$ -2 graus de liberdade (n_{A1} e n_{A2}, tamanhos das amostras retiradas das populações 1 e 2 , respectivamente). Para um dado nível de significância, α, a regra de decisão é dada por [115]:

$$\text{rejeitar } H_0 \text{ (hipótese nula) se } t > t_{n_{A1}+n_{A2}-2} \text{ ou se } t < -t_{n_{A1}+n_{A2}-2},$$

$$\text{caso contrário, não rejeitar } H_0.$$

O teste a ser realizado pode ser unicaudal ou bicaudal, dependo da aplicação, ou seja, se o objetivo é testar se as duas médias aritméticas das populações são meramente diferentes ou se uma média aritmética é maior que a outra, o que pode ser descrito da forma [115]:

Teste Bicaudal	Teste Unicaudal	Teste Unicaudal
$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 \geq \mu_2$	$H_0 : \mu_1 \leq \mu_2$
$H_1 : \mu_1 \neq \mu_2$	$H_1 : \mu_1 < \mu_2$	$H_1 : \mu_1 > \mu_2$

Pode-se calcular a estatística do *teste t de variância combinada* pela expressão [115]

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n_{A1}} + \frac{1}{n_{A2}})}}, \tag{6.6}$$

em que

$$S_p^2 = \frac{(n_{A1} - 1)S_1^2 + (n_{A2} - 1)S_2^2}{(n_{A1} - 1) + (n_{A2} - 1)}, \tag{6.7}$$

sendo S_p^2 a variância combinada, $\overline{x_1}$ e $\overline{x_2}$ as médias aritméticas das amostras retiradas das populações 1 e 2, respectivamente, S_1^2 e S_2^2 as variâncias das amostras retiradas das populações 1 e 2, respectivamente.

6.3.5 Análise estatística dos valores obtidos no SRAL

A realização de uma estatística não tendenciosa é uma tarefa bastante difícil na avaliação de um Sistema de Reconhecimento Automático de locutor (SRAL). Os índices de desempenho variam com o vocabulário utilizado na avaliação, o tipo de pronúncia, o ruído, dentre outros fatores.

A aplicação de uma análise estatística, quando da realização de uma pesquisa por amostragem, como no caso deste trabalho, é bastante importante, pois a partir dessa análise é possível inferir, com determinado nível de confiança, a cerca da generalização dos resultados obtidos com a amostra, o que permitirá a obtenção de resultados mais representativos e, conseqüentemente, mais conclusivos.

Apesar da amostra de locutores não ter sido escolhida de forma probabilística (ou aleatória) [110], evitando assim a tendência dos resultados, procurou-se utilizar uma quantidade não muito pequena de elocuções por locutor, além de locutores com características vocais bastante diferentes, procurando compor uma amostra de locutores representativa das vozes e dos sexos da "população" de locutores (adultos do sexo masculino e feminino, com idades próximas e estudantes universitários da UFPB).

Neste trabalho foi utilizada, inicialmente, uma amostra de 20 locutores e, em seguida, essa foi ampliada para 40 locutores satisfazendo assim, as condições mínimas de determinação do tamanho da amostra, principalmente na segunda etapa (n_A=40). Entretanto, para a realização do teste de significância a amostra será considerada "pequena", visto que essa foi subdividida em dois grupos (20 locutores femininos e 20 locutores masculinos), pois é importante analisar a significância dos valores médios obtidos para cada grupo. Com o objetivo de se obter resultados mais precisos, a probabilidade de confiança utilizada é de 99%. As análises são apresentadas a seguir.

6.3.4.1 Avaliação da significância e construção do intervalo de confiança da média

Visando avaliar a significância dos valores médios, assim como estabelecer, com determinado nível de confiança, os intervalos que contêm a verdadeira média populacional, a seguir serão apresentados os intervalos de confiança dos valores médios obtidos nas etapas de pré-identificação e identificação.

A Tabela 6.8 apresenta os resultados referentes à etapa de pré-identificação, que correspondem aos intervalos de confiança para a F_0 média dos locutores femininos e

masculinos, como também para cada um desses grupos. Nesse caso, como cada grupo possui 20 locutores, o número de graus de liberdade é, portanto, 19 e o valor de t correspondente é 2,8609 (Tabela A.35 - Anexo A).

Tabela 6.8: Intervalo de confiança para a Freqüência Fundamental média (em Hz) dos locutores femininos (LF) e masculinos (LM) (Li, $1 \leq i \leq 20$, indica o locutor).

locutores	LF	LM
L1	229,80 ± 6,57	154,88 ± 1,31
L2	217,05 ± 3,21	145,73 ± 4,39
L3	196,06 ± 3,49	142,84 ±2,42
L4	208,97 ± 2,81	127,69 ± 2,48
L5	218,42 ± 4,37	156,91 ± 11,72
L6	199,99 ± 2,04	131,69 ± 1,92
L7	210,14 ± 5,07	129,45 ± 4,47
L8	199,03 ± 6,15	119,97 ± 2,68
L9	238,50 ± 1,94	151,51 ± 1,74
L10	203,16 ± 8,51	133,98 ± 2,50
L11	202,52 ± 6,04	136,84 ± 2,87
L12	195,60 ± 5,26	131,64 ± 14,43
L13	199,41 ± 6,01	123,79 ± 4,73
L14	195,85 ± 9,24	126,87 ± 4,52
L15	199,20 ± 12,18	141,90 ± 3,27
L16	204,24 ± 7,53	136,67 ± 2,03
L17	214,61 ± 3,69	135,02 ± 7,36
L18	207,40 ± 3,73	121,44 ± 2,77
L19	212,06 ± 2,43	141,39 ± 2,75
L20	215,99 ± 13,01	132,47 ± 3,95
F_0 média	208,40 ± 7,38	136,13 ± 6,74

Observa-se, a partir da Tabela 6.8, que o valor da F_0 média de cada um dos locutores, é muito maior que o erro padrão correspondente. Assim, pode-se dizer que a F_0 média de cada locutor é significativa. Este fato pode ser observado até mesmo para os locutores que apresentam erros de classificação quanto ao sexo (LF10, LF15, LF20, LM5 e LM12, que fornecem erros padrões mais elevados que os demais locutores).

O intervalo de confiança de 99% afirma que se tem 99% de certeza de que a amostra selecionada é uma amostra em que a F_0 média de cada locutor, μ, está localizada dentro do intervalo. Essa confiança de 99% significa que, se todas as amostras possíveis de tamanho igual a 20 fossem selecionadas, para cada locutor, 99% dos intervalos conteriam a verdadeira média aritmética do locutor, *em algum lugar* dentro do intervalo. Pode-se concluir, por exemplo, com 99% de confiança, que a F_0 média do locutor LF1 está entre 223,23 e 236,37 Hz e a do locutor LM20 está entre 128,51 e 136,42 Hz.

A última linha da Tabela 6.8 apresenta o intervalo de confiança da F_0 média feminina e masculina, respectivamente. Conclui-se portanto, de acordo com o valor do erro padrão, que este valor médio é bastante significativo, tanto para os locutores femininos, quanto os masculinos. Pode-se concluir, por exemplo, com 99% de confiança, que a F_0 média está entre 201,02 e 215,78 Hz, para os locutores femininos e entre 129,39 e 142,87 Hz, para os locutores masculinos.

A Tabela 6.9 apresenta os resultados referentes a etapa de identificação, que correspondem aos valores do intervalo de confiança da taxa média de identificação, para os locutores femininos e masculinos, assim como para todo o grupo. O valor da taxa média de identificação para os locutores femininos e masculinos, assim como para todo o grupo, são bastante significativas visto que essas taxas médias são muito superiores aos erros padrões obtidos.

Tabela 6.9: Valores do intervalo de confiança para a taxa média de identificação dos locutores femininos (LF), masculinos (LM) e para o grupo.

LF	LM	grupo
97,3% ± 2,4%	98,3% ± 3,0%	97,8% ± 1,9%

As demais taxas médias (falsa aceitação, falsa rejeição e confiabilidade) não foram avaliadas visto que são obtidas a partir da primeira (taxa média de identificação). Portanto, as análises realizadas para a primeira, podem ser estendidas para as demais.

É possível concluir, por exemplo, com 99% de confiança, que a taxa média de identificação dos locutores femininos, nesse sistema, está entre 94,8% e 99,7%.

6.3.4.2 Teste da diferença entre as médias

As aplicações do teste *t*, para este trabalho, estão resumidas na Tabela 6.10. Nas três aplicações tem-se: graus de liberdade igual a 38 (20 + 20 -2), nível de significância,

α, igual a 0,01 (ou probabilidade de confiança igual a 99%), população 1 - locutores masculinos e população 2 - locutores femininos.

Tabela 6.10: Resumo dos resultados obtidos com as aplicações do teste *t*.

aplicação	tipo do teste	hipóteses	estatística do teste	valor crítico
F_0 média	unicaudal	$H_0 : \mu_1 \geq \mu_2$ e $H_1 : \mu_1 < \mu_2$	20,6833	2,4286
Taxa média de pré-identificação	bicaudal	$H_0 : \mu_1 = \mu_2$ e $H_1 : \mu_1 \neq \mu_2$	0,3037	2,7116
Taxa média de identificação	bicaudal	$H_0 : \mu_1 = \mu_2$ e $H_1 : \mu_1 \neq \mu_2$	0,7435	2,7116

A partir da Tabela 6.10 tem-se:

1. F_0 média - a hipótese nula (H_0) é rejeitada porque t=+20,6833 $> t_{38}$=+2,4286. Pode-se concluir que há evidências de que a média aritmética masculina não é maior, nem tão pouco igual, a média aritmética feminina. Observa-se portanto, com uma probabilidade de confiança de 99%, que a F_0 média feminina é superior a masculina.

2. Taxa média de pré-identificação - a hipótese nula (H_0) não é rejeitada porque $t = +0{,}3037 < t_{38} = +2{,}7116$. Pode-se concluir que não há evidências de uma diferença na média aritmética para os dois grupos. Verifica-se, portanto, com uma probabilidade de confiança de 99%, que não se pode afirmar que o desempenho do sistema, no que se refere a pré-identificação, é sensível ao sexo do locutor.

3. Taxa média de identificação - a hipótese nula (H_0) não é rejeitada porque $t = +0{,}7435 < t_{38} = +2{,}7116$. Pode-se concluir que não há evidências de uma diferença na média aritmética para os dois grupos. Verifica-se, portanto, com uma probabilidade de confiança de 99%, que não se pode afirmar que o desempenho do sistema, durante a etapa de identificação, é sensível ao sexo do locutor.

Diante do exposto, pode-se concluir que a aplicação de uma análise estatística de desempenho permite uma maior validação e generalização dos resultados obtidos com a amostra de locutores.

Capítulo 7

Conclusões e Sugestões

7.1 Introdução

A comunicação oral é, sem dúvida alguma, a forma mais natural de comunicação humana. Em virtude da interação homem-máquina se tornar cada vez mais comum, surge uma demanda natural por sistemas capazes de reconhecer o que está sendo dito, bem como quem se está falando [20]. O interesse nessa área se deve ao número de aplicações, bem como a existência de várias questões teóricas que ainda não foram respondidas [21].

Sistemas automáticos de reconhecimento de locutor são, provavelmente, os métodos mais econômicos e naturais para solucionar os problemas de uso autorizado de computadores e sistemas de comunicação e controle de acesso. Com a disponibilidade das linhas telefônicas e microfones acoplados aos computadores, o custo de um sistema de reconhecimento de locutor está relacionado, basicamente, ao projeto do *software*.

Sistemas biométricos reconhecem a pessoa pelo uso de traços (feições) distintos. Sua voz, assim como outras características biométricas, não pode ser esquecida ou perdida, diferentemente dos métodos de controle de acesso baseados em objetos (cartões, chaves, etc.) ou mensagens fornecidas por meio do teclado (senha, etc.). Além disso, os sistemas de reconhecimento de locutor, a partir da fala, podem ser projetados de tal forma que se tornem robustos mesmo diante de ruído e variações do canal [19, 22], de alterações humanas (*e.g.*, resfriados) e de ambientes de gravação [8].

O processo de reconhecimento automático da identidade vocal necessita de precisão,

visto que deverá ser aplicado em situações que exigem a certeza do resultado (*e.g.*, controle de acesso a ambiente restrito). Em virtude desse fato, busca-se obter os melhores métodos com o objetivo de tornar possível a elaboração de um sistema de reconhecimento automático da identidade vocal que consiga representar as características vocais dos locutores, sendo capaz de diferenciá-los de forma eficiente.

Este capítulo apresenta as conclusões deste trabalho, destacando as contribuições relevantes e indica sugestões para trabalhos futuros.

7.2 Sumário da Pesquisa

O trabalho, aqui apresentado, trata da elaboração de um sistema híbrido, que utiliza métodos paramétrico e estatístico, para a identificação automática da identidade vocal (dependente do texto) em um conjunto fechado.

Visando tornar a tarefa de identificação mais eficiente, o sistema é composto de dois estágios: pré-identificação e identificação principal.

No primeiro estágio, os locutores são separados em dois grupos gerais de acordo com o sexo (homens e mulheres), utilizando a freqüência fundamental.

O segundo estágio, a identificação proporiamente dita, é subdividido em duas etapas. Em cada etapa é construído um conjunto de padrões, um padrão para cada locutor.

A primeira etapa da identificação utiliza a Quantização Vetorial (QV) Paramétrica para construção dos padrões representativos dos locutores (vetores-código do dicionário). Os parâmetros são obtidos a partir da Análise por Predição Linear. São avaliados cinco tipos de coeficientes obtidos a partir dessa análise: coeficientes LPC, Cepestrais, Cepestrais Ponderados, Delta Cepestrais e Delta Cepestrais Ponderados. Os coeficientes Cepestrais, seguido dos Delta Cepestrais proporcionam os melhores resultados (o primeiro bastante superior), sendo, portanto, os escolhidos para compor o vetor de características que irá gerar o primeiro padrão representativo de cada locutor.

Para o projeto dos dicionários do quantizador vetorial é realizada uma análise comparativa de três técnicas. A primeira utiliza o algoritmo tradicional LBG [40]. As demais técnicas utilizam os algoritmos KMVVT (Kohonen Modificado com Vizinhança Centrada em Torno do Vetor de Treino) [41, 42] e SSC (Competitivo no Espaço

Sináptico) [43]. O SSC proporcionou os melhores resultados, sendo portanto o escolhido para o projeto dos dicionários.

A regra de decisão utilizada na primeira etapa baseia-se no cálculo de uma medida de distorção (Erro Médio Quadrático). O locutor que proporcionar o menor valor de distorção (desde que maior que um dado limiar, visando impossibilitar o acesso de impostores quando da implementação de um sistema aberto) na comparação do padrão de teste (vetor de características) com os padrões de referência (dicionários) é o locutor identificado pelo sistema.

A segunda etapa do estágio de identificação se caracteriza em um "refinamento" do processo de identificação, sendo utilizada quando os resultados obtidos na primeira etapa (valores de distorção) indicarem similaridade entre os padrões vocais dos locutores (valores das medidas de distorção próximos). Nessa etapa utiliza-se os Modelos de Markov Escondidos (HMMs) Discretos para a construção dos padrões representativos dos locutores, um HMM para cada locutor. A regra de decisão baseia-se no cálculo de uma medida de probabilidade, que corresponde à média aritmética ponderada da probabilidade obtida com a utilização dos coeficientes Cepestrais (peso - 70%) e dos coeficientes Delta Cepestrais (peso - 30%).

7.3 Contribuições

A partir dos resultados obtidos pode-se destacar algumas conclusões e contribuições relevantes deste trabalho, apresentadas a seguir (referentes às etapas de pré-identificação e identificação, respectivamente).

7.3.1 Pré-identificação dos locutores

1. Utilização de um detetor surdo-sonoro, visando a separação prévia dos sons da voz, de forma a tornar mais rápido e preciso o processo de estimação da freqüência fundamental.

2. Avaliação dos parâmetros do detetor surdo-sonoro de forma a verificar que o bom desempenho deste está diretamente associado ao ajuste dos limiares de decisão (energia, número total de picos, diferença de picos, taxa de cruzamento por zero e o coeficiente de correlação normalizado), que depende da forma de aquisição

do sinal. Não há como determinar valores "ótimos" para os limiares, que possam ser utilizados em qualquer ambiente. Esses valores são indicados de uma forma bastante genérica. Tornando-se necessário, portanto, quando da elaboração do sistema, a realização de testes empíricos (dentro dos valores indicados), de forma a determinar os valores para os limiares que melhor se adaptam à aplicação.

3. Desenvolvimento de um método para estimação da freqüência fundamental, que gera valores de F_0 média representativos, com baixas taxas de erro, tanto para os locutores masculinos quanto para os femininos, viabilizando, portanto, o uso da F_0 como parâmetro de separação prévia de locutores em grupos gerais de acordo com o sexo.

4. Um aspecto a destacar, em relação à estimação da F_0, assim como no detetor Surdo-Sonoro, refere-se à escolha dos limiares utilizados para os parâmetros temporais do sinal de voz. Sendo necessário, portanto, quando da elaboração do sistema, a realização de testes empíricos (dentro dos valores indicados), de forma a determinar os valores para os limiares que melhor se adaptam à aplicação.

5. O uso da inferência estatística, que possibilitou, com alto grau de precisão (99%), as seguintes avaliações:

   (a) verificação da significância da F_0 média e estimação o intervalo de confiança que contém a verdadeira F_0 média populacional, para os locutores masculinos e femininos;

   (b) comprovação da diferença acentuada entre os valores da F_0 média para locutores femininos e masculinos, constatando ser a primeira bastante superior à segunda;

   (c) determinação da pouca sensibilidade do desempenho do sistema, no que se refere ao sexo do locutor, mostrando que não existe diferença significativa entre as taxas de pré-identificação dos locutores femininos e masculinos.

A análise estatística possibilitou, portanto, a validação e generalização dos resultados obtidos na etapa de pré-identificação.

7.3.2 Identificação dos locutores

1. A aplicação de uma única sentença para todos os locutores é bastante útil, pois verifica o desempenho do sistema no que se refere à identificação unicamente das características vocais de um locutor, que o diferenciam dos demais, não havendo influência significativa do texto que está sendo dito.

2. Eficiência da Análise por Predição Linear na construção dos parâmetros representativos das características vocais dos locutores. Dentre os tipos de coeficientes utilizados, pode-se concluir que os coeficientes Cepestrais (seguidos dos Delta Cepestrais) são os que apresentam maior eficiência no que se refere à identificação de pessoas a partir de suas vozes, pois proporcionam as maiores taxas médias de identificação e confiabilidade, bem como as menores taxas de médias de falsa rejeição e de falsa aceitação.

3. Verificação da superioridade dos algoritmos que utilizam redes neurais, KMVVT e SSC, no projeto de dicionários de padrões acústicos para identificação de locutor, em relação ao algoritmo LBG. Observa-se também que, dentre os algoritmos KMVVT e SSC, o segundo apresenta desempenho superior, além de ser mais simples a sua implementação.

4. Eficiência da utilização de medidas adicionais na análise de desempenho do sistema. A primeira, denominada confiabilidade (razão entre o número de identificações corretas e o total de identificações), como o próprio nome diz, verifica o nível de confiança da taxa média de identificacão obtida. A segunda, denominada Coeficiente de Variação, que mede a variabilidade dos dados em torno do valor médio (a representatividade do valor médio), visto que, em se tratando da obtenção de medidas de desempenho que correspondem à taxas médias (identificação, falsa aceitação e falsa rejeição), é imprescindível a avaliação da representatividade dessas medidas.

5. Apesar do sistema ter sido desenvolvido para um grupo fechado, foi introduzido um limiar no processo de identificação, visando evitar o acesso de locutores não cadastrados, quando da implementação do grupo aberto.

6. A modelagem por HMM se mostra eficiente na discriminação de locutores com características vocais similares.

7. Observação de que o uso de múltiplas seqüências de observações no projeto do modelo, $\lambda = (\mathcal{A}, \mathcal{B}, \pi)$ para o HMM do tipo "esquerda-direita", proporciona uma melhoria no desempenho do sistema.

8. Aumento da eficiência do sistema com o uso, na fase de identificação (classificação), de uma medida de probabilidade, associada ao HMM, que corresponde à média aritmética ponderada das probabilidades referentes aos HMMs obtidos a partir de coeficientes Cepestrais e Delta Cepestrais, respectivamente. Foi atribuído ao primeiro tipo de coeficiente um peso maior, devido a sua maior eficiência na tarefa de identificação.

9. Desenvolvimento de um sistema de identificação automática da identidade vocal de locutores que utiliza uma técnica híbrida, baseada em análise por predição linear, quantização vetorial, redes neurais (no projeto dos dicionários do quantizador vetorial, utilizando o algoritmo SSC) e Modelos de Markov Escondidos. O uso dessa técnica proporciona altas taxas médias de identificação e confiabilidade, baixas taxas de falsa aceitação e falsa rejeição e baixa variabilidade para esses valores médios, tanto para locutores masculinos quanto femininos.

10. A utilização da etapa de pré-identificação, associada à identificação, apesar de reduzir, um pouco, a taxa média de identificação, proporciona uma diminuição considerável do tempo de processamento quando da identificação de um locutor, visto que este só é comparado com os locutores do mesmo sexo. Este fato se torna ainda mais relevante com a ampliação da base de dados do sistema.

11. A aplicação da inferência estatística permite concluir (com 99% de confiabilidade) que o desempenho do sistema é, de uma forma geral, pouco sensível ao sexo do locutor.

12. Elaboração de uma interface homem-máquina (apresentada no Anexo B) simples e eficiente, que possibilita uma fácil interação entre o usuário e a máquina, tornando mais fácil e atrativa a utilização, bem como a análise do sistema desenvolvido.

Diante do exposto, pode-se concluir que o uso da técnica híbrida para identificação automática da identidade vocal de locutores, se mostra bastante eficiente, sendo capaz de separar, com confiabilidade elevada, inicialmente os locutores em subgrupos, de

acordo com o sexo e, posteriormente, identificá-los corretamente, em seus respectivos subgrupos.

7.4 Sugestões para trabalhos futuros

Sendo a identificação automática da identidade vocal de locutores um problema que exige confiabilidade elevada, por se destinar, principalmente, à aplicações tais como controle de acesso a ambiente restrito pelo uso da senha verbal, vários aspectos deste trabalho podem ser aperfeiçoados visando alcançar tal objetivo. A seguir são apresentadas algumas sugestões para continuidade do trabalho ora apresentado.

1. Ampliação da base de dados do sistema, visando torná-la ainda mais representativa.

2. Desenvolvimento de um sistema formado por um conjunto aberto de locutores, obtendo-se para tanto, um nova base de dados composta por locutores não cadastrados, visando avaliar o desempenho do sistema para aplicações em que o conjunto aberto seja mais adequado.

3. Investigação da utilização de outros parâmetros (*e.g.*, coeficientes Mel Cepestrais e RASTA-PLP [116]) na composição das características acústicas representativas dos locutores.

4. Avaliação, de forma mais criteriosa, da influência da dimensão e do número de níveis do quantizador vetorial no desempenho do sistema.

5. Análise dos efeitos de modificações nos parâmetros que caracterizam o HMM, $\lambda = (\mathcal{A}, \mathcal{B}, \pi)$: N, (número de estados do modelo), M (número de símbolos do alfabeto discreto), $\mathcal{A} = [a_{ij}]$, $1 \leq i, j \leq N$ (matriz de transição de estados), $\mathcal{B}= [b_j(k)]$, $1 \leq j \leq N$ e $1 \leq k \leq M$ (matriz de função de probabilidade das observações) e $\pi = \ \pi_i = P\{q_i|t = 1\}$, $1 \leq i \leq N$ (vetor de probabilidade do estado inicial). Analisar, principalmente, os efeitos das atribuições dos valores iniciais no processo de reestimação de Baum-Welch.

6. Aplicação de novas técnicas para a identificação automática de locutor como, por exemplo, a teoria dos conjuntos nebulosos (*Fuzzy Logic*) [117] e a Transformada *Wavelet* [118].

7. Avaliação da possibilidade de utilização de dois limiares, na tarefa de aceitar ou rejeitar um dado locutor, em substituição ao único limiar utilizado, visando minimizar o efeito da transição abrupta no processo de decisão.

8. Por fim, sugere-se a implementação e otimização do sistema de identificação automática de locutores para aplicações em tempo real, avaliando a possibilidade de implementação de um *hardware* para o sistema.

Diante do exposto, o reconhecimento ou, mais especificamente neste trabalho, a identificação automática 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. Os resultados apresentados neste trabalho satisfizeram essas exigências, obtendo-se, com o sistema, índices de falsa aceitação pequenos para todos os locutores (os impostores não conseguiram se passar pelos locutores verdadeiros) e índices de falsa rejeição também pequenos (os locutores verdadeiros foram aceitos, na maioria das vezes, pelo sistema), com nível de confiabilidade elevado.

O desenvolvimento deste trabalho proporcionou a elaboração de um método para o reconhecimento automático da identidade vocal de locutores, em um grupo fechado, que apresenta grandes variações interlocutor (de forma a evitar a possibilidade dos locutores serem confundidos entre si) e pequenas variações intralocutor. No último caso, significa que as características estabelecidas, para um mesmo locutor são, na maioria das vezes, estáveis ao longo do tempo, insensíveis às variações quanto à maneira de falar, incluindo a velocidade e o nível da elocução, e robustas face às variações na qualidade da voz, devido a causas tais como "disfarce", resfriados e/ou ruídos provenientes do ambiente de gravação. Tem-se, portanto, uma interface vocal homem-máquina que, utilizando um método híbrido para modelagem das características vocais dos locutores, é capaz de discriminá-los de forma eficiente.

Anexo A

Resultados Complementares

A seguir são apresentados alguns resultados complementares do sistema de identificação automática da identidade vocal de locutores, de acordo com a seqüência apresentada a seguir.

A.1 Pré-identificação dos Locutores

A.1.1 Detetor Surdo-Sonoro

1. Parâmetros Temporais do sinal de voz para a palavra *aplausos*, com as respectivas decisões tomadas pelo detetor surdo-sonoro.

A.1.2 Detetor da Freqüência Fundamental

1. Freqüência Fundamental, Freqüência Fundamental média, Coeficiente de Variação e Taxas de Erro, dos locutores femininos (LF1 a LF5) e masculinos (LM1 a LM5), para as cinco elocuções das palavras *aplausos* e *bola*.

2. Freqüência Fundamental, Freqüência Fundamental média, Coeficiente de Variação e Taxas de Erro, dos locutores femininos (LF1 a LF4) e masculinos (LM1 a LM4), para as quarenta e cinco elocuções de todas as sentenças.

3. Freqüência Fundamental, Freqüência Fundamental média, Coeficiente de Variação e Taxas de Erro, dos locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10), para as vinte elocuções da sentença: *Quero usar a máquina.*

4. Freqüência Fundamental, Freqüência Fundamental média, Coeficiente de Variação e Taxas de Erro, dos locutores femininos (LF1 a LF20) e masculinos (LM1 a LM20), para as vinte elocuções da sentença: *Quero usar a máquina*, algoritmo AMDF (AMDF-1).

5. Freqüência Fundamental, Freqüência Fundamental média, Coeficiente de Variação e Taxas de Erro, dos locutores femininos (LF1 a LF20) e masculinos (LM1 a LM20), para as vinte elocuções da sentença: *Quero usar a máquina*, algoritmo AMDF modificado (AMDF-2).

A.2 Identificação dos Locutores

1. Taxas de identificação do SRAL, método QV-LBG (parâmetros acústicos: LPC, CEP, CEP-P, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 e LM10).

2. Taxas de falsa rejeição do SRAL, método QV-LBG (parâmetros acústicos: LPC, CEP, CEP-P, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 e LM10).

3. Taxas de falsa aceitação do SRAL, método QV-LBG (parâmetros acústicos: LPC, CEP, CEP-P, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 e LM10).

4. Taxas de identificação do SRAL, método QV-KMVVT (parâmetros acústicos: LPC, CEP, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 e LM10).

5. Taxas de falsa rejeição do SRAL, método QV-KMVVT (parâmetros acústicos: LPC, CEP, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 e LM10).

6. Taxas de falsa aceitação do SRAL, método QV-KMVVT (parâmetros acústicos: LPC, CEP, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 e LM10).

7. Taxas de identificação, falsa rejeição e falsa aceitação do SRAL, método QV-SSC (parâmetros acústicos: CEP e DCEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 e LM10).

8. Taxas de identificação, falsa rejeição e falsa aceitação do SRAL, método QV-SSC (parâmetros acústicos: CEP e DCEP), para os locutores femininos (LF1 a LF20) e masculinos (LM1 e LM20).

9. Taxas de identificação, falsa rejeição e falsa aceitação do SRAL, método QV-SSC-HMM, para os locutores femininos (LF1 a LF20) e masculinos (LM1 a LM20).

10. Taxas de identificação, falsa rejeição e falsa aceitação do SRAL, método QV-SSC-HMM, adicionada a etapa de pré-identificação, para os locutores femininos (LF1 a LF20) e masculinos (LM1 a LM20).

11. Matriz de confusão do SRAL, método QV-LBG (parâmetro acústico - LPC), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

12. Matriz de confusão do SRAL, método QV-LBG (parâmetro acústico - CEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

13. Matriz de confusão do SRAL, método QV-LBG (parâmetro acústico - CEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

14. Matriz de confusão do SRAL, método QV-LBG (parâmetro acústico - DCEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

15. Matriz de confusão do SRAL, método QV-LBG (parâmetro acústico - DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

16. Matriz de confusão do SRAL, método QV-KMVVT (parâmetro acústico - LPC), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

17. Matriz de confusão do SRAL, método QV-KMVVT (parâmetro acústico - CEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

18. Matriz de confusão do SRAL, método QV-KMVVT (parâmetro acústico - DCEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

19. Matriz de confusão do SRAL, método QV-KMVVT (parâmetro acústico - DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

20. Matriz de confusão do SRAL, método QV-SSC (parâmetro acústico - CEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

21. Matriz de confusão do SRAL, método QV-SSC (parâmetro acústico - DCEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

22. Matriz de confusão do SRAL, método QV-SSC (parâmetro acústico - CEP), para os locutores femininos (LF1 a LF20).

23. Matriz de confusão do SRAL, método QV-SSC (parâmetro acústico - CEP), para os locutores masculinos (LM1 a LM20).

24. Matriz de similaridade do SRAL, método QV-SSC (parâmetro acústico: CEP), dos locutores masculinos e femininos, para as vinte elocuções da sentença (E1 a E20).

25. Matriz de confusão do SRAL, método QV-SSC-HMM, para os locutores femininos (LF1 a LF20).

26. Matriz de confusão do SRAL, método QV-SSC-HMM, para os locutores masculinos (LM1 a LM20).

A.3 Análise estatística de desempenho

1. Tabela da Distribuição *t*-student.

Tabela A.1: Parâmetros Temporais do sinal de voz - *aplausos* (número de quadros = 149, tamanho do quadro = 200, total de amostras lidas = 29.800 - janela utilizada - Hamming).

QUADRO	E_{seg}	ρ_1	TCZ	NTP	DNP	SOM
1	59.246813	-0.057592	104	127	-7	silencio
2	59.731727	0.060750	94	119	-7	silencio
3	60.550200	-0.214977	94	120	-10	silencio
4	61.357202	-0.019860	106	127	-7	silencio
5	60.131226	-0.065488	103	123	-1	silencio
6	61.647816	-0.068789	105	121	-1	silencio
7	59.883489	-0.028079	92	110	2	silencio
8	59.772534	-0.093318	102	120	0	silencio
9	59.546978	-0.067233	105	120	2	silencio
10	61.095411	0.044595	82	104	-6	silencio
11	70.039738	0.839477	49	77	5	silencio
12	92.473193	0.713598	33	55	13	som sonoro
13	95.524738	0.701427	45	50	4	som sonoro
14	97.779094	0.618534	50	53	3	som sonoro
15	98.267923	0.337840	53	57	1	som sonoro
16	98.874017	0.323091	60	63	-3	som sonoro
17	98.574075	0.139333	64	69	3	som sonoro
18	99.395628	0.225468	60	65	5	som sonoro
19	99.683579	0.444397	48	55	-1	som sonoro
20	98.021904	0.493658	31	44	0	som sonoro
21	92.971861	0.816661	34	66	-10	som sonoro
22	80.954957	0.734930	50	88	2	som surdo
23	69.540737	0.661453	53	79	11	silencio
24	66.296796	0.465870	72	90	-10	silencio
25	63.373151	0.310064	89	106	-6	silencio

Continuação da Tabela A.1 ...

QUADRO	E_{seg}	ρ_1	TCZ	NTP	DNP	SOM
26	62.824338	0.138500	89	112	-4	silencio
27	64.827806	0.325095	72	99	-5	silencio
28	65.448459	0.521401	68	91	-3	silencio
29	63.281442	0.416537	80	100	4	silencio
30	63.042304	0.272879	87	105	3	silencio
31	63.640964	0.213872	84	107	-1	silencio
32	62.265514	0.232801	83	109	-1	silencio
33	62.539289	0.210581	84	107	-3	silencio
34	62.831297	0.141372	80	106	-6	silencio
35	87.523149	-0.189635	83	104	-6	som surdo
36	87.734723	0.777994	48	69	3	som surdo
37	95.785060	0.802699	23	43	5	som sonoro
38	97.281395	0.791049	29	46	0	som sonoro
39	96.497081	0.760036	27	46	4	som sonoro
40	95.754421	0.786662	34	47	5	som sonoro
41	95.938160	0.782005	39	48	6	som sonoro
42	98.085602	0.671345	37	50	4	som sonoro
43	100.074760	0.238293	46	67	1	som sonoro
44	99.702617	0.150140	60	78	2	som sonoro
45	100.265377	0.170550	63	76	-2	som sonoro
46	100.777584	0.350965	59	73	3	som sonoro
47	100.540098	0.364407	56	70	0	som sonoro
48	101.326927	0.298310	55	67	7	som sonoro
49	102.010805	0.135356	52	69	7	som sonoro
50	100.969149	0.317160	53	66	8	som sonoro

Continuação da Tabela A.1 ...

QUADRO	E_{seg}	ρ_1	TCZ	NTP	DNP	SOM
51	101.475107	0.304601	56	69	9	som sonoro
52	101.851003	0.280523	57	70	4	som sonoro
53	101.104487	0.177085	51	63	11	som sonoro
54	101.343461	0.389395	46	64	6	som sonoro
55	100.522374	0.345412	41	62	2	som sonoro
56	99.628709	0.435296	49	60	2	som sonoro
57	99.304361	0.287766	55	69	3	som sonoro
58	99.243119	0.122322	59	74	4	som sonoro
59	98.954961	0.030138	52	66	4	som sonoro
60	100.329457	0.355905	43	56	4	som sonoro
61	99.903526	0.276342	38	51	-3	som sonoro
62	99.693449	0.612647	33	52	-8	som sonoro
63	97.874475	0.734659	30	52	-8	som sonoro
64	96.620003	0.760287	28	48	2	som sonoro
65	95.774551	0.786999	19	44	8	som sonoro
66	94.944675	0.851099	25	48	6	som sonoro
67	94.673332	0.820615	27	49	-3	som sonoro
68	93.628164	0.824495	23	46	-2	som sonoro
69	91.812368	0.877258	23	53	7	som sonoro
70	90.351384	0.934690	15	60	12	som sonoro
71	89.446405	0.705118	33	81	1	som surdo
72	91.090807	-0.213974	69	109	-9	som sonoro
73	93.071675	-0.675950	102	133	-11	som sonoro
74	92.787390	-0.754073	125	142	-10	som sonoro
75	98.406231	-0.747070	142	148	-6	som sonoro

Continuação da Tabela A.1 ...

QUADRO	E_{seg}	ρ_1	TCZ	NTP	DNP	SOM
76	96.901748	-0.789807	154	155	-1	som sonoro
77	93.705553	-0.704679	155	156	0	som sonoro
78	95.050815	-0.704982	152	153	-1	som sonoro
79	93.374055	-0.707152	138	144	-6	som sonoro
80	88.779793	-0.492540	112	129	-7	som surdo
81	87.081188	0.173647	72	106	2	som surdo
82	87.045251	0.862393	35	74	6	som surdo
83	87.878745	0.958425	23	57	1	som surdo
84	88.831134	0.952470	20	51	-3	som surdo
85	88.920238	0.958540	19	51	-1	som surdo
86	88.005601	0.961943	18	46	-4	som surdo
87	87.206117	0.965190	25	56	10	som surdo
88	85.311352	0.959628	26	53	13	som surdo
89	83.664504	0.962226	24	60	16	som surdo
90	82.190806	0.912778	24	68	18	som surdo
91	79.757065	0.455653	61	94	4	som surdo
92	85.988147	-0.536862	116	127	-3	som surdo
93	93.498428	-0.597099	134	136	-2	som sonoro
94	94.340660	-0.592805	135	137	-1	som sonoro
95	97.173009	-0.637753	141	142	0	som sonoro
96	96.792147	-0.638983	144	144	0	som sonoro
97	92.090796	-0.667906	146	147	-1	som sonoro
98	93.018424	-0.677043	149	149	1	som sonoro
99	89.084668	-0.683388	148	149	1	som surdo
100	87.776011	-0.744270	146	148	0	som surdo

Continuação da Tabela A.1 ...

QUADRO	E_{seg}	ρ_1	TCZ	NTP	DNP	SOM
101	86.897136	-0.716629	146	149	1	som surdo
102	86.408518	-0.711641	148	152	0	som surdo
103	84.565471	-0.733373	154	157	-3	som surdo
104	85.772629	-0.790203	156	159	-3	som surdo
105	86.013901	-0.736778	160	161	-1	som surdo
106	91.528780	-0.712468	153	153	1	som sonoro
107	92.319170	-0.683533	144	145	-1	som sonoro
108	93.517587	-0.664628	142	142	0	som sonoro
109	90.447691	-0.694748	145	146	0	som sonoro
110	86.787508	-0.629303	141	146	0	som surdo
111	84.972693	-0.620755	137	145	3	som surdo
112	82.562536	-0.529534	136	143	3	som surdo
113	74.329960	-0.323255	126	134	2	som surdo
114	74.470709	0.158722	90	110	-2	som surdo
115	72.093237	0.687257	52	89	-7	som surdo
116	72.348130	0.814484	43	80	0	som surdo
117	70.596754	0.739987	39	71	1	som surdo
118	70.277173	0.769043	38	76	4	som surdo
119	69.711325	0.750306	43	76	2	silencio
120	70.585468	0.797005	44	77	-9	silencio
121	70.218360	0.827811	34	76	-4	silencio
122	69.568905	0.811394	30	72	-6	silencio
123	65.495383	0.569213	52	85	7	silencio
124	69.919275	0.821821	51	86	-6	silencio
125	67.206763	0.761242	37	77	-21	silencio

Continuação da Tabela A.1 ...

QUADRO	E_{seg}	ρ_1	TCZ	NTP	DNP	SOM
126	71.009171	0.911372	36	73	-11	silencio
127	65.775280	0.626975	42	80	-10	silencio
128	66.928666	0.674729	54	90	-10	silencio
129	67.084834	0.701668	44	86	-4	silencio
130	65.005840	0.646606	40	88	-12	silencio
131	64.554891	0.552530	67	97	-7	silencio
132	64.811454	0.719498	71	96	4	silencio
133	63.670940	0.542380	56	89	3	silencio
134	65.160293	0.541972	61	98	-2	silencio
135	67.203627	0.774742	56	101	7	silencio
136	66.353646	0.752441	58	101	13	silencio
137	62.769904	0.445670	76	108	-6	silencio
138	62.302560	0.174750	89	113	-7	silencio
139	61.827489	0.287331	85	110	2	silencio
140	62.632909	0.250518	82	105	7	silencio
141	61.128131	0.219608	77	103	-13	silencio
142	60.941041	0.253935	79	100	-2	silencio
143	60.856979	0.303969	92	110	8	silencio
144	60.906188	0.041888	99	117	3	silencio
145	60.358766	0.168208	94	116	0	silencio
146	61.123184	-0.016521	97	118	-8	silencio
147	61.128258	-0.139050	105	121	-7	silencio
148	60.979597	-0.043080	113	123	-9	silencio
149	61.098619	-0.075240	106	117	-3	silencio

Tabela A.2: Freqüência fundamental (em Hz), Freqüência Fundamental média ($\overline{F_0}$) , Coeficiente de Variação (C.V.) e Taxas de Erro, dos locutores femininos (LF1 a LF5) e masculinos (LM1 a LM5), para as cinco elocuções da palavra *aplausos* (E1 a E5).

	LF1	LF2	LF3	LF4	LF5	LM1	LM2	LM3	LM4	LM5
E1	209,48	224,13	230,66	204,87	223,54	134,90	139,03	137,65	138,24	133,72
E2	233,60	225,55	219,17	220,36	230,26	129,35	158,44	149,39	131,12	126,30
E3	222,83	163,79	218,01	198,60	208,05	126,61	139,30	141,93	131,87	126,86
E4	repetir	222,93	202,10	221,84	207,55	128,32	128,74	139,10	129,64	121,86
E5	220,09	226,33	198,29	218,34	202,98	126,35	107,84	161,96	150,38	114,06
$\overline{F_0}$	221,50	212,55	213,65	212,80	214,48	129,11	134,67	146,01	136,25	124,56
C.V.	4,47%	12,84%	6,23%	4,90%	5,48%	2,69%	13,69%	6,85%	6,28%	5,81%
Erro	0%	20%	0%	0%	0%	0%	0%	0%	0%	0%

Tabela A.3: Freqüência Fundamental (em Hz), Freqüência Fundamental média ($\overline{F_0}$), Coeficiente de Variação (C.V.) e Taxas de Erro, dos locutores femininos (LF1 a LF5) e masculinos (LM1 a LM5), para as cinco elocuções da palavra *bola* (E1 a E5).

	LF1	LF2	LF3	LF4	LF5	LM1	LM2	LM3	LM4	LM5
E1	207,35	192,41	231,95	209,19	227,07	129,88	135,36	137,32	200,76	105,82
E2	207,39	200,67	219,17	220,20	203,93	129,66	122,18	144,08	132,83	110,07
E3	228,64	196,86	226,70	212,13	134,60	124,23	237,41	129,78	134,26	115,19
E4	224,94	206,13	226,87	208,38	206,07	132,40	190,42	132,20	149,47	112,74
E5	224,78	189,25	233,11	210,88	204,13	122,58	127,39	140,92	143,79	146,24
$\overline{F_0}$	218,62	197,06	227,56	212,16	195,16	127,75	162,55	136,86	152,22	118,01
C.V.	4,75%	3,38%	2,42%	2,23%	18,05%	3,25%	30,74%	4,33%	18,39%	13,69%
Erro	0%	0%	0%	0%	20%	0%	40%	0%	20%	0%

repetir - solicitação de repetição da sentença.

Tabela A.4: Freqüência Fundamental (em Hz), Freqüência Fundamental média ($\overline{F_0}$), Coeficiente de Variação (C.V.) e Taxas de Erro, dos locutores femininos (LF1 a LF4) e masculinos (LM1 a LM4), para as quarenta cinco elocuções de todas as sentenças (E1 a E45).

	LF1	LF2	LF3	LF4	LM1	LM2	LM3	LM4
E1	233,61	242,50	209,57	256,16	151,13	123,58	159,00	138,28
E2	232,76	229,99	225,94	259,44	143,78	135,26	131,46	132,15
E3	241,76	242,88	223,17	242,85	142,50	128,86	135,05	134,83
E4	231,14	223,88	223,69	256,88	136,09	126,07	131,24	155,66
E5	227,25	240,51	230,23	256,44	130,12	122,60	161,27	143,26
E6	220,37	239,87	178,68	252,19	160,37	127,84	163,00	155,54
E7	229,87	218,03	195,53	255,88	162,77	116,30	133,23	176,28
E8	212,66	237,21	217,29	255,05	164,21	118,10	161,86	149,53
E9	230,07	237,03	184,97	250,43	162,71	126,64	173,73	188,42
E10	249,87	243,95	192,38	238,85	143,14	122,74	171,68	121,41
E11	205,53	230,88	207,58	240,75	155,84	116,17	147,58	115,01
E12	236,77	204,19	191,81	229,35	133,50	112,92	170,19	112,74
E13	233,97	218,36	197,33	241,40	132,45	119,37	151,93	132,78
E14	220,36	235,91	214,36	236,86	125,30	109,83	157,64	117,75
E15	209,33	228,71	193,62	243,19	124,96	110,80	157,16	112,00
E16	226,67	220,80	240,23	226,29	155,12	128,71	142,67	124,71
E17	219,80	200,57	218,87	201,43	172,31	171,53	146,24	122,51
E18	224,82	208,61	212,65	220,72	162,34	127,79	150,78	122,44
E19	216,33	210,47	223,00	229,67	121,89	128,25	150,78	110,85
E20	225,46	215,79	201,26	256,44	161,21	166,32	149,53	115,92
E21	253,04	237,24	248,73	218,02	169,41	143,58	135,66	135,53
E22	204,49	252,84	221,88	158,76	159,73	132,07	157,00	134,83
E23	205,98	227,37	232,30	230,40	158,23	135,90	134,19	131,74
E24	231,55	213,64	234,71	229,37	202,61	119,10	145,66	116,00
E25	231,34	244,58	180,89	195,74	182,00	165,00	144,46	193,08

Continuação da Tabela A.4 ...

E26	239,01	232,51	234,04	249,97	179,65	195,91	151,97	188,53
E27	258,87	235,45	231,86	244,65	170,74	192,98	196,95	130,99
E28	252,42	209,12	220,79	239,52	189,82	157,33	217,82	210,47
E29	239,26	206,58	244,44	230,39	152,43	177,27	192,52	185,20
E30	244,67	251,99	219,02	243,68	175,22	178,88	201,56	187,95
E31	220,60	237,63	238,02	233,85	132,20	140,34	143,00	118,26
E32	220,42	222,45	247,81	227,58	138,19	128,27	148,12	147,90
E33	180,84	230,30	220,11	229,99	125,59	141,43	148,89	148,75
E34	214,82	229,17	213,92	227,27	137,53	136,75	155,46	173,17
E35	215,28	229,13	257,32	235,14	135,27	106,79	149,71	146,90
E36	220,29	220,86	204,18	252,02	126,37	127,15	136,86	144,94
E37	220,26	222,79	182,40	240,45	126,09	161,78	152,27	155,17
E38	225,74	174,59	199,58	184,92	122,31	125,78	154,45	136,99
E39	228,78	225,42	153,52	209,01	127,25	167,97	144,42	167,60
E40	223,51	173,77	205,01	161,77	124,45	122,98	148,25	157,45
E41	245,19	255,28	234,29	244,99	162,63	159,91	160,35	141,46
E42	248,84	260,90	241,96	259,05	206,04	143,23	202,94	129,08
E43	186,45	246,44	240,50	245,65	180,24	170,23	160,70	130,65
E44	249,70	258,39	236,47	245,26	194,81	129,98	151,73	121,71
E45	242,98	257,82	235,56	248,90	188,72	161,91	181,88	145,83
$\overline{F_0}$	227,39	228,59	216,92	234,15	153,55	139,16	156,95	143,61
C.V.	7,26%	8,41%	10,24%	9,91%	15,33%	16,61%	12,66%	17,57%
Erro	0,0%	4,4%	2,2%	4,4%	17,8%	8,9%	13,3%	13,3%

Tabela A.5: Freqüência Fundamental (em Hz), Freqüência Fundamental média ($\overline{F_0}$), Coeficiente de Variação (C.V.) e Taxas de Erro, dos locutores femininos (LF1 a LF20), para as vinte elocuções da sentença *Quero usar a máquina* (E1 a E20), algoritmo AMDF (AMDF-1).

	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10
E1	233,61	206,47	196,47	218,41	215,00	197,65	211,49	181,17	240,48	187,34
E2	232,78	211,14	157,30	182,06	215,43	198,28	132,07	177,81	232,63	181,65
E3	241,76	205,69	204,50	171,23	repetir	196,68	211,80	189,47	236,48	209,44
E4	231,14	210,48	189,62	209,94	213,72	199,57	160,04	185,12	235,25	181,28
E5	227,25	214,50	186,83	219,20	167,36	195,37	208,07	182,20	235,28	198,29
E6	220,37	219,78	190,69	210,38	210,77	200,25	225,39	171,47	238,92	191,19
E7	229,86	222,43	188,14	210,91	199,08	203,12	220,94	185,71	235,36	189,66
E8	212,66	219,42	195,33	209,93	222,71	201,09	167,29	174,63	244,17	179,79
E9	230,07	219,52	193,42	210,14	198,45	202,12	207,68	196,87	240,88	180,43
E10	249,87	215,91	193,40	183,32	213,22	206,34	199,61	185,39	238,35	169,35
E11	211,14	208,12	185,45	195,27	219,78	202,27	183,87	190,60	240,29	175,85
E12	203,40	127,55	186,47	208,06	209,02	199,16	200,38	186,70	235,51	198,29
E13	209,57	218,97	189,21	200,72	170,53	195,04	197,40	182,30	235,01	181,40
E14	214,16	210,64	187,52	208,55	210,29	189,40	151,85	177,95	240,19	190,78
E15	251,88	219,94	191,30	209,00	211,89	193,04	204,89	179,57	238,18	183,56
E16	234,58	221,62	186,28	203,64	187,37	153,93	199,85	182,03	214,54	188,55
E17	229,96	218,03	194,43	204,13	214,96	191,87	196,31	189,17	240,63	192,38
E18	231,66	223,02	187,85	206,34	216,63	202,74	205,11	185,70	239,78	172,76
E19	221,93	222,09	185,10	203,76	215,28	199,56	201,51	183,89	241,36	173,50
E20	230,23	218,92	182,03	169,26	214,29	204,47	199,99	181,64	239,82	183,84
$\overline{F_0}$	227,39	211,81	188,57	201,71	206,66	196,60	194,28	183,47	238,51	185,47
C.V.	5,6%	9,7%	4,7%	7,1%	7,6%	5,6%	12,2%	3,2%	1,2%	5,3%
Erro	0%	5%	5%	10%	10%	5%	20%	10%	0%	15%

Continuação da Tabela A.5 ...

	LF11	LF12	LF13	LF14	LF15	LF16	LF17	LF18	LF19	LF20
E1	225,27	202,10	196,94	194,26	180,74	193,03	214,22	203,21	216,43	214,19
E2	196,43	190,28	219,55	192,03	196,89	195,87	213,61	212,26	209,16	209,24
E3	201,54	180,16	190,51	200,58	194,03	194,10	209,64	199,06	202,08	204,36
E4	200,27	189,88	188,29	198,76	181,31	198,11	205,95	200,70	207,64	224,27
E5	201,99	208,75	205,37	192,92	180,27	202,32	202,98	207,55	208,13	219,50
E6	180,33	217,30	201,81	193,11	187,95	195,31	209,22	208,85	215,55	225,60
E7	198,21	193,16	198,19	204,53	202,66	204,81	214,42	204,25	208,04	226,89
E8	213,64	215,00	202,91	184,89	186,68	206,19	214,21	210,02	206,27	211,16
E9	197,79	201,99	171,49	208,98	195,47	202,87	212,44	206,92	212,61	222,51
E10	206,27	180,47	189,26	203,44	209,34	199,05	210,94	234,46	212,57	203,88
E11	212,68	187,27	204,15	182,82	180,51	189,18	219,60	218,93	207,40	199,28
E12	189,09	179,08	205,02	194,39	167,25	199,19	205,49	208,25	217,25	208,86
E13	212,74	193,49	207,51	173,31	194,13	197,52	224,22	213,70	217,63	196,68
E14	194,49	190,17	194,63	175,12	186,36	188,59	215,14	202,79	213,66	216,33
E15	184,95	199,73	185,58	183,04	193,51	185,98	214,56	204,85	214,11	202,46
E16	204,90	207,62	197,69	205,93	195,02	196,94	211,12	214,66	206,97	213,82
E17	204,93	194,31	202,47	188,18	208,23	202,86	225,88	205,59	211,11	202,53
E18	205,77	182,17	194,75	205,82	192,66	197,61	211,19	203,72	213,30	198,57
E19	205,85	186,23	202,38	215,69	210,76	201,55	208,19	200,63	214,51	218,65
E20	200,12	188,93	190,92	187,41	202,89	227,45	207,99	187,27	211,31	199,89
$\overline{F_0}$	201,86	194,40	197,47	194,26	192,33	198,93	212,55	207,38	211,29	210,93
C.V.	5,1%	5,9%	5,1%	5,8%	5,8%	4,3%	2,7%	4,5%	2,0%	4,6%
Erro	0%	0%	5%	5%	5%	0%	0%	0%	0%	0%

Tabela A.6: Freqüência Fundamental (em Hz), Freqüência Fundamental média ($\overline{F_0}$), Coeficiente de Variação (C.V.) e Taxas de Erro, dos locutores masculinos (LM1 a LM20), para as vinte elocuções da sentença *Quero usar a máquina* (E1 a E20), algoritmo AMDF (AMDF-1).

	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10
E1	165,06	157,39	135,64	163,93	181,65	134,07	163,02	123,02	153,64	143,82
E2	166,64	149,97	148,65	179,44	168,25	129,25	153,59	116,07	154,38	164,79
E3	156,03	156,15	146,63	170,03	174,79	127,19	147,78	114,28	172,39	147,76
E4	173,63	157,71	146,51	126,60	171,52	138,44	136,81	111,54	156,63	155,04
E5	152,76	136,89	142,96	130,06	170,89	146,44	163,38	124,57	152,59	170,99
E6	173,52	152,69	146,09	124,74	174,11	141,22	143,38	116,72	155,97	170,29
E7	162,04	154,98	144,63	125,61	173,67	153,16	145,59	117,26	154,22	168,56
E8	157,00	156,72	149,50	121,18	165,81	155,19	150,93	125,78	171,56	157,14
E9	168,99	149,18	140,46	122,26	163,64	127,36	144,17	117,27	152,37	151,07
E10	156,80	142,65	142,13	131,15	168,09	151,70	134,27	154,92	153,12	161,93
E11	161,87	150,68	142,49	146,06	182,18	126,71	130,32	117,78	163,03	164,39
E12	156,55	151,45	124,57	126,78	215,87	186,78	134,97	119,67	157,63	211,16
E13	156,03	149,40	138,53	140,20	182,96	147,01	136,14	116,21	164,68	210,48
E14	164,27	170,68	137,39	128,84	187,56	140,23	163,11	111,54	156,95	138,80
E15	167,25	141,38	141,27	128,67	172,82	209,35	149,12	109,15	162,64	169,45
E16	175,86	131,08	145,20	121,93	182,20	137,84	147,26	112,19	156,32	148,03
E17	174,99	155,91	137,39	125,47	174,81	155,87	143,62	111,09	184,31	207,37
E18	154,62	157,67	140,01	127,04	174,46	136,00	138,96	111,29	158,36	146,88
E19	153,38	151,32	144,40	132,25	167,38	130,77	127,97	118,34	155,19	160,33
E20	151,38	141,51	144,04	132,92	173,72	139,05	153,87	119,48	161,75	161,52
$\overline{F_0}$	162,43	150,76	141,93	135,26	176,32	145,68	145,41	118,41	159,89	165,49
C.V.	4,9%	5,8%	3,9%	12,4%	6,4%	14,1%	7,3%	8,2%	5,1%	12,8%
Erro	5%	0%	0%	5%	30%	10%	0%	0%	5%	15%

Continuação da Tabela A.6 ...

	LM11	LM12	LM13	LM14	LM15	LM16	LM17	LM18	LM19	LM20
E1	154,33	135,46	139,66	154,02	166,24	167,00	110,99	160,17	162,63	168,22
E2	139,72	120,01	125,93	161,23	124,89	154,79	116,03	repetir	143,43	138,23
E3	142,88	110,79	135,97	135,21	141,66	147,67	117,25	125,36	140,52	138,08
E4	138,13	120,23	119,10	149,54	138,68	143,85	110,03	122,76	169,98	154,26
E5	141,45	125,17	169,43	160,82	133,10	149,48	120,86	118,12	144,33	136,18
E6	138,02	138,88	149,52	143,55	144,59	158,43	111,42	129,85	156,16	160,89
E7	132,19	126,35	139,36	125,58	262,50	146,56	105,52	124,21	150,68	139,77
E8	148,80	125,96	160,10	115,76	159,71	161,92	110,59	118,68	140,05	166,51
E9	141,40	118,27	116,50	121,94	145,93	132,95	109,64	122,16	155,37	123,84
E10	126,97	122,68	132,42	134,33	141,27	148,73	111,53	119,65	169,06	119,81
E11	0,00	131,77	113,99	125,72	159,11	143,69	113,94	128,48	151,65	145,09
E12	130,22	107,92	117,52	125,83	158,75	164,69	106,17	114,96	156,09	162,24
E13	145,36	113,27	repetir	127,71	152,92	145,38	111,51	122,94	159,75	146,08
E14	143,34	132,45	133,06	129,48	147,97	146,58	105,19	113,57	140,51	149,87
E15	154,33	repetir	118,24	133,37	154,12	159,78	116,14	147,64	151,30	140,55
E16	149,68	114,24	137,24	162,82	169,59	173,15	114,29	126,97	160,92	152,90
E17	134,29	136,44	120,86	138,51	216,40	154,33	106,41	167,30	164,78	158,27
E18	140,32	123,69	117,02	141,72	155,21	174,41	105,06	114,52	162,98	164,39
E19	139,11	122,39	119,37	131,60	168,79	173,85	113,91	115,05	155,17	135,06
E20	147,99	118,36	134,64	123,88	163,58	156,44	107,88	118,42	137,30	132,93
$\overline{F_0}$	134,43	123,39	131,58	137,13	160,25	155,18	111,22	126,88	153,63	146,66
C.V.	24,2%	7,1%	11,8%	10,3%	19,0%	7,4%	4,0%	11,9%	6,5%	9,6%
Erro	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%

Tabela A.7: Freqüência Fundamental (em Hz), Freqüência Fundamental média ($\overline{F_0}$), Coeficiente de Variação (C.V.) e Taxas de Erro, dos locutores femininos (LF1 a LF20), para as vinte elocuções da sentença *Quero usar a máquina* (E1 a E20), algoritmo AMDF modificado (AMDF-2).

	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10
E1	242,03	209,84	199,49	219,15	214,01	197,65	213,60	198,50	238,99	204,97
E2	238,50	212,38	200,72	213,83	215,43	199,59	214,27	195,95	232,63	214,06
E3	242,30	211,83	209,20	208,71	218,08	196,68	216,32	192,04	236,48	227,69
E4	233,31	213,26	190,67	211,85	215,87	199,57	220,06	200,91	234,53	202,05
E5	230,52	218,01	189,45	218,33	214,36	198,29	210,26	219,52	235,2	208,85
E6	222,42	220,92	201,37	209,98	209,83	199,88	218,59	217,00	238,92	203,26
E7	229,77	222,43	190,86	211,02	233,94	203,12	221,18	218,16	235,36	207,14
E8	223,73	220,36	195,33	209,93	228,15	201,09	221,47	195,67	244,17	199,14
E9	230,49	219,52	199,47	209,41	228,81	202,11	208,60	210,05	240,88	182,78
E10	235,52	215,91	196,31	207,19	222,34	206,34	204,00	193,55	238,35	193,43
E11	217,24	206,56	196,59	203,40	218,05	202,27	203,55	193,21	241,42	205,29
E12	208,52	213,59	189,67	209,61	208,96	201,49	206,61	187,47	235,51	217,70
E13	215,16	218,97	190,06	207,08	213,40	201,12	198,69	194,98	235,01	210,10
E14	219,51	210,64	195,26	207,08	213,35	193,91	206,59	198,13	240,19	211,76
E15	248,50	219,94	202,40	209,00	212,75	199,93	206,59	197,58	238,18	199,67
E16	241,02	222,46	190,05	203,96	211,98	197,47	200,18	188,47	241,54	208,57
E17	233,63	216,94	196,59	206,04	222,41	193,80	200,76	198,12	240,63	203,66
E18	231,66	224,59	195,25	206,55	223,53	202,74	204,46	191,51	240,81	195,13
E19	221,93	220,96	190,52	203,39	225,33	198,52	223,77	198,97	241,35	162,94
E20	230,23	221,98	201,86	203,95	217,80	204,19	203,19	190,83	239,82	205,11
$\overline{F_0}$	229,80	217,05	196,06	208,97	218,42	199,99	210,14	199,03	238,50	203,16
C.V.	4,5%	2,3%	2,8%	2,1%	3,1%	1,6%	3,8%	4,8%	1,3%	6,6%
Erro	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%

Continuação da Tabela A.7 ...

	LF11	LF12	LF13	LF14	LF15	LF16	LF17	LF18	LF19	LF20
E1	225,27	200,97	196,94	191,62	217,28	193,03	220,65	203,21	212,89	212,18
E2	196,43	190,28	219,55	188,70	221,16	196,56	219,90	212,26	209,15	139,14
E3	201,54	185,30	190,50	203,86	200,02	199,67	211,55	195,49	205,61	229,56
E4	200,26	191,67	182,25	213,35	186,01	198,11	208,13	217,69	213,62	232,65
E5	201,99	205,41	203,46	192,92	152,23	215,96	205,17	201,19	208,13	218,25
E6	180,33	217,30	207,03	193,11	190,10	226,93	212,13	208,85	211,87	223,10
E7	198,21	196,91	198,19	204,53	200,69	223,13	216,95	203,75	219,52	226,60
E8	213,64	199,56	202,91	184,89	195,56	203,55	213,09	211,81	209,22	227,13
E9	197,79	200,10	215,49	207,82	198,75	210,33	215,38	207,06	217,46	221,31
E10	206,27	194,98	189,98	210,14	213,44	202,44	218,00	206,97	210,38	228,62
E11	212,68	187,37	200,94	182,81	186,79	194,12	221,33	218,93	206,34	217,71
E12	190,94	193,56	210,22	206,13	157,07	198,59	209,25	207,79	216,95	209,57
E13	212,74	194,29	191,38	156,25	220,95	194,57	221,05	213,70	215,31	210,94
E14	200,28	190,17	194,63	175,12	205,80	190,14	211,80	204,07	213,66	222,65
E15	193,32	198,38	190,25	191,98	208,06	187,69	214,67	204,46	214,11	239,26
E16	202,16	206,43	196,32	205,93	191,52	200,66	217,29	212,44	206,97	212,11
E17	204,93	198,76	202,47	199,03	207,36	223,44	227,89	205,59	211,09	207,06
E18	205,77	182,17	201,21	205,82	202,25	204,80	208,89	203,69	213,30	201,77
E19	205,85	188,09	205,42	215,69	226,15	201,13	207,06	200,45	214,51	218,16
E20	200,12	190,33	189,10	187,41	202,83	219,90	212,07	208,68	211,09	222,02
$\overline{F_0}$	202,52	195,60	199,41	195,85	199,20	204,24	214,61	207,40	212,06	215,99
C.V.	4,7%	4,2%	4,7%	7,4%	9,6%	5,8%	2,7%	2,8%	1,8%	9,4%
Erro	0%	0%	0%	5%	10%	0%	0%	0%	0%	5%

Tabela A.8: Freqüência Fundamental (em Hz), Freqüência Fundamental média ($\overline{F_0}$), Coeficiente de Variação (C.V.) e Taxas de Erro, dos locutores masculinos (LM1 a LM20), para as vinte elocuções da sentença *Quero usar a máquina* (E1 a E20), algoritmo AMDF modificado (AMDF-2).

	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10
E1	154,13	154,51	135,64	123,05	196,43	137,91	133,64	123,02	153,64	140,40
E2	150,74	149,97	148,65	122,95	143,83	132,18	134,45	116,07	150,19	138,74
E3	154,42	152,61	146,63	128,41	151,52	131,66	127,80	115,85	149,98	133,42
E4	154,31	153,82	146,51	126,60	149,11	130,98	128,37	118,28	152,64	131,16
E5	153,67	136,89	142,96	130,06	143,82	133,85	122,25	129,08	148,28	135,57
E6	153,83	152,69	146,10	126,55	152,68	127,73	123,39	117,44	148,33	135,22
E7	158,97	142,31	144,63	125,61	147,97	133,66	134,57	117,97	151,62	131,32
E8	157,55	150,49	149,50	121,18	149,77	128,57	136,25	127,96	147,48	135,72
E9	156,92	149,19	140,46	122,26	150,72	134,12	131,56	117,27	149,21	129,06
E10	156,28	142,65	142,13	122,35	154,81	133,83	137,39	121,06	151,61	128,07
E11	152,77	150,68	142,49	128,19	149,62	129,44	116,73	121,29	160,20	143,52
E12	156,23	147,84	142,41	132,86	220,41	128,27	126,63	119,67	151,07	130,75
E13	154,81	147,08	138,53	130,64	153,31	135,31	126,97	116,21	153,54	134,95
E14	154,94	150,62	137,79	130,02	152,86	133,22	141,27	124,15	150,84	133,35
E15	155,13	145,70	141,27	128,67	152,60	133,08	121,02	118,14	150,60	133,91
E16	156,07	131,08	145,20	127,55	154,25	131,03	138,94	112,19	150,46	134,09
E17	156,14	138,77	137,39	133,90	153,85	125,04	122,95	124,54	151,55	135,24
E18	156,43	143,26	140,01	127,83	155,03	131,70	123,05	121,38	152,83	135,70
E19	153,31	133,03	144,40	132,25	153,05	129,06	124,23	118,34	153,09	129,06
E20	151,01	141,51	144,04	132,92	152,64	133,17	137,58	119,48	153,04	130,34
$\overline{F_0}$	154,88	145,73	142,84	127,69	156,91	131,69	129,45	119,97	151,51	133,98
C.V.	1,3%	4,7%	2,6%	3,0%	11,7%	2,3%	5,4%	3,5%	1,8%	2,9%
Erro	0%	0%	0%	0%	10%	0%	0%	0%	0%	0%

Continuação da Tabela A.8 ...

	LM11	LM12	LM13	LM14	LM15	LM16	LM17	LM18	LM19	LM20
E1	143,83	120,06	128,63	133,61	135,14	138,77	144,77	127,96	142,78	136,22
E2	138,61	126,79	120,79	136,93	135,58	137,79	128,68	122,19	139,00	140,29
E3	137,99	140,56	129,86	121,37	135,64	133,94	138,42	125,36	140,31	135,11
E4	139,91	124,00	113,99	131,02	134,60	136,53	135,09	123,15	139,65	136,52
E5	141,74	136,24	122,74	142,46	135,59	138,76	131,14	118,39	143,33	124,83
E6	139,50	122,60	137,13	132,14	145,67	137,58	152,88	128,56	143,48	135,17
E7	130,86	137,49	124,44	123,91	142,78	134,40	132,73	124,21	145,04	127,05
E8	133,66	116,25	114,93	119,10	149,51	141,58	142,55	118,68	139,34	125,86
E9	138,78	124,89	125,42	123,00	145,93	132,95	136,70	122,16	142,77	121,08
E10	134,65	122,52	137,83	117,71	143,09	134,55	127,09	129,09	144,87	131,56
E11	140,37	139,13	122,65	122,85	140,11	137,62	126,75	124,49	136,66	137,65
E12	131,72	124,01	127,10	126,65	138,87	141,76	138,64	117,09	135,99	145,01
E13	132,97	128,17	135,78	119,93	140,86	140,33	158,47	119,82	138,61	135,16
E14	132,94	122,93	118,37	122,93	141,34	136,55	125,42	113,63	140,43	135,67
E15	137,48	121,96	115,06	123,75	138,19	140,54	122,04	121,64	147,17	132,28
E16	147,57	117,83	117,30	139,92	149,96	135,55	145,44	117,79	141,78	131,70
E17	133,88	222,74	126,80	129,18	146,50	129,39	106,50	118,99	145,03	135,84
E18	131,70	132,01	122,97	125,84	142,63	132,93	129,53	122,14	147,70	132,34
E19	133,66	129,76	119,18	124,93	148,19	135,89	142,86	114,97	144,36	128,99
E20	134,94	122,88	114,74	120,24	147,73	135,94	134,70	118,56	129,44	121,00
$\overline{F_0}$	136,84	131,64	123,79	126,87	141,90	136,67	135,02	121,44	141,39	132,47
C.V.	3,3%	17,1%	6,0%	5,6%	3,6%	2,3%	8,5%	3,6%	3,0%	4,7%
Erro	0%	5%	0%	0%	0%	0%	0%	0%	0%	0%

Tabela A.9: Taxas de identificação do SRAL, método QV-LBG (parâmetros acústicos: LPC, CEP, CEP-P, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOCUTOR	LPC	CEP	CEP-P	DCEP	DCEP-P
LF1	70%	75%	65%	80%	40%
LF2	45%	95%	75%	80%	95%
LF3	100%	100%	100%	95%	75%
LF4	100%	100%	100%	90%	85%
LF5	90%	95%	95%	100%	100%
LF6	85%	100%	100%	90%	100%
LF7	95%	95%	85%	90%	60%
LF8	65%	70%	95%	85%	95%
LF9	85%	100%	100%	100%	100%
LF10	85%	100%	100%	95%	95%
LM1	90%	95%	50%	100%	85%
LM2	100%	100%	100%	100%	100%
LM3	90%	100%	100%	100%	100%
LM4	65%	85%	65%	95%	75%
LM5	20%	50%	15%	40%	5%
LM6	95%	100%	100%	100%	100%
LM7	45%	85%	85%	75%	70%
LM8	100%	100%	65%	45%	100%
LM9	100%	100%	100%	95%	95%
LM10	80%	95%	100%	100%	100%

Tabela A.10: Taxas de falsa rejeição do SRAL, método QV-LBG (parâmetros acústicos: LPC, CEP, CEP-P, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOCUTOR	LPC	CEP	CEP-P	DCEP	DCEP-P
LF1	30%	25%	35%	20%	60%
LF2	30%	5%	25%	0%	0%
LF3	0%	0%	0%	0%	0%
LF4	0%	0%	0%	5%	15%
LF5	10%	5%	5%	0%	0%
LF6	0%	0%	0%	0%	0%
LF7	5%	5%	15%	10%	40%
LF8	0%	0%	0%	5%	0%
LF9	5%	0%	0%	0%	0%
LF10	0%	0%	0%	0%	0%
LM1	0%	5%	10%	0%	5%
LM2	0%	0%	0%	0%	0%
LM3	5%	0%	0%	0%	0%
LM4	0%	15%	35%	5%	25%
LM5	80%	50%	85%	55%	90%
LM6	5%	0%	0%	0%	0%
LM7	35%	15%	15%	25%	30%
LM8	0%	0%	10%	0%	0%
LM9	0%	0%	0%	0%	0%
LM10	20%	5%	0%	0%	0%

Tabela A.11: Taxas de falsa aceitação do SRAL, método QV-LBG (parâmetros acústicos: LPC, CEP, CEP-P, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOCUTOR	LPC	CEP	CEP-P	DCEP	DCEP-P
LF1	0%	0%	0%	0%	0%
LF2	25%	0%	0%	20%	5%
LF3	0%	0%	0%	5%	25%
LF4	0%	0%	0%	5%	0%
LF5	0%	0%	0%	0%	0%
LF6	15%	0%	0%	10%	0%
LF7	0%	0%	0%	0%	0%
LF8	35%	30%	5%	10%	5%
LF9	10%	0%	0%	0%	0%
LF10	15%	0%	0%	5%	5%
LM1	10%	0%	0%	0%	10%
LM2	0%	0%	0%	0%	0%
LM3	5%	0%	0%	0%	0%
LM4	35%	0%	0%	0%	0%
LM5	0%	0%	0%	5%	5%
LM6	0%	0%	0%	0%	0%
LM7	20%	0%	0%	0%	0%
LM8	0%	0%	25%	55%	0%
LM9	0%	0%	0%	5%	5%
LM10	0%	0%	0%	0%	0%

Tabela A.12: Taxas de identificação do SRAL, método QV-KMVVT (parâmetros acústicos: LPC, CEP, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOCUTOR	LPC	CEP	DCEP	DCEP-P
LF1	85%	80%	70%	45%
LF2	40%	85%	90%	90%
LF3	100%	100%	75%	80%
LF4	100%	100%	95%	100%
LF5	100%	100%	100%	100%
LF6	85%	100%	100%	100%
LF7	95%	95%	95%	80%
LF8	60%	80%	80%	90%
LF9	90%	100%	100%	100%
LF10	95%	100%	95%	95%
LM1	80%	100%	100%	85%
LM2	100%	100%	100%	95%
LM3	100%	100%	100%	100%
LM4	90%	100%	95%	90%
LM5	75%	100%	55%	45%
LM6	95%	100%	100%	100%
LM7	90%	95%	90%	70%
LM8	100%	100%	30%	90%
LM9	100%	100%	90%	95%
LM10	85%	95%	100%	80%

Tabela A.13: Taxas de falsa rejeição do SRAL, método QV-KMVVT (parâmetros acústicos: LPC, CEP, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOCUTOR	LPC	CEP	DCEP	DCEP-P
LF1	15%	20%	30%	35%
LF2	5%	10%	0%	0%
LF3	0%	0%	0%	0%
LF4	0%	0%	5%	0%
LF5	0%	0%	0%	0%
LF6	0%	0%	0%	0%
LF7	5%	5%	5%	20%
LF8	0%	0%	0%	0%
LF9	0%	0%	0%	0%
LF10	0%	0%	0%	0%
LM1	0%	0%	0%	10%
LM2	0%	0%	0%	5%
LM3	0%	0%	0%	0%
LM4	0%	0%	5%	10%
LM5	25%	0%	40%	55%
LM6	5%	0%	0%	0%
LM7	10%	5%	10%	30%
LM8	0%	0%	0%	0%
LM9	0%	0%	0%	0%
LM10	15%	5%	0%	20%

Tabela A.14: Taxas de falsa aceitação do SRAL, método QV-KMVVT (parâmetros acústicos: LPC, CEP, DCEP e DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOCUTOR	LPC	CEP	DCEP	DCEP-P
LF1	0%	0%	0%	20%
LF2	55%	5%	10%	10%
LF3	0%	0%	25%	20%
LF4	0%	0%	0%	0%
LF5	0%	0%	0%	0%
LF6	15%	0%	0%	0%
LF7	0%	0%	0%	0%
LF8	40%	20%	20%	10%
LF9	10%	0%	0%	0%
LF10	5%	0%	5%	5%
LM1	20%	0%	0%	5%
LM2	0%	0%	0%	0%
LM3	0%	0%	0%	0%
LM4	10%	0%	0%	0%
LM5	0%	0%	5%	0%
LM6	0%	0%	0%	0%
LM7	0%	0%	0%	0%
LM8	0%	0%	70%	10%
LM9	0%	0%	10%	5%
LM10	0%	0%	0%	0%

Tabela A.15: Taxas de identificação, falsa rejeição e falsa aceitação do SRAL, método QV-SSC (parâmetros acústicos: CEP e DCEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOCUTOR	identificação		falsa rejeição		falsa aceitação	
	CEP	DCEP	CEP	DCEP	CEP	DCEP
LF1	90%	60%	10%	40%	0%	0%
LF2	100%	95%	0%	0%	0%	5%
LF3	100%	90%	0%	0%	0%	10%
LF4	100%	95%	0%	5%	0%	0%
LF5	100%	100%	0%	0%	0%	0%
LF6	100%	100%	0%	0%	0%	0%
LF7	100%	80%	0%	20%	0%	0%
LF8	85%	85%	0%	0%	15%	15%
LF9	100%	100%	0%	0%	0%	0%
LF10	100%	100%	0%	0%	0%	0%
LM1	95%	95%	5%	0%	0%	5%
LM2	100%	100%	0%	0%	0%	0%
LM3	100%	100%	0%	0%	0%	0%
LM4	100%	95%	0%	5%	0%	0%
LM5	90%	25%	10%	75%	0%	0%
LM6	100%	100%	0%	0%	0%	0%
LM7	95%	80%	5%	20%	0%	0%
LM8	100%	70%	0%	0%	0%	30%
LM9	100%	100%	0%	0%	0%	0%
LM10	100%	100%	0%	0%	0%	0%

Tabela A.16: Taxas de identificação, falsa rejeição e falsa aceitação do SRAL, método QV-SSC (parâmetros acústicos: CEP e DCEP), para os locutores femininos (LF1 a LF20) e masculinos (LM1 a LM20).

LOCUTOR	identificação		falsa rejeição		falsa aceitação	
	CEP	DCEP	CEP	DCEP	CEP	DCEP
LF1	90%	55%	10%	40%	0%	5%
LF2	100%	95%	0%	0%	0%	5%
LF3	100%	90%	0%	0%	0%	10%
LF4	100%	95%	0%	5%	0%	0%
LF5	100%	100%	0%	0%	0%	0%
LF6	100%	100%	0%	0%	0%	0%
LF7	95%	80%	5%	20%	0%	0%
LF8	85%	75%	0%	0%	15%	25%
LF9	100%	100%	0%	0%	0%	0%
LF10	100%	100%	0%	0%	0%	0%
LF11	100%	80%	0%	0%	0%	20%
LF12	100%	95%	0%	0%	0%	5%
LF13	100%	100%	0%	0%	0%	0%
LF14	100%	95%	0%	0%	0%	5%
LF15	100%	100%	0%	0%	0%	0%
LF16	100%	100%	0%	0%	0%	0%
LF17	100%	100%	0%	0%	0%	0%
LF18	95%	85%	0%	0%	5%	15%
LF19	100%	100%	0%	0%	0%	0%
LF20	100%	100%	0%	0%	0%	0%

Continuação da Tabela A.16 ...

LOCUTOR	identificação		falsa rejeição		falsa aceitação	
	CEP	DCEP	CEP	DCEP	CEP	DCEP
LM1	95%	90%	5%	0%	0%	10%
LM2	100%	100%	0%	0%	0%	0%
LM3	100%	100%	0%	0%	0%	0%
LM4	100%	95%	0%	5%	0%	0%
LM5	90%	25%	10%	75%	0%	0%
LM6	100%	100%	0%	0%	0%	0%
LM7	95%	80%	5%	20%	0%	0%
LM8	100%	70%	0%	0%	0%	30%
LM9	100%	100%	0%	0%	0%	0%
LM10	100%	100%	0%	0%	0%	0%
LM11	100%	100%	0%	0%	0%	0%
LM12	100%	75%	0%	0%	0%	25%
LM13	100%	100%	0%	0%	0%	0%
LM14	100%	100%	0%	0%	0%	0%
LM15	100%	90%	0%	0%	0%	10%
LM16	100%	100%	0%	0%	0%	0%
LM17	90%	100%	0%	0%	10%	0%
LM18	95%	100%	0%	0%	5%	30%
LM19	100%	100%	0%	0%	0%	0%
LM20	100%	95%	0%	0%	0%	5%

Tabela A.17: Taxas de identificação, falsa rejeição e falsa aceitação do SRAL, método QV-SSC-HMM, para os locutores femininos (LF1 a LF20) e masculinos (LM1 a LM20).

LOCUTOR	identificação	falsa rejeição	falsa aceitação
LF1	90%	10%	0%
LF2	100%	0%	0%
LF3	100%	0%	0%
LF4	100%	0%	0%
LF5	100%	0%	0%
LF6	100%	0%	0%
LF7	95%	5%	0%
LF8	90%	0%	10%
LF9	100%	0%	0%
LF10	100%	0%	0%
LF11	100%	0%	0%
LF12	100%	0%	0%
LF13	100%	0%	0%
LF14	100%	0%	0%
LF15	100%	0%	0%
LF16	100%	0%	0%
LF17	100%	0%	0%
LF18	95%	0%	5%
LF19	100%	0%	0%
LF20	100%	0%	0%

Continuação da Tabela A.17 ...

LOCUTOR	identificação	falsa rejeição	falsa aceitação
LM1	95%	5%	0%
LM2	100%	0%	0%
LM3	100%	0%	0%
LM4	100%	0%	0%
LM5	90%	10%	0%
LM6	100%	0%	0%
LM7	95%	5%	0%
LM8	100%	0%	0%
LM9	100%	0%	0%
LM10	100%	0%	0%
LM11	100%	0%	0%
LM12	100%	0%	0%
LM13	100%	0%	0%
LM14	100%	0%	0%
LM15	100%	0%	0%
LM16	100%	0%	0%
LM17	100%	0%	0%
LM18	100%	0%	0%
LM19	100%	0%	0%
LM20	100%	0%	0%

Tabela A.18: Taxas de identificação, falsa rejeição e falsa aceitação do SRAL, método QV-SSC-HMM, adicionada a etapa de pré-identificação, para os locutores femininos (LF1 a LF20) e masculinos (LM1 a LM20).

LOCUTOR	identificação	falsa rejeição	falsa aceitação
LF1	90%	10%	0%
LF2	100%	0%	0%
LF3	100%	0%	0%
LF4	100%	0%	0%
LF5	100%	0%	0%
LF6	100%	0%	0%
LF7	95%	5%	0%
LF8	90%	0%	10%
LF9	100%	0%	0%
LF10	95%	5%	0%
LF11	100%	0%	0%
LF12	100%	0%	0%
LF13	100%	0%	0%
LF14	95%	5%	0%
LF15	90%	5%	5%
LF16	100%	0%	0%
LF17	100%	0%	0%
LF18	95%	0%	5%
LF19	100%	0%	0%
LF20	95%	0%	5%

Continuação da Tabela A.18 ...

LOCUTOR	identificação	falsa rejeição	falsa aceitação
LM1	95%	5%	0%
LM2	100%	0%	0%
LM3	100%	0%	0%
LM4	100%	0%	0%
LM5	80%	20%	0%
LM6	100%	0%	0%
LM7	95%	5%	0%
LM8	100%	0%	0%
LM9	100%	0%	0%
LM10	100%	0%	0%
LM11	100%	0%	0%
LM12	95%	0%	5%
LM13	100%	0%	0%
LM14	100%	0%	0%
LM15	100%	0%	0%
LM16	100%	0%	0%
LM17	100%	0%	0%
LM18	100%	0%	0%
LM19	100%	0%	0%
LM20	100%	0%	0%

Tabela A.19: Matriz de confusão do SRAL, método QV-LBG (parâmetro acústico - LPC), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	06
LF2	0	09	0	0	0	0	0	0	05	0	0	0	0	0	0	0	0	0	0	0	06
LF3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF5	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	02
LF6	0	0	01	0	0	17	0	0	0	02	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF8	01	0	0	0	0	0	0	13	0	06	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	01	0	0	0	0	17	01	0	0	0	0	0	0	0	0	0	0	01
LF10	03	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	02	0	0	0
LM2	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM3	0	0	0	0	0	0	0	0	0	0	01	0	18	0	0	0	0	0	0	0	01
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	07	0	0	0
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	07	0	0	0	0	0	13
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	01
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	04	09	0	0	0	07
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
LM9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	04

Tabela A.20: Matriz de confusão do SRAL, método QV-LBG (parâmetro acústico - CEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	05
LF2	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF5	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF8	0	0	02	0	0	0	0	14	0	04	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	01
LM2	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM3	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	03
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	06
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	03
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
LM9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	01

Tabela A.21: Matriz de confusão do SRAL, método QV-LBG (parâmetro acústico - CEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	07
LF2	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	05
LF3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF5	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	03
LF8	01	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	08	0	0	02
LM2	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM3	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	07
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	06	0	0	0	0	0	14
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	0	0	0	03
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	05	0	0	0	13	0	0	02
LM9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0

Tabela A.22: Matriz de confusão do SRAL, método QV-LBG (parâmetro acústico - DCEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	04
LF2	0	16	0	0	0	0	0	0	04	0	0	0	0	0	0	0	0	0	0	0	0
LF3	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01	0	0
LF4	0	0	0	18	0	0	0	0	01	0	0	0	0	0	0	0	0	0	0	0	01
LF5	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF6	0	0	0	0	0	18	0	0	0	01	0	0	0	0	0	0	0	0	01	0	0
LF7	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	02
LF8	01	0	0	0	0	0	0	17	01	0	0	0	0	0	0	0	0	0	0	0	01
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	01	0	19	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
LM2	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM3	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	01
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	10
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	05
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	03	01	0	0	09	0	07	0
LM9	0	0	01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0

Tabela A.23: Matriz de confusão do SRAL, método QV-LBG (parâmetro acústico - DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12
LF2	0	19	0	0	0	01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	0	0	15	0	0	05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	03
LF5	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0	08
LF8	0	0	01	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	01	0	19	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	02	01
LM2	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM3	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	05
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	03	0	0	0	0	0	17
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	06
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
LM9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	01	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0

Tabela A.24: Matriz de confusão do SRAL, método QV-KMVVT (parâmetro acústico - LPC), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	03
LF2	0	08	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF5	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF6	0	0	02	0	0	17	0	01	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF8	0	0	01	0	0	0	0	12	0	07	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	01	01	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0
LF10	01	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	04	0	0	0
LM2	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM3	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	02	0	0	0
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	05
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	01
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	02
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
LM9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	03

Tabela A.25: Matriz de confusão do SRAL, método QV-KMVVT (parâmetro acústico - CEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	04
LF2	0	17	0	0	01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	02
LF3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF5	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF8	01	0	02	0	0	0	0	16	0	01	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
LM2	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM3	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	01
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
LM9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	01

Tabela A.26: Matriz de confusão do SRAL, método QV-KMVVT (parâmetro acústico - DCEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	06
LF2	01	18	0	0	0	0	0	0	01	0	0	0	0	0	0	0	0	0	0	0	0
LF3	0	0	15	0	0	05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF5	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF8	04	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	01	0	19	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	04	0	0	0
LM2	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM3	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	01
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	01	08
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	02
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	06	0	13	01
LM9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01	01	18	0	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0

Tabela A.27: Matriz de confusão do SRAL, método QV-KMVVT (parâmetro acústico - DCEP-P), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	09	0	0	0	0	01	01	01	0	01	0	0	0	0	0	0	0	0	0	0	07
LF2	0	18	0	0	0	02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	0	0	16	0	0	04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF5	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF8	0	0	0	0	0	02	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	01	0	19	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	02
LM2	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	01
LM3	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	02
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	09	0	0	0	0	0	11
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	01	0	0	14	0	0	0	05
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	01	01	0	0	18	0	0	0
LM9	0	0	0	0	0	0	01	0	0	0	0	0	0	0	0	0	0	0	19	0	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	04

Tabela A.28: Matriz de confusão do SRAL, método QV-SSC (parâmetro acústico - CEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	02
LF2	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF5	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF8	02	0	01	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	01
LM2	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM3	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	02
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	01
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
LM9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0

Tabela A.29: Matriz de confusão do SRAL, método QV-SSC (parâmetro acústico - DCEP), para os locutores femininos (LF1 a LF10) e masculinos (LM1 a LM10).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	frej
LF1	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	08
LF2	0	19	0	0	0	0	0	0	01	0	0	0	0	0	0	0	0	0	0	0	0
LF3	0	0	18	0	0	02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF5	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	04
LF8	0	0	0	0	0	0	0	17	01	0	0	0	0	0	0	0	0	0	01	0	01
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
LM1	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	01	0	0	0
LM2	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM3	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM4	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	01
LM5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	05	0	0	0	0	0	15
LM6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	04
LM8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	06	0
LM9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LM10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0

Tabela A.30: Matriz de confusão do SRAL, método QV-SSC (parâmetro acústico - CEP), para os locutores femininos (LF1 a LF20).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LF11	LF12	LF13	LF14	LF15	LF16	LF17	LF18	LF19	LF20	frej
LF1	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	02
LF2	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF5	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF8	02	0	01	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
LF11	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
LF12	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LF13	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LF14	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0
LF15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0
LF16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LF17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
LF18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	01	0	0
LF19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LF20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0

Tabela A.31: Matriz de confusão do SRAL, método QV-SSC (parâmetro acústico - CEP), para os locutores masculinos (LM1 a LM20).

LOC	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	LM11	LM12	LM13	LM14	LM15	LM16	LM17	LM18	LM19	LM20	frej
LM1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LM2	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LM3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LM4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LM5	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	02
LM6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LM7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LM8	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
LM9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LM10	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
LM11	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
LM12	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM13	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM14	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0
LM15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0
LM16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01	0	18	01	0	0	0
LM18	0	0	0	0	0	0	0	0	0	0	0	01	0	0	0	0	0	19	0	0	0
LM19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LM20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0

Tabela A.32: Matriz de similaridade do SRAL, método QV-SSC (parâmetro acústico: CEP), dos locutores masculinos e femininos, para as vinte elocuções da sentença (E1 a E20).

ELOC	LF3	LF8	LF11	LF13	LF18	LF20	LM8	LM17	LM18
E1	0	0	0	0	LF12,LF15,LF16,LF19,LF20,LM11,LM12,LM14,LM15,LM16,LM17,LM19	0	0	LM13,LM15	0
E2	0	0	0	0	0	0	0	LM15	LM12
E3	0	0	0	0	0	0	0	0	0
E4	0	LF1,LF3,LF10	0	0	0	0	0	0	0
E5	0	0	0	0	0	0	0	LM15	0
E6	0	0	0	0	0	0	0	LM15	0
E7	0	LF10	LF13	0	0	0	0	LM13,LM15,LM18	0
E8	0	LF1,LF3,LF10	0	0	0	0	0	LM13,LM15,LM17,LM18	0
E9	LF6,LF8	LF3	LF16	LF18	0	0	0	LM13,LM18	0
E10	0	LF1,LF3,LF5,LF6,LF10	LF16	0	0	LF11,LF16	0	LM13,LM18	0
E11	0	0	0	0	0	0	0	LM10,LM13,LM15,LM20	0
E12	0	0	0	0	0	0	0	LM13,LM20	0
E13	0	LF1,LF10	0	0	0	0	0	0	0
E14	0	LF1	0	0	0	0	0	LM15	0
E15	0	0	0	0	0	0	0	0	0
E16	0	0	0	0	0	0	0	0	0
E17	0	0	0	0	0	0	0	0	0
E18	0	0	0	0	LF12,LF13,LF16	0	0	LM13	0
E19	0	LF1,LF3,LF5,LF6,LF10	0	0	0	0	0	LM15	0
E20	0	LF1,LF3,LF10	0	0	0	0	LM4	LM15	0

Tabela A.33: Matriz de confusão do SRAL, método QV-SSC-HMM, para os locutores femininos (LF1 a LF20).

LOC	LF1	LF2	LF3	LF4	LF5	LF6	LF7	LF8	LF9	LF10	LF11	LF12	LF13	LF14	LF15	LF16	LF17	LF18	LF19	LF20	frej
LF1	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	02
LF2	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF5	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LF7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LF8	02	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0
LF9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LF10	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
LF11	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
LF12	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LF13	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LF14	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0
LF15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0
LF16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LF17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
LF18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	01	0	0
LF19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LF20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0

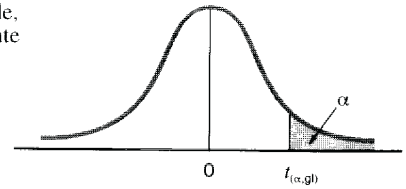
Tabela A.34: Matriz de confusão do SRAL, método QV-SSC-HMM, para os locutores masculinos (LM1 a LM20).

LOC	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	LM11	LM12	LM13	LM14	LM15	LM16	LM17	LM18	LM19	LM20	frej
LM1	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LM2	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LM3	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LM4	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LM5	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	02
LM6	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LM7	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	01
LM8	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
LM9	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
LM10	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
LM11	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0
LM12	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0
LM13	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0
LM14	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0
LM15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0
LM16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0
LM17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
LM18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0
LM19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0
LM20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0

Tabela A.35: Distribuição *t*-Student.

Valores Críticos de *t*

Para um determinado número de graus de liberdade, o dado representa o valor crítico de *t* correspondente a uma determinada área da cauda superior, (α).



Graus de Liberdade	Áreas da Cauda Superior (α)					
	0,25	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**
1	1,0000	3,0777	6,3138	12,7062	31,8207	63,6574
2	0,8165	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,7649	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,7407	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,7267	1,4759	2,0150	2,5706	3,3649	4,0322
6	0,7176	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,7111	1,4149	1,8946	2,3646	2,9980	3,4995
8	0,7064	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,7027	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,6998	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,6974	1,3634	1,7959	2,2010	2,7181	3,1058
12	0,6955	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,6938	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,6924	1,3450	1,7613	2,1448	2,6245	2,9768
15	0,6912	1,3406	1,7531	2,1315	2,6025	2,9467
16	0,6901	1,3368	1,7459	2,1199	2,5835	2,9208
17	0,6892	1,3334	1,7396	2,1098	2,5669	2,8982
18	0,6884	1,3304	1,7341	2,1009	2,5524	2,8784
19	0,6876	1,3277	1,7291	2,0930	2,5395	2,8609
20	0,6870	1,3253	1,7247	2,0860	2,5280	2,8453
21	0,6864	1,3232	1,7207	2,0796	2,5177	2,8314
22	0,6858	1,3212	1,7171	2,0739	2,5083	2,8188
23	0,6853	1,3195	1,7139	2,0687	2,4999	2,8073
24	0,6848	1,3178	1,7109	2,0639	2,4922	2,7969
25	0,6844	1,3163	1,7081	2,0595	2,4851	2,7874
26	0,6840	1,3150	1,7056	2,0555	2,4786	2,7787
27	0,6837	1,3137	1,7033	2,0518	2,4727	2,7707
28	0,6834	1,3125	1,7011	2,0484	2,4671	2,7633
29	0,6830	1,3114	1,6991	2,0452	2,4620	2,7564
30	0,6828	1,3104	1,6973	2,0423	2,4573	2,7500
31	0,6825	1,3095	1,6955	2,0395	2,4528	2,7440
32	0,6822	1,3086	1,6939	2,0369	2,4487	2,7385
33	0,6820	1,3077	1,6924	2,0345	2,4448	2,7333
34	0,6818	1,3070	1,6909	2,0322	2,4411	2,7284
35	0,6816	1,3062	1,6896	2,0301	2,4377	2,7238
36	0,6814	1,3055	1,6883	2,0281	2,4345	2,7195
37	0,6812	1,3049	1,6871	2,0262	2,4314	2,7154
38	0,6810	1,3042	1,6860	2,0244	2,4286	2,7116
39	0,6808	1,3036	1,6849	2,0227	2,4258	2,7079
40	0,6807	1,3031	1,6839	2,0211	2,4233	2,7045
41	0,6805	1,3025	1,6829	2,0195	2,4208	2,7012
42	0,6804	1,3020	1,6820	2,0181	2,4185	2,6981
43	0,6802	1,3016	1,6811	2,0167	2,4163	2,6951
44	0,6801	1,3011	1,6802	2,0154	2,4141	2,6923
45	0,6800	1,3006	1,6794	2,0141	2,4121	2,6896
46	0,6799	1,3002	1,6787	2,0129	2,4102	2,6870
47	0,6797	1,2998	1,6779	2,0117	2,4083	2,6846
48	0,6796	1,2994	1,6772	2,0106	2,4066	2,6822
49	0,6795	1,2991	1,6766	2,0096	2,4049	2,6800
50	0,6794	1,2987	1,6759	2,0086	2,4033	2,6778

Anexo B

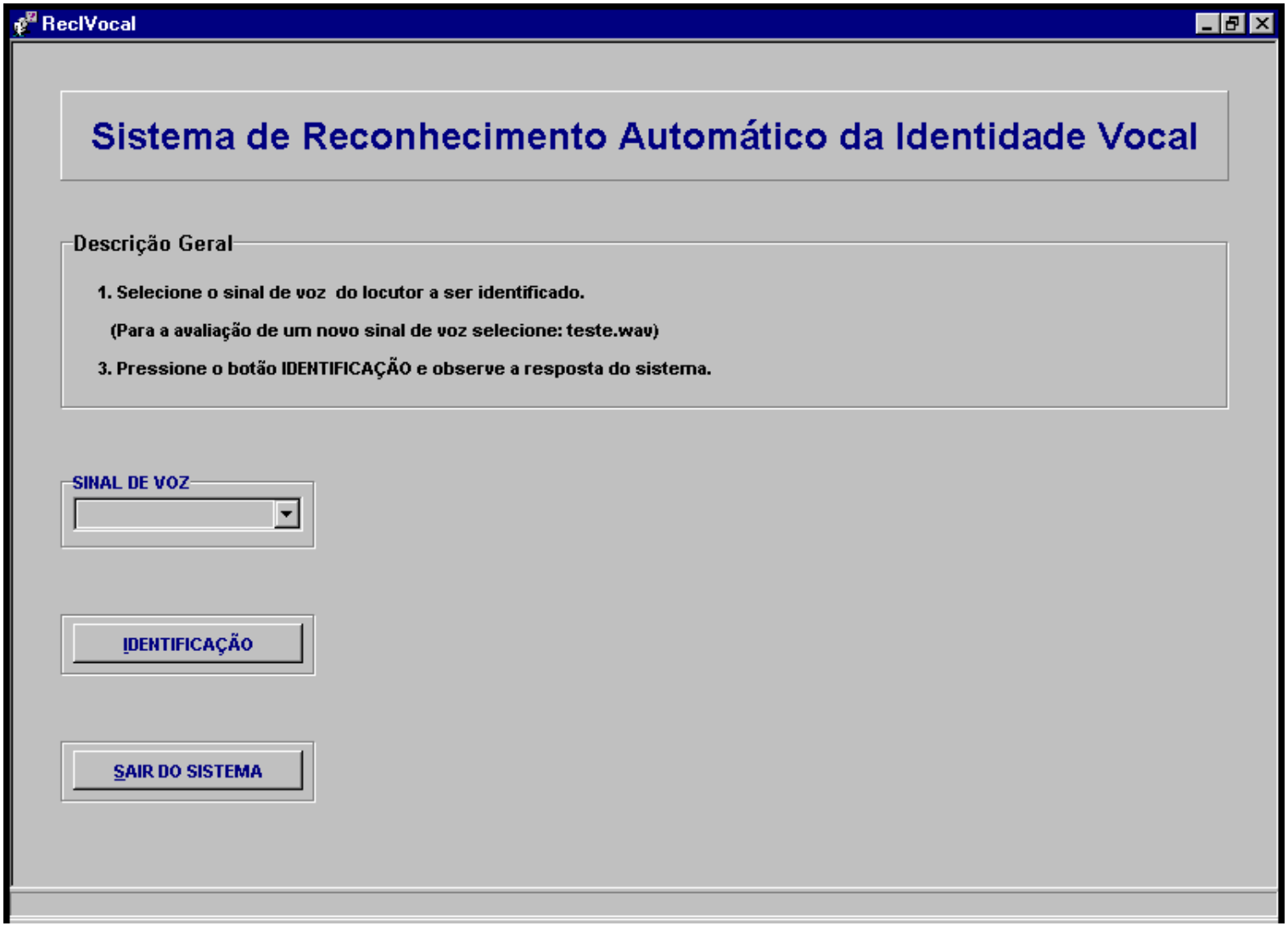
Interface do Sistema

A interface utilizada no sistema de reconhecimento (identificação) automático da identidade vocal de locutores apresentado neste trabalho, foi implementada utilizando a linguagem de programação C++ Builder.

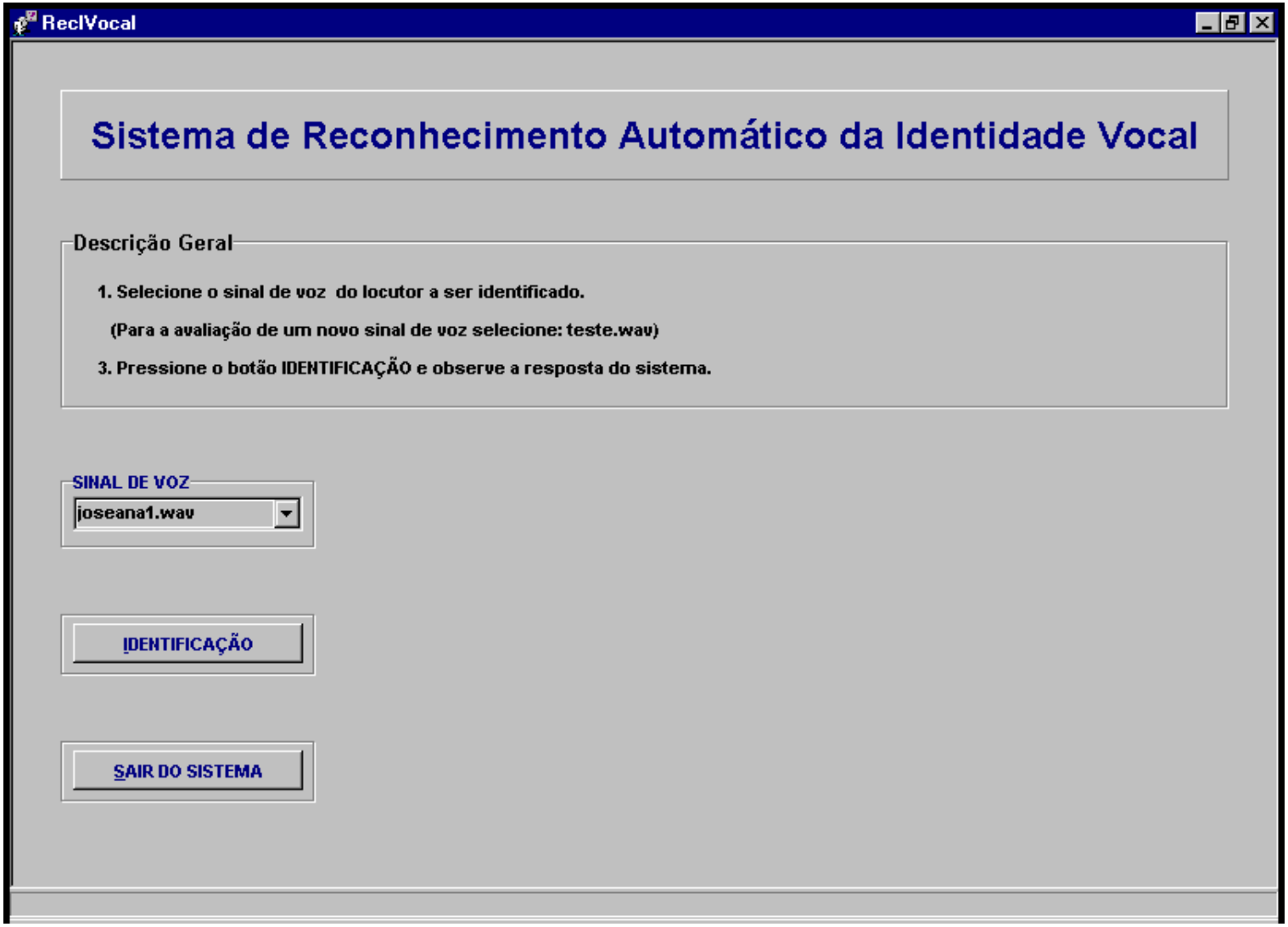
Nas páginas seguintes, serão apresentadas as "Telas" correspondentes à cada opção do sistema de acordo com a seguinte seqüência:

1. "Tela" principal - Descrição geral;
2. Escolha do sinal de voz (ou elocução da sentença de acesso) do locutor a ser identificado;
3. Processamento da informação;
4. Resposta do Sistema.

A resposta do sistema é fornecida nas formas textual, gráfica e vocal. As respostas textuais possíveis são: USUÁRIO "nome" CADASTRADO, ACESSO PERMITIDO; USUÁRIO NÃO CADASTRADO, ACESSO NEGADO e REPITA A SENTENÇA, se o locutor é aceito, rejeitado ou se é solicitada a repetição da sentença, respectivamente. A resposta textual evidencia também o sexo do locutor, sendo utilizado o termo "USUÁRIO" ou "USUÁRIA", se o locutor é do sexo masculino ou feminino, respectivamente. As respostas vocais, associadas a cada uma das respostas textuais, são: "acesso permitido", "acesso negado" e "repita a sentença", respectivamente. A resposta gráfica utiliza um semáforo para indicar se o locutor é aceito (luz verde), rejeitado (luz vermelha) ou se é solicitada a repetição da sentença (luz amarela).



"Tela" principal: descrição geral do sistema.



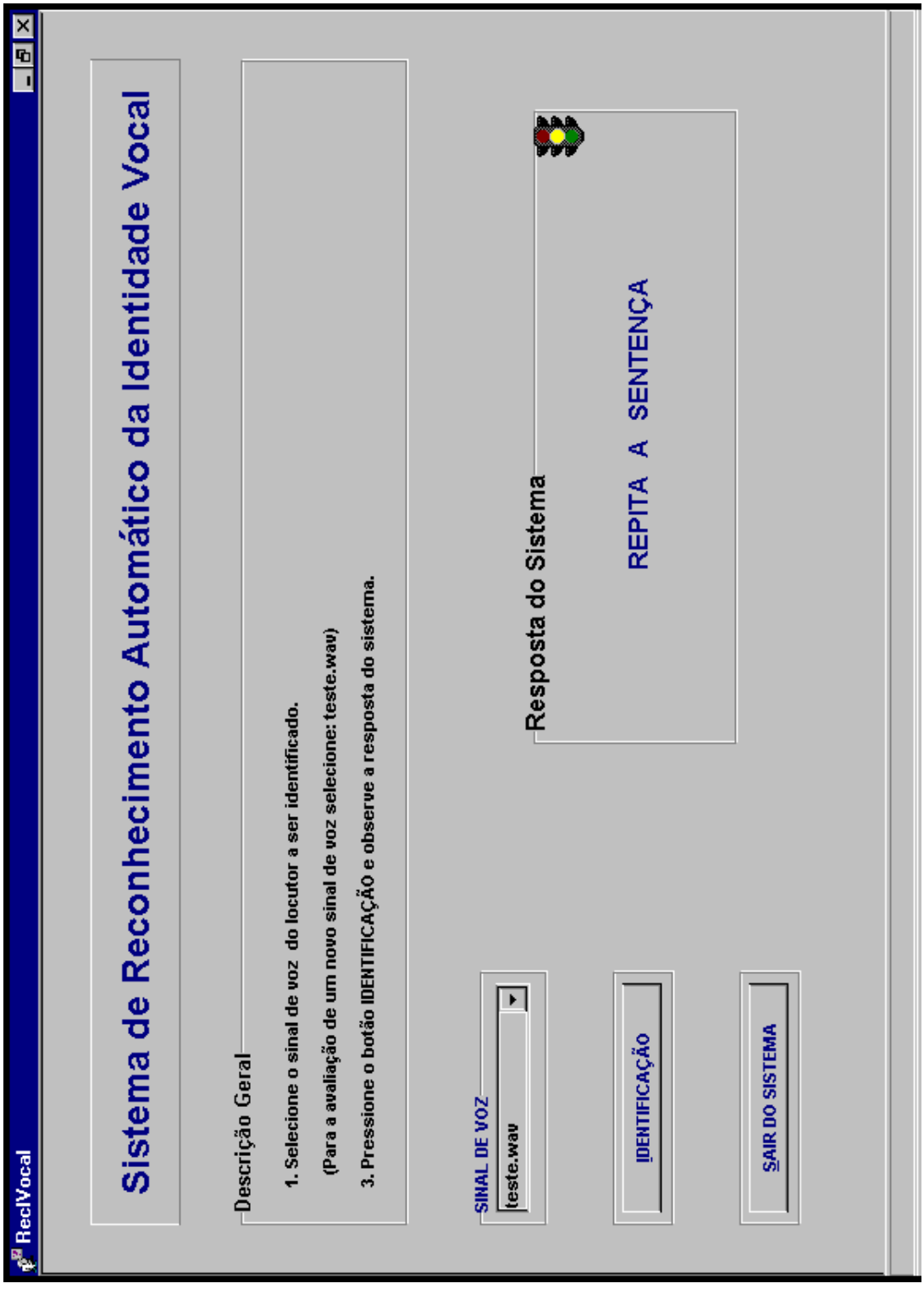
Escolha do sinal de voz do locutor a ser identificado.

Elocução da sentença para identificação do locutor.

Processamento da informação.

Resposta do sistema: locutor identificado.

Resposta do sistema: locutor rejeitado.



Resposta do sistema: solicitação de repetição da sentença.

Bibliografia

[1] Rabiner, L. R. and Schafer, R. W. *Digital Processing of Speech Signals*. Prentice Hall, Upper Saddle River, New Jersey, 1978.

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

[3] Doddington, G.R. Speaker Recognition - Identifying People by their Voices. *Proceedings of the IEEE, Vol. 73, No. 11*, pages 1651–1664, November 1985.

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

[5] Davis et al, K. H. Automatic Recognition of Spoken Digits. *Journal of the Acoustical Society of America, Vol. 24, No. 6*, pages 637–642, 1952.

[6] Koenig, W. The Sound Spectrograph. *Journal of the Acoustical Society of America, Vol. 17*, pages 19–49, 1946.

[7] Lee, K., Hauptmann, A. G. and Rudnicky, A. The Spoken Word. *Byte*, pages 225–232, July 1990.

[8] Campbell, J. P. Speaker Recognition: A Tutorial. *Proceedings of the IEEE, Vol. 85, No. 9*, pages 1437–1462, September 1997.

[9] Immendörfer, M. Applications for Speech Processing in Telecommunication and Office Equipment. *Electrical Communication, Vol. 60*, pages 71–78, 1986.

[10] Rabiner, L. R. Special Issue on Man-machine Communication by Voice. *Proceedings of the IEEE, Vol. 64*, pages 403–404, 1976.

[11] H. Gu, C. Tseng and L. Lee. Isolated-utterance Speech Recognition Using Hidden Markov Models with Bounded State Durations. *IEEE Transactions on Signal Processing*, pages 698–713, 1991.

[12] Vergin, R. and O'Shaugnessy, D. On the use of Some Divergence Measures in Speaker Recognition. *International Conference on Acoustics, and Signal Processing (ICASSP99)*, 1999.

[13] Atal, B. S. Automatic Recognition of Speakers from Their Voices. *Proceedings of the IEEE, Vol. 64, No. 4*, pages 460–475, April 1976.

[14] Demirekler, M. and Haydar, A. Feature Selection using Genetics-Based Algorithm and its Application to Speaker Identification. *International Conference on Acoustics, and Signal Processing (ICASSP99)*, 1999.

[15] Siohan, O., Lee, C., Surendran, A., and Li, Q. Background Model Design for Flexible and Portable Speaker Verification Systems. *International Conference on Acoustics, and Signal Processing (ICASSP99)*, 1999.

[16] O'Shaughnessy, D. Speaker Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing Magazine*, pages 4–17, October 1986.

[17] O'Shaughnessy. Speech Communication, Human and Machine. *Digital Signal Processing*, Reading, MA: Addison-Wesley, 1987.

[18] Rosenberg, A. E. Automatic Speaker Verification: A Review. *Proceedings of the IEEE, Vol. 64, No. 4*, pages 475–487, April 1976.

[19] Mammone, R. J., Zhang, X., and Ramachandran, R. P. Robust Speaker Recognition – A Feature-Based Approach. *IEEE Signal Processing Magazine, Vol. 13, No. 5*, pages 58–71, September 1996.

[20] Vassali, M. R., de Seixas, J. M. e Espain, C. Reconhecimento de Voz em Tempo Real Baseado na Tecnologia dos Processadores Digitais de Sinais. *XVIII Simpósio Brasileiro de Telecomunicações*, Setembro 2000.

[21] de Lima, A. A., Francisco, M. S., Netto, S. L. e F. Resende Jr., G. V. Análise Comparativa de Parâmetros em Sistemas de Reconhecimento de Voz. *XVIII Simpósio Brasileiro de Telecomunicações*, Setembro 2000.

[22] Reynolds, D. A. and Rose, R. C. Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1*, pages 72–83, January 1995.

[23] Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE, Vol. 77, No. 2*, pages 257–286, February 1989.

[24] Rabiner, L. R., Levinson, S. E., and Sondhi, M. M. On the Application of Vector Quantization and Hidden Markov Models to Speaker-independent, Isolated Word Recognition. *The Bell System Technical Journal, Vol. 62, No. 4*, pages 1075–1105, April 1983.

[25] Douglas, A. R. and Rose, R. C. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Tansactions on Speech and Audio Processing*, pages 72–82, 1995.

[26] Vassilis, D. D. and Vassilios, V. D. Maximum-Likelihood Stochastic-Transformation Adaptation of Hidden Markov Models. *IEEE Tansactions on Speech and Audio Processing*, pages 177–187, 1999.

[27] Deller Jr., J. R., Proakis, J. G., and Hansen, J. H. L. *Discrete-time Processing of Speech Signals*. Macmillan Publishing Co., 1993.

[28] Editor - Kosko, B. Neural Network for Signal Processing. *Prentice Hall International, Englewood Cliffs*, 1992.

[29] Oglesby, J. and Mason, J. S. Optimisation of Neural Models for Speaker Identification. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'90)*, pages 261–264, 1990.

[30] Gray, R. M. Vector Quantization. *IEEE ASSP Magazine*, pages 4–29, April 1984.

[31] He, J., Liu, L., and Palm, G. A Discriminative Training Algorithm for VQ-based Speaker Identification. *IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 3*, pages 353–356, May 1999.

[32] Gersho, A. and Gray, R. M. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, MA, 1992.

[33] Bennani, Y., Fogelman Soulie, F., and Gallinari, P. Text-Dependent Speaker Identification Using Learning Vector Quantization. *International Neural Network Conference*, 1990.

[34] Yuan, Z.-X., Xu, B.-L., and Yu, C.-Z. Binary Quantization of Feature Vectors for Robust Text-Independent Speaker Identification. *IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 1*, pages 70–78, July 1996.

[35] Furui, S. Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 29, No. 2*, pages 254–272, April 1981.

[36] Juang, B. H., Wong, D. Y., and Gray, Jr., A. H. Distortion Performance of Vector Quantization for LPC Voice Coding. *IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 30, No. 2*, pages 294–304, April 1982.

[37] Levinson, S. E., Rabiner, L. R., and Sondhi M. M. An Introduction to the Application of the Theory of Probabilist Functions of a Markov Process to Automatic Speech Recognition. *The Bell System Technical Journal, Vol. 62, No. 4*, pages 1035–1068, April 1983.

[38] Martins, J. A. and Violaro, F. Performance of Speaker-Independent Recognizers Based on Hidden Markov Models. *XVII Simpósio Brasileiro de Telecomunicações*, pages 554–557, 1999.

[39] Comeford, R., Makhoul, J., and Schwartz, R. The Voice of the Computer is Heard in the land (and it Listends too !). *IEEE Speech Recognition*, pages 39–47, 1997.

[40] Linde, Y., Buzo, A., and Gray, R. M. An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications, Vol. COM - 28, No. 1*, pages 84–95, January 1980.

[41] Madeiro, F., Fechine, J. M., and Aguiar Neto, B. G. Algoritmo Modificado de Kohonen Aplicado ao Projeto de Dicionários de Padrões Acústicos para Reconhecimento de Locutor. *Anais do V Simpósio Brasileiro de Redes Neurais (SBRN'98), Belo Horizonte, MG, Brasil*, pages 22–26, Dezembro, 1998.

[42] Madeiro, F., Vilar, R. M., and Aguiar Neto, B. G. Avaliação de Desempenho de um Algoritmo Modificado de Kohonen em Quantização Vetorial. *Anais do V*

Simpósio Brasileiro de Redes Neurais (SBRN'98), Belo Horizonte, MG, Brasil, pages 41–46, Dezembro 1998.

[43] Madeiro, F., Vajapeyam, M. S., Morais, M. R., Aguiar Neto, B. G., and Alencar, M. S. Multiresolution Codebook Design for Wavelet/VQ Image Coding. *Proceedings of the 15th International Conference on Pattern Recognition (ICPR'2000), Vol. 3, Barcelona, Spain*, pages 79–82, September 2000.

[44] Flanagan, L. J. *Speech Analysis Synthesis and Perception.* Murray Hill Second Edition, New Jersey, 1978.

[45] Russo, I e Behlau, M. *Percepção da Fala: Análise Acústica.* Editora Lovise, 1993.

[46] Fellbaum, K. *Sprachsignalverarbeitung and Sprachübertragung.* Springer-verlag, Berlin, 1984.

[47] Rabiner, L. R. and Juang, B. *Fundamentals on Speech Recognition.* 1996.

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

[49] Fechine, J. M. Verificação de Locutor Utilizando Modelos de Markov Escondidos (HMMs) de Densidades Discretas. Universidade Federal da Paraíba - Dissertação de Mestrado, Abril 1994.

[50] Papoulis, A. *Signal Analysis.* McGraw-Hill, 1985.

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

[52] Rabiner, L. R. Digital Formant Synthesizer for Speech Synthesis Studies. *J. Acoust. Soc. Am., Vol. 43, No. 4*, pages 822–828, April 1968.

[53] Winham, G. and Steiglitz, K. Input Generators for Digital Sound Synthesis. *J. Acoust. Soc. Am., Vol. 47, No. 2*, pages 665–666, February 1970.

[54] Chengalvarayan, R. Hierarchical Subband Linear Predictive Cepstral (HSLPC) Features for HMM-Based Speech Recognition. *International Conference on Acoustics, and Signal Processing (ICASSP99)*, 1999.

[55] Farrell, K. R., Mammone, R. J., and Assaleh, K. T. Speaker Recognition Using Neural Networks and Conventional Classifiers. *IEEE Transactions on Speech and Audio Processing (Special Issue on Neural Networks for Speech), Vol. 2*, pages 194–205, January 1994.

[56] Gopalan, K., Anderson, T. R., and Cupples, E. J. A Comparison of Speaker Identification Results Using Features Based on Cepstrum and Fourier-Bessel Expansion. *IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 3*, pages 289–294, May 1999.

[57] Tolba, H. and O'Shaughnessy, D. Voiced-Unvoiced Classification Using the First Mel Frequency Cepstral Coefficient. *International Conference on Speech Processing, Vol. 1*, pages 137–142, August 1997.

[58] Merwe, C. J. and Preez, J. A. Calculation of LPC-based Cepstrum Coefficients using Mel-Scale Frequency Warping. *IEEE Transactions on Acoustics, Speech and, Signal Processing*, July 1991.

[59] Ramachandran, R. P., Zilovic, M. S., and Mammone, R. J. A Comparative Study of Robust Linear Predictive Analysis Methods with Applications to Speaker Identification. *IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 2*, pages 117–125, March 1995.

[60] Costas, S. X. and Papanastasiou, C. Split Matrix Quantization of LPC Parameters. *IEEE Transactions on Speech and Audio Processing*, pages 113–125, 1999.

[61] Costa, W. C. de A. Reconhecimento de Fala Utilizando Modelos de Markov Escondidos (HMM's) de Densidades Contínuas. Universidade Federal da Paraíba - Dissertação de Mestrado, Junho 1994.

[62] Rass, M. J. et al. Average Magnitude Difference Function Pitch Extractor. *IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 22*, pages 353–362, October 1974.

[63] Atal, B. S. and Hanauer, S. L. Speech Analysis and Synthesis by Linear prediction of the Speech Wave. *J. Acoust. Soc. Am., Vol. 50, No. 2 (Part 2)*, pages 637–655, August 1971.

[64] Makhoul, J. Linear Prediction: A Tutorial Review. *Proceedings of the IEEE, Vol. 63, No. 4*, pages 561–580, April 1975.

[65] Aguiar Neto, B. G. *Signalaufbereitung in Digitalen Sprachübertragungssystemen.* Vom Fachbereich Elektrotechnik der Technischen Universität Berlin zur Verleihung des akademischen Grades Doktor-Ingenieur genehmigte Dissertation, 1987.

[66] Young, S., Jansen, J., Odell, J., Ollason, D., and Woodland, P. The HTK BOOK. *http://tcw2.ppsw.rug.nl/ tjeerd/spraak/HTKBook/HTKBook.html.*

[67] Violaro, F., Kaspar, B., and Martins, J. A. Isolated Word Recognition Using Hidden Markov Models. *VII Simpósio Brasileiro de Microondas e Optoeletrônica e XIV Simpósio Brasileiro de Telecomunicações, Vol. 2*, pages 533–538, Julho 1996.

[68] Liu, J. *Zur Untersuchung und Optimierung von Spracherkennungssystemen fr Isoliert Gesproche Wrter.* VDI VERLAG, Dsseldorf, 1989.

[69] Kanungo, T. *Hidden Markov Models.* Language and Media Processing Lab, Center for Automation Research, University of Maryland, http://www.cfar.umd.edu/ kanungo, 1998.

[70] Rabiner, L. R., Juang, B. H., Levinson, S. E., and Sondhi, M. M. Recognition of Isolated Digits Using Hidden Markov Models With Continuos Mixture Densities. *AT & T Technical Journal, Vol. 64, No. 6*, pages 1211–1234, July-August 1985.

[71] Makhoul, J., Roucos, S., and Gish, H. Vector Quantization in Speech Coding. *Proceedings of the IEEE, Vol. 73, No. 11*, pages 1551–1588, November 1985.

[72] He, J., Liu, L., and Palm, G. A Discriminative Training Algorithm for VQ-based Speaker Identification. *IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 3*, pages 353–356, May 1999.

[73] Madeiro, F. M. Quantização Vetorial Aplicada à Compressão de Sinais de Voz e Imagem. Universidade Federal da Paraíba - Dissertação de Mestrado, Março 1998.

[74] Gray, R. M. and Karnin, E. D. Multiple Local Optima in Vector Quantizers. *IEEE Transactions on Information Theory, Vol. IT-28, No. 2*, pages 256–261, March 1982.

[75] Madeiro, F. Avaliação de Desempenho de Algoritmos para Projeto de Quantizadores Vetoriais Aplicados à Compressão de Imagens. *Relatório Técnico, Universidade Federal da Paraíba, Departamento de Engenharia Elétrica*, 1998.

[76] Pan, J. S., McInnes, F. R., and Jack, M. A. VQ Codebook Design Using Genetic Algorithms. *Electronics Letters, Vol. 31, No. 17*, pages 1418–1419, 17th August 1995.

[77] Chen, C.-Q., Koh, S.-N., and Sivaprakasapillai. Codebook Generation for Vector Quantisation. *Electronics Letters, Vol. 31, No. 7*, pages 522–523, 30th March 1995.

[78] Chen, C.-Q., Koh, S.-N., and Sivaprakasapillai, P. VQ Codebook Design Algorithm Based on Partial GLA. *Electronics Letters, Vol. 31, No. 21*, pages 1803–1806, 12th October 1995.

[79] Lee, D., Baek, S., and Sung, K. Modified K-means Algorithm for Vector Quantizer Design. *IEEE Signal Processing Letters, Vol. 4, No. 1*, pages 2–4, January 1997.

[80] Yair, E., Zeger, K., and Gersho, A. Competitive Learning and Soft Competition for Vector Quantizer Design. *IEEE Transactions on Signal Processing, Vol. 40, No. 2*, pages 294–309, February 1992.

[81] Kohonen, T. The Self-Organizing Map. *Proceedings of the IEEE, Vol. 78, No. 9*, pages 1464–1480, September 1990.

[82] Haykin, S. *Neural Networks - A Comprehensive Foundation*. IEEE Press, Englewood Cliffs - NJ, 1994.

[83] Kohonen, T. *Self-Organization and Associative Memory (3rd ed)*. Springer-Verlag, Berlin, 1989.

[84] Freeman, J. A. and Skapura, D. M. Neural Networks - Algorithms, Applications and Programming Techniques. 1991.

[85] Hertz, J., Krogh, A., and Palmer, R. G. Introduction to the Theory of Neural Computation. 1992.

[86] Vilar França, R. M. Abordagem Neural da Quantização Vetorial de Sinais de Voz. Universidade Federal da Paraíba - Proposta de Tese de Doutorado, 1996.

[87] Beale, R. and Jackson, T. *Neural Computing: An Introduction.* Institute of Physics Publishing, Bristol and Philadelphia, 1990.

[88] Freeman, J. A. and Skapura, D. M. *Neural Networks - Algorithms, Applications and Programming Techniques.* Addison-Wesley, Reading, MA, 1991.

[89] Grossberg, S. Competitive Learnig: from Interactive Activation to Adaptative Resonance. *Cognitive Science*, pages 23–63, 1987.

[90] Kosko, B. – Editor. *Neural Networks for Signal Processing.* Prentice-Hall International, Englewood Cliffs, NJ, 1992.

[91] Madeiro, F., Vilar, R. M., and Aguiar Neto, B. G. A Self-Organizing Algorithm for Image Compression. *Proceedings of the Vth Brazilian Symposium on Neural Networks (IEEE SBRN'98), Belo Horizonte - MG, Brazil*, pages 146–150, December 1998.

[92] Madeiro, F., Vilar, R. M., Fechine, J. M., and Aguiar Neto, B. G. A Self-Organizing Algorithm for Vector Quantizer Design Applied to Signal Processing. *International Journal of Neural Systems, Vol. 9, No. 3, Special Issue on Neural Networks in Brazil: V Brazilian Symposium on Neural Networks*, pages 219–226, June 1999.

[93] Vilar França, R. M. and Aguiar Neto, B. G. Voice Waveform Vector Quantization Using a Competitive Algorithm. *Records of the IEEE GLOBECOM'94*, pages 872–875, November 1994.

[94] Madeiro, F., Vilar, R. M., Aguiar Neto, B. G., and de Assis, F. Designing Codebooks for Speech Compression through a Neural Network Algorithm. *Proceedings of the 2nd Conference on Telecommunications (Conftele'99), Sesimbra, Portugal*, pages 675–679, April 1999.

[95] Satish, L. and Gururaj, B. I. Use of Hidden Markov Models for Partial Discharge Pattern Classification. *IEEE Transactions on Electrical Insulation, Vol. 28, No. 2*, pages 172–182, April 1993.

[96] Auckenthaler, R., Parris, E. S., and Carey, M. J. Improving a GMM Speaker Verification System by Phonetic Weighting. *International Conference on Acoustics, and Signal Processing (ICASSP99)*, 1999.

[97] Rabiner, L. R., Juang, B. H., Levinson, S. E., and Sondhi, M. M. Some Properties of Continuos Hidden Markov Model Representations. *AT & T Technical Journal, Vol. 64, No. 6*, pages 1251–1270, August 1985.

[98] Definition of Hidden Markov Model. *http://www.jedlik.phy.bme.hu/ gerjanos/HMM.*

[99] Hidden Markov Models. *University of Leeds, http://www.scs.leeds.ac.uk/scs-only/tea...s/HiddenMarkovModels.*

[100] Rabiner, L. R. and Juang, B. H. An Introduction to Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 3, No. 1*, pages 4–16, February 1986.

[101] Rabiner, L. R. and Levinson, S. E. 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, Vol. 33, No. 3*, pages 561–573, June 1985.

[102] Savic, M. and Gupta, S. K. Variable Parameter Speaker Verification System Based on Hidden Markov Modeling. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'90)*, pages 281–284, 1990.

[103] Viterbi, A. J. Error Bounds for Convolutional Codes and an Asymptotically Optimal Decodind Algorithm. *IEEE Transactions Information Theory, Vol. 13*, pages 260–269, 1967.

[104] Forney, G. D. The Viterbi Algorithm. *Proceedings of the IEEE, Vol. 61*, pages 268–278, 1973.

[105] CREATIVE - Audio Products. Sound Blaster 16 Pro CSP (CT2290). *http://support.soundblaster.com/specs/audio/sb16/ct2290.html.*

[106] Fechine, J. M. Estimação dos Sons Básicos da Voz através da Análise de Parâmetros Temporais. Relatório Técnico, Universidade Federal da Paraíba, Departamento de Engenharia Elétrica, 1994.

[107] Sayed, A. H. e Alens, N. Simulador de Reconhecedores de Palavras Isoladas. *Revista da Sociedade Brasileira de Telecomunicações, Vol. 6, No. 1*, Dezembro 1991.

[108] Fechine, J. M. Caracterização de Sinais de Voz. Relatório Técnico, Universidade Federal da Paraíba, Departamento de Engenharia Elétrica, 1994.

[109] Williams, C. S. *Designing Digital Filters.* Prentice-Hall, New Jersey, 1986.

[110] Costa Neto, P. L. O. *Estatística.* Editora Edgard Blucher, Ltda, 1977.

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

[112] Suen, C. Y., Nadal, C., Legault, R., Mai, T. A., and Lam, L. Computer Recognition of Unconstrained Handwritten Numerals. *Proceedings of the IEEE, Vol. 80*, pages 1162–1180, 1992.

[113] Correia, S. E. N. *Reconhecimento de Caracteres Numéricos Manuscritos Usando a Transformada Wavelet.* Universidade Federal da Paraíba - Dissertação de Mestrado, 2000.

[114] Veloso, L. R. *Reconhecimento de Caracteres Numéricos Manuscritos.* Universidade Federal da Paraíba - Dissertação de Mestrado, 1998.

[115] Levine, D. M., Berenson, M. L. e Stephan, D. *Estatística: Teoria e Aplicações.* LTC - Livros Técnicos e Científicos Editora S. A., 2000.

[116] Hermansky, H., Morgan, N., Bayya, A., and Kohn, P. RASTA-PLP Speech Analysis. *Eurospeech*, pages 2–6, 1991.

[117] Mendell, J. M. Fuzzy Logic Systems for Engineering: A Tutorial. *Proceedings of the IEEE*, pages 345–375, March 1995.

[118] Burrus, C. S., Gopinath, R. A., and Guo, H. *Introduction to Wavelets and Wavelet Transforms.* Prentice-Hall, New Jersey, 1998.