UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Sistema de Reconhecimento de Locutor utilizando Redes Neurais Artificiais

por

ANDRÉ GUSTAVO ADAMI

Dissertação submetida à avaliação, como requisito parcial
para a obtenção do grau de Mestre em
Ciência da Computação

Prof. Dr. Dante Augusto Couto Barone
Orientador

Porto Alegre, maio de 1997.

CIP - CATALOGAÇÃO NA PUBLICAÇÃO

Adami, André Gustavo

Sistema de Reconhecimento de Locutor Utilizando Redes Neurais Artificiais / André Gustavo Adami.- Porto Alegre: CPGCC da UFRGS, 1997.
76 f.: il.

Dissertação (mestrado) - Universidade Federal do Rio Grande do Sul, Curso de Pós-Graduação em Ciência da Computação, Porto Alegre, 1997. Orientador: Barone, Dante Augusto Couto

1. Reconhecimento de Voz. 2. Processamento de Sinais Digitais. 3. Redes Neurais. I. Barone, Dante Augusto Couto. II. Título

Aos meus Pais,

por todo amor e compreensão

devotado sem medidas.

Agradecimentos

Gostaria de mencionar e agradecer a algumas pessoas, que contribuíram para a realização deste trabalho:

- Ao *Prof. Dante Barone*, por sua orientação, confiança, incentivo e contribuição efetiva para meu crescimento profissional.
- Ao *Prof. Altamiro Suzim*, pelos esclarecimentos dados sobre a área de Processamento de Sinais.
- Ao *Prof. Euvaldo Cabral Filho* da USP e *Prof. Fábio Violaro* da UNICAMP por terem elucidado algumas questões referentes à pesquisa.
- Ao amigo e *Prof. Ricardo Dorneles* pelo total apoio ao trabalho , incansável paciência e incentivo para ir sempre muito mais além.
- À *Prof. Adriana Miorelli* pela sua paciência nas correções realizadas deste trabalho, que ajudaram muito na confecção do mesmo.
- Ao *Prof. Heitor Strogulski* e ao amigo *Alex Pellin* pelo apoio e material fornecido para a conclusão deste trabalho.
- À minha família, pela paciência e apoio, principalmente nas horas de intenso trabalho.
- Aos amigos e colegas do curso, que de uma forma ou outra auxiliaram ou acompanharam o decorrer deste trabalho.
- Aos acadêmicos do curso de Computação da UCS que cederam material (voz) para a pesquisa.
- E finalmente, ao CNPq pelo financiamento da bolsa e ao CPGCC pela obtenção deste financiamento.

Sumário

Lista de Abreviaturas

A/D	Analógico/Digital
ANN	Artificial Neural Networks
CNPq	Conselho Nacional de Pesquisa
CPGCC	Curso de Pós-Graduação em Ciência da Computação
DFT	Discrete Fourier Tranform
DSP	Digital Signal Processing
DTW	Dynamic Time Warping
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
HMM	Hidden Markov Models
IC	Inteligência Computacional
IIR	Infinite Impulse Response
LMS	Least Mean-Square
LPC	Linear Predictive Coding
LVQ	Learning Vector Quantization
MLP	Multi-Layer Perceptron
RAF	Reconhecimento Automático da Fala
RL	Reconhecimento de Locutor
RNA	Rede Neural Artificial
ROC	Receptor Operations Curve
TDNN	Time-Delay Neural Network
UCS	Universidade de Caxias do Sul
UFRGS	Universidade Federal do Rio Grande do Sul
UNICAMP	Universidade de Campinas
USP	Universidade de São Paulo
VQ	Vector Quantization
ZCT	Zero Crossing Threshold

Lista de Figuras

Lista de Tabelas

Resumo

Este trabalho envolve o emprego de recentes tecnologias ligadas à promissora área de Inteligência Computacional e à tradicional área de Processamento de Sinais Digitais. Tem por objetivo o desenvolvimento de uma aplicação específica na área de Processamento de Voz: o reconhecimento de locutor.

Inúmeras aplicações, ligadas principalmente à segurança e controle, são possíveis a partir do domínio da tecnologia de reconhecimento de locutor, tanto no que diz respeito à identificação quanto à verificação de diferentes locutores.

O processo de reconhecimento de locutor pode ser dividido em duas grandes fases: extração das características básicas do sinal de voz e classificação.

Na fase de extração, procurou-se aplicar os mais recentes avanços na área de Processamento Digital de Sinais ao problema proposto. Neste contexto, foram utilizadas a freqüência fundamental e as freqüências formantes como parâmetros que identificam o locutor. O primeiro foi obtido através do uso da autocorrelação e o segundo foi obtido através da transformada de Fourier.

Estes parâmetros foram extraídos na porção da fala onde o trato vocal apresenta uma coarticulação entre dois sons vocálicos. Esta abordagem visa extrair as características desta mudança do aparato vocal.

Existem dois tipos de reconhecimento de locutor: identificação (busca-se reconhecer o locutor em uma população) e verificação (busca-se verificar se a identidade alegada é verdadeira).

O processo de reconhecimento de locutor é dividido em duas grandes fases: extração das características (envolve aquisição, pré-processamento e extração dos parâmetros característicos do sinal) e classificação (envolve a classificação do sinal amostrado na identificação/verificação do locutor ou não).

São apresentadas diversas técnicas para representação do sinal, como análise espectral, medidas de energia, autocorrelação, LPC (*Linear Predictive Coding*), entre outras. Também são abordadas técnicas para extração de características do sinal, como a freqüência fundamental e as freqüências formantes.

Na fase de classificação, pode-se utilizar diversos métodos convencionais: Cadeias de Markov, Distância Euclidiana, entre outros. Além destes, existem as Redes Neurais Artificiais (RNAs) que são consideradas poderosos classificadores. As RNAs já vêm sendo utilizadas em problemas que envolvem classificações de sinais de voz. Neste trabalho serão estudados os modelos mais utilizados para o problema de reconhecimento de locutor.

Assim, o tema principal da Dissertação de Mestrado deste autor é a implementação de um sistema de reconhecimento de locutor utilizando Redes Neurais Artificiais para classificação do locutor.

Neste trabalho também é apresentada uma abordagem para a implementação de um sistema de reconhecimento de locutor utilizando as técnicas convencionais para o processo de classificação do locutor. As técnicas utilizadas são Dynamic Time Warping (DTW) e Vector Quantization (VQ).

PALAVRAS-CHAVE: Reconhecimento de Voz, Processamento de Sinais Digitais, Reconhecimento de Locutor, Redes Neurais Artificiais, Inteligência Computacional.

TITLE: "ARTIFICIAL NEURAL NETWORKS SPEAKER RECOGNITION SYSTEM"

Abstract

This work deals with the application of recent technologies related to the promising research domain of Intelligent Computing (IC) and to the traditional Digital Signal Processing area. This work aims to apply both technologies in a Voice Processing specific application which is the **speaker recognition** task.

Many security control applications can be supported by speaker recognition technology, both in identification and verification of different speakers.

The speaker recognition process can be divided into two main phases: basic characteristics extraction from the voice signal and classification. In the extraction phase, one proposed goal was the application of recent advances in DSP theory to the problem approached in this work. In this context, the fundamental frequency and the formant frequencies were employed as parameters to identify the speaker. The first one was obtained through the use of autocorrelation and the second ones were obtained through Fourier transform.

These parameters were extracted from the portion of speech where the vocal tract presents a coarticulation between two voiced sounds. This approach is used to extract the characteristics of this apparatus vocal changing.

In this work, the Multi-Layer Perceptron (MLP) ANN architecture was investigated in conjunction with the backpropagation learning algorithm. In this sense, some main characteristics extracted from the signal (voice) were used as input parameters to the ANN used. The output of MLP, trained previously with the speakers features, returns the authenticity of that signal.

Tests were performed with 10 different male speakers, whose age were in the range from 18 to 24 years. The results are very promising.

In this work it is also presented an approach to implement a speaker recognition system by applying conventional methods to the speaker classification process. The methods used are Dynamic Time Warping (DTW) and Vector Quantization (VQ).

KEYWORDS: Voice Recognition, Digital Signal Processing, Speaker Recognition, Artificial Neural Networks, Intelligent Computing.

1 Introdução

A tecnologia do processamento da fala vem se desenvolvendo nos últimos anos devido a fatores como o avanço da área de processamento de sinais digitais (DSP - *Digital Signal Processing*), tanto no desenvolvimento de técnicas quanto na criação de *hardware* específico, e a crescente capacidade de processamento e armazenamento dos computadores digitais.

Esta tecnologia está possibilitando o desenvolvimento de numerosas aplicações. Tais aplicações podem ser compreendidas em três classificações de sistemas de comunicação homem-máquina pela voz:

1. Sistemas de Resposta pela Voz;
2. Sistemas de Processamento da Fala.
3. Sistemas de Reconhecimento;

Os sistemas de resposta pela voz são projetados para responder, através do uso da fala, a uma requisição do usuário ou do sistema. A comunicação destes sistemas é uni-direcional, isto é, da máquina com o homem. Nestes tipos de sistemas, utilizam-se as técnicas de síntese de voz, as quais têm por objetivo gerar sinais de voz com qualidades humanas para comunicação.

Os sistemas de processamento de fala têm por objetivo realizar operações sobre o sinal de voz a fim de fornecer resultados que não envolvam o reconhecimento do conteúdo da mensagem, do locutor ou da língua. Nestes tipos de sistemas enquadram-se aplicações como:

- **Melhoramento da Fala**: técnicas para a separação do locutor de um ruído de fundo. O objetivo é melhorar ou aumentar a compreensão do sinal de voz que, dependendo do ambiente e meio (tipo de microfone) de aquisição do sinal, pode conter estes ruídos **[ASC 86] [EPH 90].**

- **Separação da Fala**: técnicas que tentam separar o sinal de voz de cada locutor[1] de uma locução na qual múltiplos locutores estão presentes. Há três décadas que se pesquisa a solução do problema de separação de locutores **[DOR 93] [ZIS 89]**. Assemelha-se muito ao *Melhoramento da Fala*, com a diferença que o objetivo é separar um locutor do outro. Segundo **[MOR 91],** a separação de locutores é muito difícil.

- **Codificação da Fala**: técnicas que tentam extrair as características mais importantes do sinal para a transmissão do mesmo de forma mais rápida e segura. É usada em uma variedade de aplicações como: transmissão de voz sobre canais de comunicação com largura de banda

[1] Diz-se da pessoa que fala.

limitada **[ATA 84] [GRI 87]**, compressão de voz para armazenamento digital e segurança em comunicações.

Os sistemas de reconhecimento podem ser subdivididos em sub-classificações que têm por objetivo o reconhecimento de uma informação específica da fala. Tais sub-classificações podem ser enumeradas como:

1. **Reconhecimento de Locutor (RL)**: tem por objetivo reconhecer o indivíduo que está falando. Este tipo de aplicação pode ser classificada em verificação e identificação de locutor. Na verificação de locutor, a aplicação deve decidir se um locutor é a pessoa que o mesmo diz ser. Diferente disto, a aplicação de identificação de locutor deve decidir qual locutor, entre um conjunto de locutores, realizou uma declaração.

2. **Identificação da Língua**: técnicas para diferenciação de uma língua. É um problema muito difícil pois a complexidade de encontrar características que tornem uma língua única é bastante elevada. Este grau de dificuldade é acentuado ainda pelos diferentes locutores e sotaques. Estudos anteriores do problema de identificação da língua têm se concentrado nos métodos de identificar fonemas específicos da linguagem, segmentos de som, entre outros.

3. **Reconhecimento Automático de Fala (RAF)**: são as aplicações que têm por função o reconhecimento do conteúdo lingüístico da fala, ou seja, a informação da fala. Existem três tipos de aplicações para RAF: reconhecimento de palavras isoladas (palavras pronunciadas isoladamente com pausas bem delineadas) **[LUF 94]**, reconhecimento de fala conectada (cada palavra é claramente articulada e não há pausas entre as palavras) e reconhecimento de fala contínua (não há pausas entre as palavras e a articulação das mesmas nem sempre são feitas claramente).

Dentre as aplicações definidas nos sistemas de reconhecimento, o presente trabalho abordará a de reconhecimento de locutor. Esta aplicação permite a sistemas de computação verificar a identidade do locutor, operações bancárias pelo telefone, compras pelo telefone, serviços de acessos a banco de dados, serviços de informação, correio de voz, controle de segurança para áreas de informação confidencial ou residenciais (prédios com elevadores de segurança), acesso remoto de computadores.

A aplicação de RL envolve duas grandes etapas:

1. Processamento do sinal e extração das características: tem por objetivos obter uma representação discreta dos sinais de voz, identificar os componentes-chave desta representação e eliminar informação redundante;

2. Classificação de padrões: após a fala ter sido transformada em algum espaço de características, alguma técnica de classificação é aplicada para identificar ou verificar a identidade do locutor.

Em paralelo ao desenvolvimento da tecnologia de processamento da fala, foram se desenvolvendo pesquisas na área de Redes Neurais Artificiais, que são

modelos matemáticos com a finalidade de simular o comportamento das redes neurais biológicas.

As redes neurais normalmente estão associadas a reconhecimento de padrões como voz ou caracteres **[DOR 93]**, devido à sua capacidade de realizar mapeamentos complexos entre padrões discretos. Por isso, serão apresentados alguns modelos utilizados para o reconhecimento de locutor a fim de utilizar um dos mesmos neste trabalho.

Este trabalho tem por objetivo realizar a implementação de um sistema de identificação de locutor utilizando como técnica de classificação um modelo de RNAs.

1.1 O Processo de Produção da Fala

A fala é um complexo som produzido pelo aparato vocal humano, que consiste de órgãos primariamente usados para respiração e alimentação: os pulmões, traquéia, laringe, garganta, boca, e nariz, como mostra a figura 1.1.

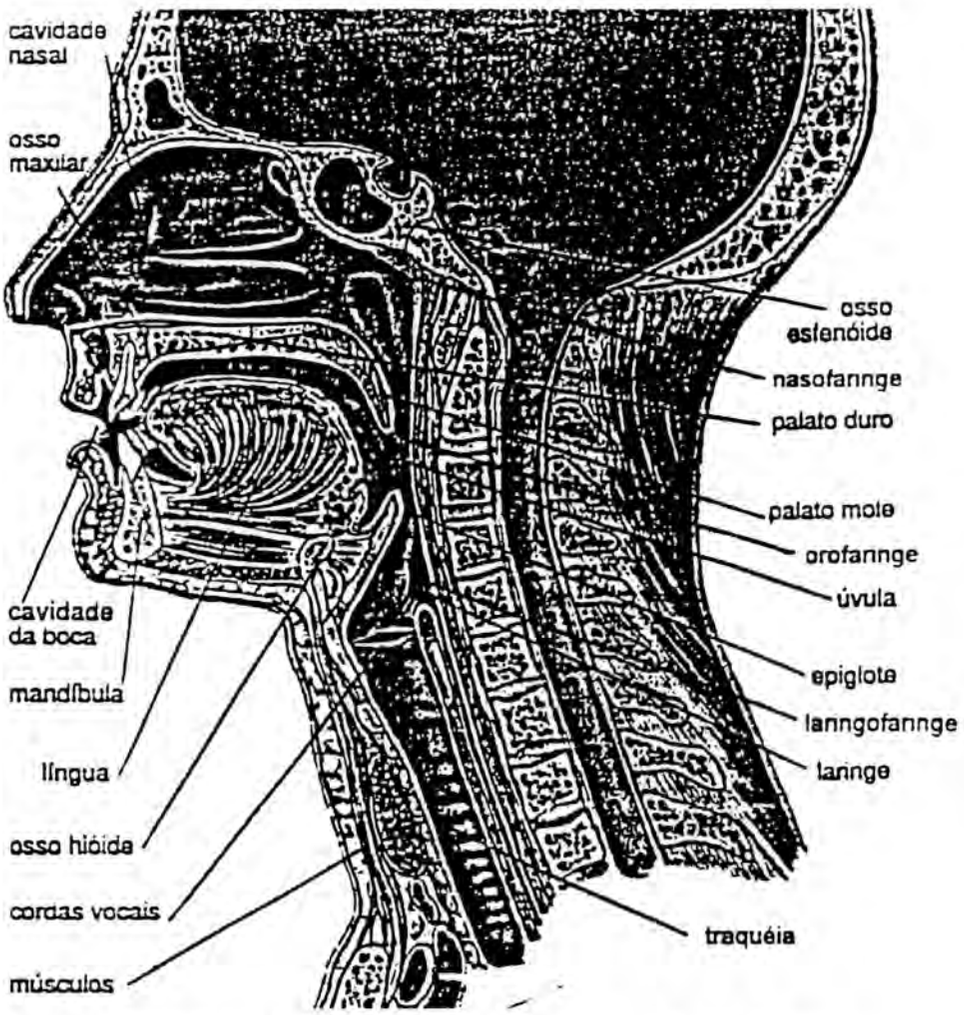

FIGURA 1.1 - Mecanismo humano da produção da fala.

Cada pessoa tem características vocais, as quais são suficientemente únicas tal que pode-se reconhecer a pessoa somente pela sua voz. Estas características estão diretamente relacionadas à fisiologia do locutor **[MOR 91]**.

O som é gerado de duas maneiras. Se o ar é forçado através da laringe com as pregas vocais apropriadamente posicionadas e tensionadas, é definida uma oscilação, tal que as pregas liberam porções de ar em uma forma quase-periódica de taxas em torno de 80 a 200 Hz para locutores masculinos, e mais rápido para mulheres e crianças. Esta forma de onda quase-periódica é conhecida como freqüência fundamental ou *pitch*. Esta fonte glotal é rica em harmônicas, e excita as ressonâncias acústicas do trato vocal acima da laringe, que filtra o som.

As freqüências ressonantes da função de transferência do trato vocal, chamadas de formantes, são determinadas pela forma da mandíbula, boca, e se o *velum* está aberto, a cavidade nasal. É ao longo da manipulação da forma do trato vocal pelos articuladores (língua, maxilar, lábios e velum) que são controladas as freqüências formantes que diferenciam os vários sons vocálicos da fala (vogais e consoantes nasais). Para a percepção da fala, são exigidas normalmente não mais que cinco formantes nas freqüências de 100 Hz a 5kHz. A característica intrínseca que diferencia as vogais entre si depende principalmente das três primeiras formantes. As duas últimas não contribuem essencialmente à compreensão dos sons da fala, mas determinam muito a naturalidade da fala e reconhecibilidade do locutor.

A outra fonte de som usada na fala é um ruído turbulento, produzido pelo ar forçado através de algumas constrições (tais como entre a língua e os dentes em /s/ - sons fricativos) ou por uma liberação abrupta de pressão formada em algum ponto no trato vocal (tais como atrás dos lábios em um /p/ ou /t/ - sons plosivos). Os picos espectrais associados com estes sons fricativos geralmente permanecem entre 2 e 8 kHz e são primariamente determinados pela posição e a forma da constrição. Alguns sons, denominados fricativos sonoros, tais como /z/ e /v/, têm excitação turbulenta e vocálica.

Diferenças distinguíveis no sinal de voz podem ser produzidas por uma pequena mudança no modo que o trato vocal é manipulado, tal que um grande número de sons pode ser produzido. Para a comunicação de linguagem, entretanto, somente um restrito número de sons (fonemas), ou mais precisamente, classes de sons são usados.

1.2 Estrutura do Trabalho

Esta monografia está dividida em nove capítulos, dos quais esta introdução é o primeiro e, no último são apresentadas as conclusões obtidas neste trabalho, decorrentes dos resultados das implementações realizadas.

No capítulo 2 são apresentadas algumas definições e técnicas sobre as aplicações que envolvem o reconhecimento da fala.

O capítulo 3 apresenta as principais técnicas de processamento de sinais para a representação dos mesmos em uma forma conveniente para a extração das características.

O capítulo 4 apresenta as técnicas e algoritmos para a extração das características do sinal.

O capítulo 5 aborda algumas técnicas convencionais para classificação de padrões e baseadas em Redes Neurais Artificiais.

No capítulo 6 é abordada a aplicação de reconhecimento de locutor, principais métodos, técnicas e estado da arte.

O capítulo 7 e 8 mostram os passos da implementação do sistema de reconhecimento de locutor baseado em RNAs e métodos convencionais DTW (Dynamic Time Warping) e VQ (Vector Quantization), respectivamente. Neste capítulo, também são apresentados e discutidos os testes realizados e os resultados obtidos.

2 Reconhecimento de Voz

Um modelo muito utilizado na tecnologia de reconhecimento de voz é apresentado na figura 2.1. Algumas variações podem ocorrer de pesquisador para pesquisador, mas o princípio do modelo é o mesmo. Este modelo mostra uma estruturação para as soluções do problema de reconhecimento de voz. Tal estruturação é muito utilizada em outros problemas, como reconhecimento de locutor **[SOR 95] [PRA 95] [ADA 96]** e separação de locutores **[DOR 93]**, entre outros.

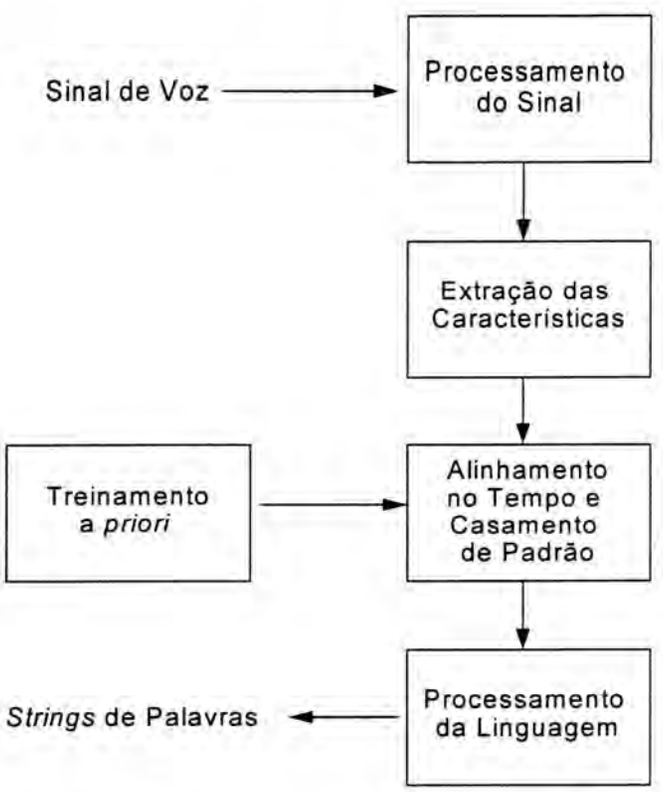

FIGURA 2.1 - Diagrama de blocos do Modelo Geral de Reconhecimento de Voz.

Os componentes descritos no diagrama, conforme mostra a figura 2.1, são:

1. **Processamento do Sinal:** a sua função é obter uma representação do sinal de voz, isto é, converter o sinal analógico em sinal digital. Nesta conversão de sinal, estão embutidos alguns processamentos para a melhoria da qualidade do sinal, devido à degradação do mesmo por alguns fatores como: variação da amplitude, ruído no canal, *stress* (estado emocional ou físico alterado), entre outros.

2. **Extração das Características:** tem por objetivo identificar os componentes-chave da representação e eliminar informações redundantes. Tais componentes podem ser entendidos como informações únicas para o próximo módulo poder realizar um casamento destas informações. Deve ser independente das partes da fala e hábil na captura das transições, pois nelas podem haver informações fonéticas do sinal.

3. **Alinhamento Temporal e Casamento de Padrão:** sua função é casar uma palavra falada com uma representação dada (amostra) daquela palavra. O casamento de padrão é utilizado para comparar as representações de palavras. O alinhamento temporal refere-se ao alinhamento dos eventos acústicos ou fonéticos que têm sido modelados, devido à taxa de fala. Esta variação temporal ocorre naturalmente pois afeta a duração das palavras, o que dificulta o casamento dos padrões. Existem vários algoritmos que são utilizados neste modelo, os quais usualmente são dependentes do paradigma de treinamento. Estes algoritmos têm que estar preparados para problemas de pronunciação e dialetos.

4. **Processamento da Linguagem:** responsável pela seleção da palavra final, isto é, a montagem da sentença falada. A entrada deste módulo é constituída por palavras, e não pelo sinal de voz, onde será feita a seleção das palavras para o contexto. Este módulo é muito utilizado em reconhecedores de grandes vocabulários, onde a similaridade de palavras é grande.

Este modelo não ressalta tarefas específicas para serem executadas e nem mesmo para a representação do sinal de voz. Uma razão para isso é a constante evolução da área de processamento da fala, que sofre determinadas trocas na abordagem do problema de RAF e novos algoritmos para as fases do reconhecimento. Os algoritmos que são utilizados podem requerer novas representações do sinal e assim mudar o modelo geral. Segundo **[MOR 91]**, existe pouco consenso em uma representação específica do sinal a ser utilizado no processo de reconhecimento.

2.1 Processamento de Sinal e Extração das Características

O sinal produzido na fala é analógico e contínuo, por isso o mesmo deve ser amostrado de uma forma digital. O processamento digital de sinais preocupa-se com a obtenção de uma representação discreta dos sinais de voz com a mínima influência do ruído de fundo ou do canal, características do canal, variações da amplitude e *stress* do locutor.

O sinal de voz digitalizado pode ser então convertido para parâmetros digitais, que representam as suas características acústicas, a fim de que se possa analisar para uma determinada aplicação.

Uma abordagem comumente **[RAB 78] [EMB 91] [MOR 91]** utilizada é a união dos módulos de *Processamento de Sinal* e *Extração de Características* como somente uma única fase, pois ambos os módulos utilizam técnicas de processamento de sinais digitais. Nesta seção serão vistas as etapas de processamento de sinal e extração de características, bem como as principais técnicas de processamento de sinais utilizadas.

2.1.1 Amostragem

Os sinais que representam a fala podem ser definidos como uma seqüência de valores contínuos (sinal analógico) definidos no tempo. Para a manipulação desta seqüência através de computadores, é preciso representá-los em uma forma digital. Esta conversão torna-se necessária pelo fato de que os sinais analógicos têm uma precisão ilimitada (diferente dos computadores digitais).

A conversão analógico/digital (A/D) é um processo de amostragem periódica do sinal analógico e quantização de cada amostra. Após este processo o sinal digital poderá ser então processado por um filtro digital (como é mostrado na figura 2.2).

FIGURA 2.2 - Processamento de sinais analógicos por filtro digital.

O sinal analógico, representado por $x_a(t)$, é amostrado a cada T segundos. Na saída do amostrador é obtido um sinal digital $x(n)$ segundo a equação 2.1.

$$x(n) = x_a(nT) \qquad -\infty < n < \infty \qquad (2.A)$$

A figura 3.3 mostra uma forma de onda da locução "quinto andar" e o conjunto de amostras com o período T = 1/22050 segundos.

FIGURA 2.3 - Representação de um sinal de voz

Na amostragem é imposta uma condição: a utilização de sinais com uma largura de banda limitada. Este processo fica caracterizado pelo Teorema da Amostragem **[RAB 78] [EMB 91]**, o qual é enunciado da seguinte forma:

> "Qualquer função do tempo ***f(t)***, cuja freqüência mais elevada seja ***W*** **Hz**, pode ser completamente determinada por amostragem da sua amplitude em intervalos de tempo espaçados de 1/(2W)".

Esta freqüência mínima de amostragem é chamada de *Nyquist*. Se a amostragem não for desta maneira, o que pode ocorrer é um efeito chamado *aliasing* (as amostras se sobrepõem impedindo uma boa amostragem do sinal).Na figura 2.3, o sinal foi amostrado a 22kHz, o qual contém, pela teoria de *Nyquist*, freqüências de até 11kHz.

O filtro utilizado é um passa-baixa, que deverá ser inicializado abaixo da freqüência de *Nyquist* para que a transformada de *Fourier* do sinal, com uma largura de banda limitada, não contenha qualquer *aliasing* do sinal na banda base de 2πn/T intervalos **[EMB 91]**.

2.1.2 Filtros Digitais

Segundo **[EMB 91]**, filtros digitais são operadores lineares que, aplicados a seqüências (sinais digitais) permitem a certas freqüências da entrada, no domínio freqüência, passar sem alterações para a saída, enquanto bloqueiam as freqüências não relevantes. Estes operadores lineares podem ser representados pela equação 2.2.

$$y(n) = \sum_{q=0}^{M} b_q x(n-q) - \sum_{p=1}^{N} a_p y(n-p) \qquad (2.B)$$

onde *x[n]* é o estímulo para o filtro, *y[n]* é o resultado ou saída do filtro e os coeficientes b_q e a_p são os coeficientes do sinal de entrada e saída do filtro, respectivamente.

Existem duas grandes classes de filtros digitais. A primeira classe de filtros digitais tem a_p igual a zero para todo p na equação 2.2. O nome utilizado para este tipo é filtro de resposta impulso finita (*Finite Impulse Response - FIR*), pois sua resposta a um impulso extingue-se em um número finito de amostras. Segundo **[RAB 78]**, devido às propriedades destes tipos de filtros, estas tornam-se úteis para aplicações em processamento de fala, onde o preciso alinhamento temporal é essencial. Os filtros FIR são representados pela equação 2.3.

$$y(n) = \sum_{q=0}^{M} b_q x(n-q) \qquad (2.3)$$

Existem essencialmente três bem conhecidas classes de técnicas utilizadas para um filtro FIR: janelas, amostragem de freqüência e mínimo erro **[RAB 78] [EMB 91]**. Somente a primeira destas três técnicas é analítica, isto é, um conjunto fechado de equações podem ser solucionadas para obter os coeficientes do filtro. A segunda e a terceira técnica são métodos de otimização, os quais usam abordagens iterativas para obter o filtro desejado.

Os filtros do tipo janela são utilizados para reduzir a amplitude das bordas de um segmento do sinal. Estes filtros simplesmente reduzem a amplitude destas bordas. Isto é feito de uma maneira gradual e suave tal que nenhuma nova descontinuidade

será produzida e o resultado é uma substancial redução. Na tabela 2.1 são mostrados alguns exemplos de janelas, onde *M* é o tamanho da seqüência a ser analisada.

TABELA 2.A- Exemplos de algumas janelas e suas equações.

NOME	**FUNÇÃO**	
Hamming	0.54-0.46*cos(2*PI*n/M)	0≤n≤M
Hanning	0.5-0.5*cos(2*PI*n/M)	0≤n≤M
Barlett (triângulo)	2*n/M	0≤n≤(M-1)/2
	2-2*n/M	(M-1)/2≤n≤M
Blackman (3 termos)	0.42-0.5*cos(2*PI*n/M)+0.08*cos(4*PI*n/M)	0≤n≤M
Blackman-Harris (4 termos)	0.35875-0.48829*cos(2*PI*n/M)+0.14128*	
	cos(4*PI*n/M)-0.01168*cos(6*PI*n/M)	0≤n≤M

Segundo **[RAB 78]**, a janela é o método de menor complexidade, porém, o terceiro método é também largamente utilizado **[OPP 75]**.

A figura 2.4 mostra a representação em diagrama de blocos de um filtro FIR. Tal diagrama mostra graficamente as operações necessárias para calcular cada valor da seqüência de saída dos valores da seqüência de entrada. Os elementos básicos do diagrama mostram as adições, multiplicações da seqüência de valores pelas constantes e armazenamento dos valores passados da seqüência de entrada.

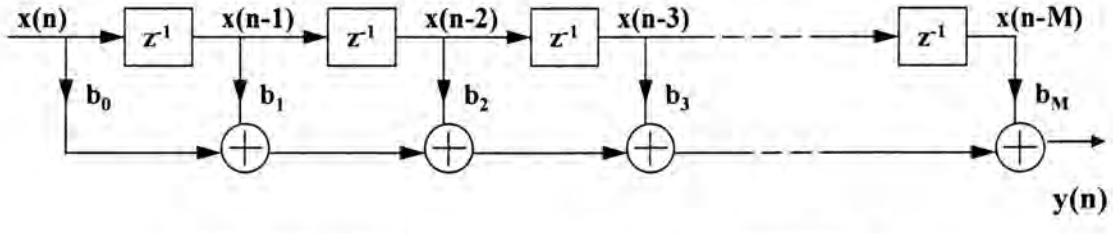

FIGURA 2.4 - Diagrama de blocos de um sistema FIR.

A segunda classe de filtros é caracterizada pela equação 2.2, onde a seqüência de saída é utilizada no cálculo dos próximos valores. Isto é, este sistema é recursivo e por isso é chamado de resposta impulso infinita (*Infinite Impulse Response - IIR*).

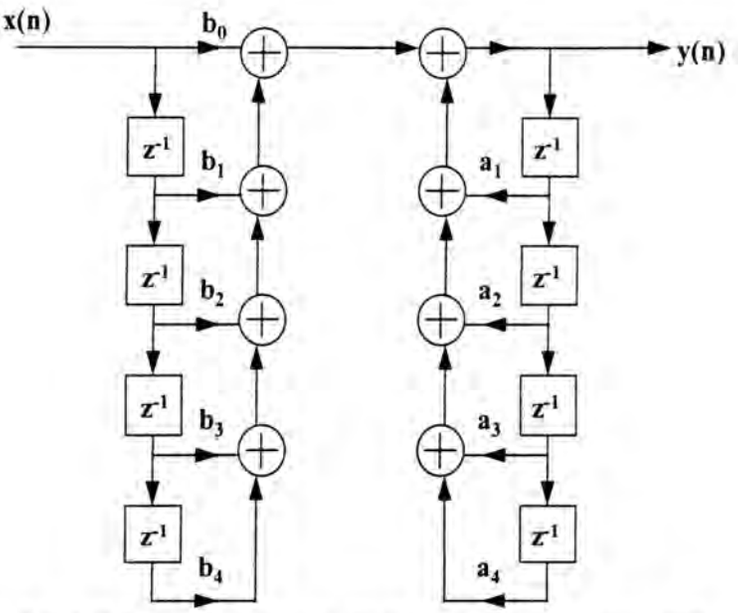

FIGURA 2.5 - Diagrama de blocos de um sistema IIR.

Estes tipos de filtros podem ser representados em diagramas de bloco como mostra a figura 2.5. Neste sistema, **M** e **N** são inicializados com o valor 4.

2.1.3 Representação digital do sinal de voz

O princípio dos métodos de representação digital de sinais de voz é amostrar o sinal de uma maneira tão precisa quanto possível, de modo que um sinal poderia ser reconstruído a partir de sua representação. No caso de reconhecimento de locutor, não existe interesse em reconstruir o sinal de voz; a única preocupação é representar o sinal de modo que as características do locutor permaneçam inalteradas na forma digital para que se possa posteriormente identificá-las.

O sinal de voz é não estacionário, porém a propriedade básica na qual são baseados os métodos de representação da voz usados nos sistemas de reconhecimento, é que as propriedades da forma de onda podem ser consideradas invariantes durante um período de tempo na ordem de 10 a 30ms **[SOR 95]**. A voz é geralmente representada como uma seqüência de parâmetros obtidos da análise de segmentos curtos do sinal de voz espaçados uniformemente no tempo. Na prática, a periodicidade com a qual é feita esta análise do sinal varia de 5ms a 20ms.

A representação destas seqüências pode ser feita através de forma de onda ou paramétrica **[WHE 92] [RAB 78]**. O modelo utilizado para a produção digital da fala é mostrado na figura 2.6. Neste modelo, a saída do filtro variável no tempo é representado pelas amostras do sinal de voz. Este filtro variável aproxima a função de transferência do trato vocal.

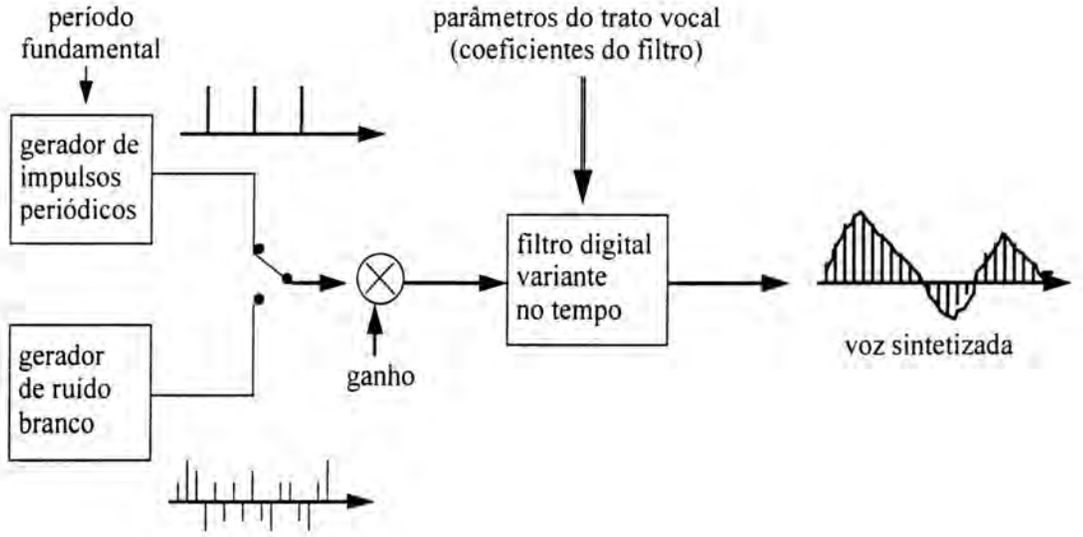

FIGURA 2.6 - Modelo Digital de geração de voz [WEI 92]

Segundo **[WEI 92]**, o trato vocal varia lentamente durante a fala contínua; portanto, pode-se assumir que o filtro digital da figura 2.4 tenha características fixas durante um intervalo de tempo da ordem de 10ms a 30ms.

Há muitas possibilidades para representações discretas de sinais de voz. Como mostra a figura 2.7, estas representações podem ser classificadas em dois grandes grupos chamados representações em forma de onda e representações paramétricas **[RAB 78] [ADA 96]**.

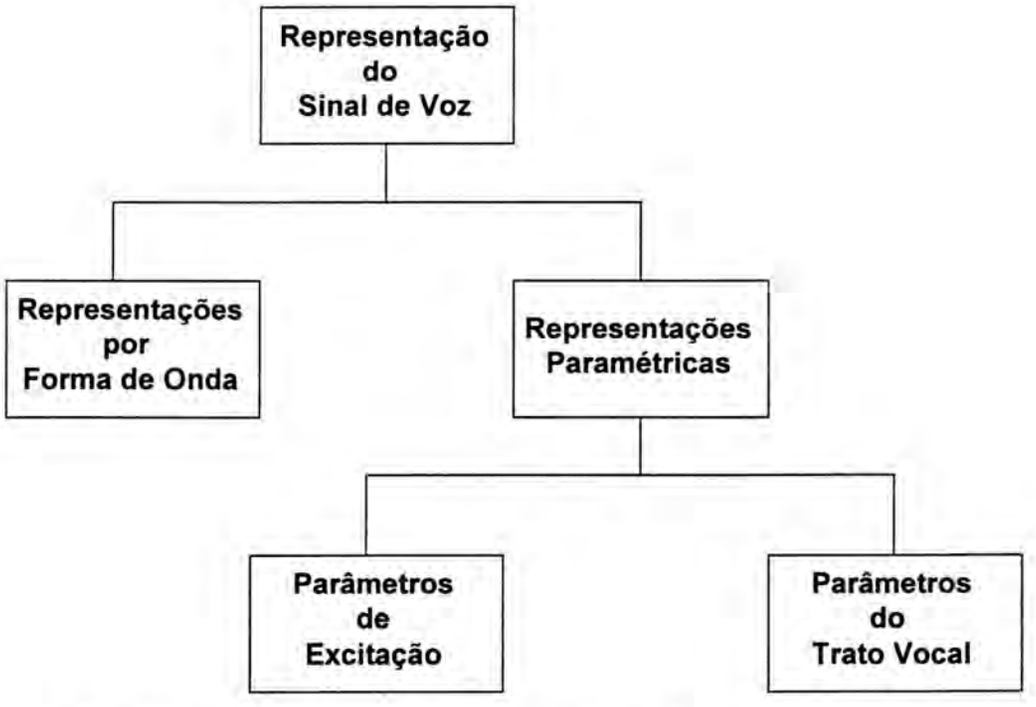

FIGURA 2.7 - Classificação das representações do sinal de voz [RAB 78]

Na codificação da forma de onda, o sinal é representado diretamente pela seqüência de valores amostrados, enquanto que, na codificação paramétrica, a representação se dá em termos dos parâmetros variáveis no tempo do modelo básico de geração de voz.

Segundo **[WEI 92]**, a escolha da forma de codificação é dirigida por um conjunto de fatores:

- qualidade do sinal amostrado;
- complexidade computacional;
- taxa de armazenamento (ou transmissão);
- robustez da estrutura de codificação e
- custo de implementação.

A complexidade computacional do processo de codificação vai definir a quantidade de processamento necessária para se codificar e recuperar o sinal. Isso influenciará na decisão de se executar em tempo real ou não.

A robustez da representação é uma medida de insensibilidade do codificador aos ruídos ambientais normalmente presentes no sinal de voz a ser codificado.

A qualidade do sinal de voz a ser codificado é o fator fundamental que irá reger as relações de compromisso entre os outros fatores de importância. Em geral, um aumento na qualidade implica em aumento da taxa de armazenagem e/ou complexidade computacional.

2.1.4 Extração de características

O processo de extração das características consiste em obter parâmetros característicos que possam ser utilizados para classificar o sinal. É o ponto chave do problema de reconhecimento de padrões. No caso de reconhecimento do locutor, as características extraídas do sinal de voz devem ser invariantes para um mesmo locutor, mas de grande diferença para um impostor.

A escolha das características únicas do locutor de um sinal de voz incorre no desempenho do sistema e qualidade de reconhecimento. Por isso, é complexo obter um sistema com um bom desempenho e uma boa qualidade. Os sistemas de reconhecimento de locutor têm usado diversos tipos de características, como por exemplo:

- medida de energia da saída de um banco de filtros **[BRI 71][PRU 63]**;
- *pitch* (freqüência fundamental) **[ATA 68]**;
- formantes **[DOD 71]**;

- intensidade **[LUM 73]**;
- *log-area ratio* **[FUR 73]**;
- coeficientes da predição linear **[BEN 91] [LEE 95]**;
- coeficientes cepstrais **[BEN 91] [FUR 81] [MAG 95] [SOR 95]**.

No capítulo três serão abordados os algoritmos para a extração das formantes e da freqüência fundamental (*pitch*) como características identificadoras do locutor.

2.2 Classificação das Características

Esta etapa do reconhecimento do locutor é responsável pela classificação do locutor a ser reconhecido entre os vários a serem verificados ou identificados. A complexidade desta etapa é relacionada ao tamanho do vocabulário, à taxa de fala, e à variabilidade do locutor (*stress* emocional e/ou físico). Os métodos convencionais comumente utilizados podem ser enumerados como:

- Distância Euclidiana **[ADA 94] [MOR 90]**: calcula-se a distância das características da fala das amostras de referência (previamente conhecidas) e de teste (a ser reconhecida).
- *Vector Quantization* (VQ) **[MAK 85]**: os padrões de características são agrupados em classes através de um algoritmo iterativo. Após isso, é verificado a que classe o padrão de referência (a ser reconhecido) pertence.
- Cadeias de Markov (HMM - *Hidden Markov Models*) **[HWA 93] [TIS 91]**: é um método para modelar sistemas com comportamentos discretos e dependentes do tempo, caracterizados por "processos", de curto-espaço e comuns, e as transições entre os mesmos. Este método pode ser comparado a uma máquina de estados finita.

Os classificadores convencionais mencionados nesta seção não são os únicos classificadores. Há outros métodos conhecidos, tais como o modelo misturado Gaussiano **[REY 75]**, distribuições discretas, vizinho mais próximo **[HIG 93]**, entre outros.

A manipulação de uma grande massa de dados sempre foi um problema para os algoritmos convencionais. A partir disto, pensou-se na utilização de Redes Neurais devido às suas características como processamento paralelo, elementos de processamento simples (soma e multiplicação), e ótimos classificadores **[MOR 91]**. Estes classificadores serão abordados mais profundamente no capítulo 5.

3 Representação do Sinal de Voz

O objetivo no processamento do sinal de voz é obter uma representação mais conveniente ou mais útil da informação contida no sinal.

A precisão necessária desta representação é definida pelas informações específicas no sinal de voz, que devem ser preservadas, ou em algum caso, mais proeminentes. Por exemplo, se a aplicação é definir se em um dado momento o sinal é vocálico, não-vocálico ou silêncio, serão aplicadas somente técnicas que descartem as informações não relevantes e coloque as características claramente em evidência em uma representação mais detalhada.

Este capítulo apresentará um conjunto de técnicas de processamento, as quais abrangem o domínio tanto da freqüência como do tempo.

3.1 Técnicas de processamento no domínio tempo

Alguns exemplos de representações do sinal de voz em termos de medições no domínio tempo incluem taxa de cruzamento por zero, energia, função de autocorrelação.

Tais representações são atrativas porque o processamento digital necessário é muito simples de implementar, e, a despeito desta simplicidade, as representações resultantes provêm uma útil base para estimar importantes características do sinal de voz.

3.1.1 Medida de Energia

A medida de energia é útil para mostrar as características da variação temporal da potência do sinal. Sendo *x[n]* a n-ésima amostra do sinal *x*, a energia *E* pode ser definida como:

$$E = \sum_{n=-\infty}^{\infty} x^2[n] \tag{3.1}$$

No caso do sinal de voz, o qual não é estacionário, a variação temporal da energia pode ser calculada da seguinte forma:

$$E[n] = \sum_{m=0}^{N-1} \left[w[m]x[n-m] \right]^2 \tag{3.2}$$

onde *w[m]* é a janela aplicada ao sinal *x[n]* e *N* é o número de amostras na janela.

A escolha de *N* deve ser adequada, pois se *N* for muito pequeno (menor que o período fundamental), *E[n]* representará muitas flutuações. Se *N* for grande demais, *E[n]* terá uma variação muito pequena, e não irá refletir de maneira adequada a

variação da amplitude do sinal de voz. Uma escolha adequada na prática para a janela é um período da ordem de 10ms a 20ms **[WEI 92]**.

O cálculo da energia pela equação 3.2 tem a propriedade de dar grande ênfase aos sinais de maior amplitude. Esse efeito pode ser minimizado usando os valores absolutos ao invés dos quadrados, tendo-se então:

$$\hat{E}[n] = \sum_{m=0}^{N-1} |w[m]x[n-m]| \tag{3.3}$$

A medida de energia pode ser usada na determinação dos limites da palavra. Neste caso, é escolhido um limiar de energia abaixo do qual o sinal é classificado como silêncio. Outra alternativa seria computar o logaritmo de *E[n]*. Esta representação facilita a observação dos sinais de baixa amplitude, como por exemplo, o ruído de fundo.

Pode-se utilizar, além do valor absoluto de energia, a energia diferencial ou transitória **[LEE 89]** como característica do locutor, pois esta medida mostra as variações relativas na amplitude do sinal. A energia diferencial *DE* é calculada em função das energias *E[n]* do sinal, como mostra a equação 3.4.

$$DE[n] = E[n+\delta] - E[n-\delta] \tag{3.4}$$

onde δ é um valor inteiro escolhido arbitrariamente **[LUF 94]**.

3.1.2 Taxa de Cruzamentos por Zero

No contexto de sinais discretos no tempo, um cruzamento por zero ocorre se sucessivas amostras têm diferentes sinais algébricos. A taxa na qual cruzamentos por zero (*zero-crossing*) ocorrem é uma medida simples do conteúdo em freqüência de um sinal. Considerando o sinal amostrado *x[n]*, pode-se definir matematicamente a função *u[n]* como sendo:

$$u[n] = \frac{x[n]}{|x[n]|} \tag{3.5}$$

onde $|x[n]|$ é o valor absoluto de *x[n]*. Desta forma, *u[n]* representa a polaridade ou o sinal de x[n]. Após definido *u[n]*, um cruzamento por zero ocorre entre os instantes de amostragem *n* e *n-1* se:

$$u[n] \neq u[n-1]. \tag{3.6}$$

A técnica mais simples de representação utilizando cruzamento por zero é a medida do número de vezes que ocorre um cruzamento por zero em um determinado intervalo de tempo (janela).

Este método é simples de ser implementado na prática, já que se resume na comparação dos sinais de duas amostras sucessivas. Entretanto, a medida de

cruzamentos por zero é extremamente sensível à presença de ruído no sinal. Uma alternativa para diminuir a influência do ruído é utilizar um detector de cruzamento por zero com um limiar **[LUF 94]**. Neste caso, a função u[n] tem um comportamento conforme é mostrado na figura 3.1, onde S é o valor de comparação, o qual deve estar acima da amplitude do ruído do sistema.

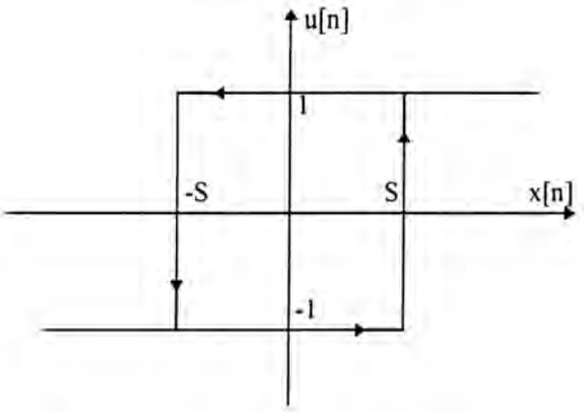

FIGURA 3.1- Função u[n] para o detector de cruzamento por zero com histerese.

A taxa de cruzamento por zero também pode ser usada para estimar as freqüências formantes do sinal de voz **[FUR 89]**. A estimativa é feita passando o sinal por uma série de filtros passa-banda e medindo-se a taxa de cruzamento por zero dos sinais da saída dos filtros que apresentam os maiores níveis de energia.

A medida de cruzamentos por zero é também aplicada na detecção dos limites das palavras e pode ser utilizada como representação para o sinal de voz em sistemas simples de reconhecimento **[RAB 78][LUF 94]**.

3.1.3 Autocorrelação

A função de autocorrelação de um sinal determinístico discreto no tempo é definida como

$$R[i] = \sum_{n=-\infty}^{\infty} s[n]s[n+i] \tag{3.7}$$

onde *s[n]* é o sinal de entrada e *i* é o deslocamento para o cálculo da correlação.

A representação da função de autocorrelação do sinal é um modo conveniente de mostrar certas propriedades do sinal. Por exemplo, se o sinal é periódico com período de P amostras, então pode ser facilmente demonstrado que

$$R[i]=R[i+P] \tag{3.8}$$

isto é, a função de autocorrelação de um sinal periódico é também periódico com o mesmo período. Outras importantes propriedades da função autocorrelação são:

1. É uma função par, isto é, *R[i] = R[-i]*;

2. O seu valor máximo se encontra em i = 0, isto é, $/R[i]/ \leq R[0]$ para todo i;
3. O valor de R[0] é igual à energia para sinais determinísticos ou a potência média para sinais periódicos ou aleatórios.

Assim, a função de autocorrelação contém a função energia como um caso especial. Se considerar a equação 3.8 juntamente com as propriedades 1 e 2, percebe-se que, para os sinais periódicos, a função autocorrelação fornece um máximo nas amostras 0, ±P, ±2P, ...

3.2 Técnicas de processamento no domínio freqüência

Alguns exemplos de representações do sinal de voz em termos de medições no domínio freqüência são a análise espectral, a análise cepstral e a codificação preditiva linear.

3.2.1 Análise espectral

A medida das freqüências contidas no sinal de voz é uma das técnicas mais importantes na análise das caraterísticas acústicas da fala. Para a análise em freqüência do sinal é utilizada a Transformada de Fourier, dada por:

$$X\left[e^{j\omega}\right] = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}. \tag{3.9}$$

onde ω é a freqüência.

Como o sinal em um curto espaço de tempo é invariante, a aplicação da equação 3.9 seria impossível devido ao intervalo infinito em que é realizada a transformada. Para isso, o cálculo é realizado pela Transformada de Fourier Discreta (DFT - *Discrete Fourier Transform*), dada pela expressão:

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi kn/N}, k = 0,1,\ldots,N-1. \tag{3.10}$$

A DFT pode ser calculada de forma eficiente por algoritmos de Transformada Rápida de Fourier (FFT - *Fast Fourier Transform*) **[RAB 78]**.

O sinal x[n] é geralmente multiplicado por uma janela no tempo w[n], com a duração de N amostras. Na equação 3.10 foi usada uma janela retangular (isto é, w[n] = 1 para 0 ≤ n < N). Outras janelas como a janela de *Hamming* e *Hanning* **[EMB 91]** são geralmente utilizadas.

A transformada rápida de Fourier (FFT) é um algoritmo muito eficiente para a computação da DFT (*Discrete Fourier Transform* - Transformada Discreta de Fourier) de uma seqüência. Ele leva a vantagem pelo fato de que muitas computações são repetidas na DFT devido à natureza periódica da transformada discreta de Fourier: $e^{-j2\pi kn/N}$. A forma da DFT é:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N}. \qquad (3.A1)$$

Assumindo que $W^{nk} = e^{-j2\pi nk/N}$, a equação 3.11 pode ser rescrita como:

$$X(k) = \sum_{n=0}^{N-1} x(n) W^{nk}. \qquad (3.12)$$

agora, $W^{(N+qN)(k+nN)} = W^{nk}$ para todo **q**, **n**, os quais são inteiros devido à periodicidade do princípio de Fourier.

Após isto, pode-se quebrar a DFT em duas partes

$$X(k) = \sum_{n=0}^{N/2-1} x(2n) W_N^{2nk} + \sum_{n=0}^{N/2-1} x(2n+1) W_N^{(2n+1)k} \qquad (3.13)$$

onde o subscrito **N** no cerne de Fourier representa o tamanho da seqüência. Se representar todos os elementos pares da seqüência x(n) por x_{ev} e os elementos ímpares por x_{od}, então a equação pode ser rescrita conforme mostra a equação 3.14.

$$X(k) = \sum_{n=0}^{N/2-1} x_{ev}(n) W_{N/2}^{2nk} + W_{N/2}^{k} \sum_{n=0}^{N/2-1} x_{od}(n) W_{N/2}^{nk} \qquad (3.14)$$

Agora tem-se duas expressões na forma de DFTs, que podem ser escritas como

$$X(k) = X_{ev}(k) + W_{N/2}^{k} X_{od}(k) \qquad (3.15)$$

Nota-se que somente DFTs de N/2 pontos necessitam ser calculadas para encontrar o valor de X(k). Sabendo que o índice **k** deve variar até N-1 e utilizando a propriedade de periodicidade das DFTs pares e ímpares , temos que:

$$X_{ev}(k) = X_{ev}(K - n/2) \qquad \text{para } n/2 \le K \le N-1 \qquad (3.16)$$

O processo de divisão das DFTs resultantes em metades par e ímpar pode ser repetido até que uma seja deixada com somente dois pontos de DFTs para avaliar

$$\begin{aligned} \Lambda(k) &= \lambda(0) + \lambda(1) e^{-j2\pi k/2} && \text{para todo k} \\ &= \lambda(0) + \lambda(1) && \text{para todo k par} \\ &= \lambda(0) - \lambda(1) && \text{para todo k ímpar .} \end{aligned}$$

Assim, para DFTs de dois pontos nenhuma multiplicação é necessária, somente adições e subtrações. Para computar a DFT completa, ainda necessita-se efetuar multiplicações de DFTs individuais de dois pontos pelos apropriados fatores de **W**, de intervalo de W^0 à $W^{N/2-1}$.

Para a DFT original, **N** multiplicações complexas são necessárias para cada um dos **N** valores de **k**. Também, N - 1 adições são necessárias para cada **k**.

Em uma FFT cada função da forma

$$\lambda(0) = W^{p}\lambda(1)$$

(chamada uma borboleta devido a forma do fluxo do grafo) requer uma multiplicação e duas adições. Pode-se generalizar que o número de borboletas é

$$NúmerodeBorboletas = \frac{N}{2}\log_2(N)$$

porque existem N/2 linhas de borboletas (desde que cada borboleta tenha duas entradas) e existem $\log_2(N)$ colunas de borboletas.

A tabela 3.1 fornece uma listagem de adições e multiplicações, de elementos complexos, para vários tamanhos de FFTs e DFTs. Essa tabela mostra que a FFT é o método de análise espectral mais prático em muitos casos onde a computação de DFT consumiria muito mais tempo.

TABELA 3.A- Comparação do Número de Operações Borboletas em DFT e FFT.

Tamanho da Transformada (N)	Operações DFT (N^2)	Operações FFT $N\text{LOG}_2(N)$
8	64	24
16	256	64
32	1024	160
64	4096	384
128	16384	896
256	65536	1024
512	262144	4608
1024	1048576	10240
2048	4194304	22528

Um dos usos do cálculo do espectro do sinal de voz é produzir um espectrograma, o qual mostra a energia do sinal em função da freqüência e do tempo. Esta representação pode ser usada para o reconhecimento de locutor, pois permite obter a freqüência fundamental e as formantes, entre outros.

As componentes espectrais obtidas através da FFT estão distribuídas linearmente em intervalos iguais de freqüência. Em sistemas de reconhecimento de voz, estes valores são geralmente redistribuídos em uma escala logarítmica de freqüência, levando em consideração as características do sistema auditivo humano.

Duas escalas propostas para este fim são a escala de Bark e a escala de Mel. Elas são definidas respectivamente pelas equações 3.17 e 3.18 **[LUF 94]**.

$$B = 13\arctan(0.76f) + 3.5\arctan\left(\frac{f}{7.5}\right)^2 \tag{3.17}$$

$$Mel = 1000\log_2(1+f) \tag{3.18}$$

onde f é a freqüência em kiloHertz.

3.2.2 Análise Cepstral

O cepstro (coeficientes cepstrais) é definido como sendo a transformada inversa de Fourier do logaritmo da magnitude da transformada de Fourier, isto é,

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X(e^{j\omega})\right| e^{j\omega n} d\omega, \qquad -\infty < n < \infty \tag{3.19}$$

Técnicas baseadas no cepstro podem ser usadas em sistemas que obedecem ao princípio da superposição, como é o caso dos sistemas lineares.

Como o sinal de voz pode ser produzido por um sistema linear, pode-se utilizar a análise cepstral para separar a excitação g(t) da resposta ao impulso do trato vocal h(t). Então, o sinal de voz s(t) é dado pela convolução (*) de g(t) com h(t):

$$s(t) = g(t) * h(t) \tag{3.20}$$

que é equivalente a

$$S(w) = G(w) * H(w) \tag{3.21}$$

onde S(w), G(w) e H(w) são as transformadas de Fourier de s(t), g(t) e h(t), respectivamente.

Tomando o logaritmo dos módulos obtém-se

$$\log|S(w)| = \log|G(w)| + \log|H(w)| \tag{3.22}$$

Os coeficientes cepstrais de s(t) são então obtidos por

$$c(\tau) = F^{-1}\log|G(w)| + F^{-1}\log|H(w)| \tag{3.23}$$

onde F^{-1} é a transformada inversa de Fourier.

Calculando os coeficientes cepstrais para uma janela de N amostras tem-se

$$c[n] = \frac{1}{N}\sum_{k=0}^{N-1} \log|X[k]| e^{j2\pi kn/N} \tag{3.24}$$

para $0 \leq n \leq N - 1$.

A figura 3.2 **[RAB 78]** exemplifica o processo de análise cepstral aplicada ao sinal de voz.

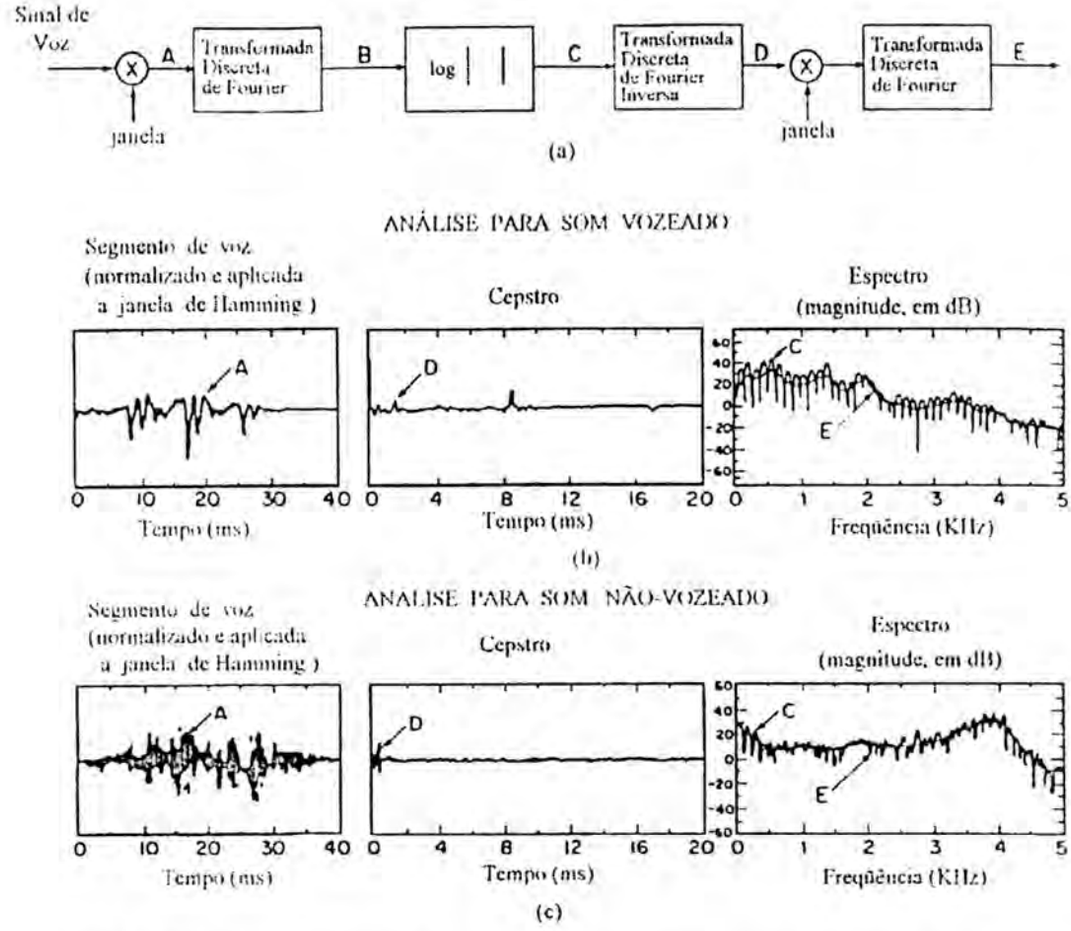

FIGURA 3.2 - Análise cepstral de um segmento de um sinal de voz (a) Operações básicas. (b) Análise para um sinal vocálico. (c) Análise para um sinal não-vocálico.

Os resultados das operações apresentadas na figura 3.2a podem ser vistos em (b) e (c) para sons vocálicos e não vocálicos respectivamente. A curva C é o logaritmo do espectro do sinal de entrada A. Esta curva apresenta duas componentes, uma que contém a variação suave das componentes espectrais relacionada à função de transferência do trato vocal, e outra que varia rapidamente em função da freqüência causada pela excitação. A parcela que varia lentamente em A produz as componentes baixas na escala de tempo do cepstro D. No caso (b), onde o segmento de voz é vocálico, a componente periódica que aparece em C reflete-se no cepstro como um forte pico; neste caso, aproximadamente em 8ms (*pitch*). Em (c) este pico não aparece devido à natureza aleatória da excitação. Deste modo, a componente de variação rápida no espectro não é periódica. Devido a estas características, o cepstro serve para estimar o período fundamental da voz e para determinar quando um determinado segmento de fala é vocálico ou não-vocálico.

A função de transferência do trato vocal, também chamada de envelope espectral, pode ser obtida através do cepstro, multiplicando o mesmo por uma janela de freqüência que deixe passar somente componentes de variação lenta no domínio tempo, e calculando a DFT do cepstro resultante. O resultado deste processo é a curva E.

Como os primeiros coeficientes cepstrais estão relacionados com o envelope espectral, os mesmos podem ser utilizados para representar o sinal de voz. Além de representar a função de transferência do trato vocal, estes coeficientes não contêm informação da fonte de excitação do trato g(t), o que torna esta análise não totalmente confiável para o processo de reconhecimento de locutor, pois a excitação é bastante variável com relação ao locutor.

Apesar da representação cepstral da voz ser utilizada em sistemas de reconhecimento, o método de cálculo apresentado na definição acima geralmente não é empregado em sistemas de reconhecimento de locutor em tempo real devido à complexidade computacional. Na prática, eles são derivados da representação do trato vocal obtidos a partir da técnica de predição linear, o que será apresentado na seção 3.2.3.

3.2.3 Codificação Linear Preditiva (LPC)

Na representação do sinal de voz, em vez de extrair parâmetros que descrevem o sinal de voz, extrai-se características que modelem o trato vocal. Como já mostrado, pode-se modelar a geração da fala com um filtro que é excitado por uma seqüência quase periódica de impulsos ou por uma fonte de ruído aleatório. Os parâmetros do filtro determinam as características espectrais do som emitido. Portanto, considerando o modelo da figura 2.6, pode-se representar a voz pelos parâmetros do filtro *H(z)*.

A idéia básica da análise preditiva linear é que a amostra de voz pode ser aproximada como uma combinação linear das amostras de voz passadas. Pela minimização da soma dos quadrados das diferenças (sobre um intervalo finito) entre a amostra de voz real e a linearmente predita, um único conjunto de coeficientes do preditor (coeficientes ponderados usados na combinação linear) pode ser determinado.

De acordo com a figura 2.6, o filtro digital variante no tempo poderia ser *H(z)*, a entrada para o filtro seria *u[n]* e a saída gerada *s[n]* para um sistema hipotético. Assim, pode-se assumir que o sinal *s[n]* é dado por uma combinação linear das saídas anteriores e da entrada atual *u[n]*:

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Gu[n] \qquad (3.25)$$

onde a_k (coeficientes do preditor) e G (fator de ganho para a entrada) são os parâmetros do sistema. Este sistema é excitado por uma seqüência quase periódica de impulsos ou por uma fonte de ruído aleatório. Assim, os parâmetros do filtro (classificação de som vocálico e não-vocálico, período do *pitch* para som vocálico,

parâmetro de ganho G, e os coeficientes $\{a_k\}$ do filtro digital) determinam as características espectrais do som emitido.

A função de transferência H(z), no domínio freqüência, pode ser obtida pela transformada Z, como mostra a equação 3.26.

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.26}$$

onde G é o fator de ganho para a entrada. A figura 3.3 mostra este modelo digital no domínio tempo.

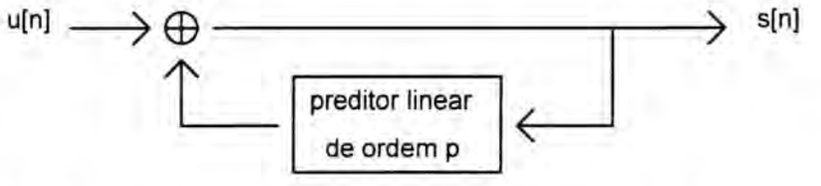

FIGURA 3.3 - Modelo digital de produção de voz, no domínio tempo.

Observando a figura 3.3, percebe-se que, com exceção da primeira amostra de cada período de *pitch*, todas as amostras de sons vocálicos são linearmente preditas em função das últimas p amostras. Esta propriedade é utilizada para determinar os parâmetros do preditor. O erro de predição *e[n]* é definido como a diferença entre a amostra de voz s[n] e o sinal predito $\tilde{s}[n]$ onde:

$$\tilde{s}[n] = \sum_{k=1}^{p} a_k s[n-k] \tag{3.27}$$

e

$$e[n] = s[n] - \tilde{s}[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k] \tag{3.28}$$

Os parâmetros a_k podem ser obtidos pelo método dos mínimos quadrados, onde o erro é minimizado para cada um dos parâmetros. O erro quadrático total E é:

$$E = \sum_n e^2[n] = \sum_n \left(s[n] + \sum_{k=1}^{p} a_k s[n-k] \right)^2 \tag{3.29}$$

Os valores de a_k que minimizam E são obtidos quando

$$\frac{\delta E}{\delta a_i} = 0, \qquad i = 0,1,\ldots,p \tag{3.30}$$

Substituindo a equação 3.29 em 3.30 obtém-se **[MAK 75]**

$$\sum_{k=1}^{p} a_k \sum_{n} s[n-k]s[n-i] = \sum_{n} s[n]s[n-i] \qquad i = 1,2,\ldots,p. \tag{3.31}$$

Os coeficientes a_k são calculados resolvendo-se este sistema de p equações e p incógnitas. Para o cálculo da equação 3.31 deve ser definida a faixa de valores n sobre a qual é efetuada a soma. Por causa da natureza variante no tempo do sinal de voz, a análise deve ser feita em intervalos curtos de tempo no qual as características do mesmo podem ser consideradas quase constantes. Neste caso aplica-se uma janela $w(m)$ para diminuir o efeito dos cortes nos limites do segmento, como mostra a equação 3.32.

$$s_n(m) = s(m+n)w(m) \tag{3.32}$$

onde $w(m)$ é igual a 0 (zero) fora do intervalo $0 \leq m \leq N-1$, e dentro do intervalo o valor é definido pelo tipo de janela (a tabela 2.1 apresenta alguns tipos de janelas).

Para a solução do sistema pode-se utilizar três métodos: autocorrelação, covariância e Lattice **[RAB 75]** .

A função de autocorrelação do sinal s[n] é dada pela equação 3.7, que é uma função par. Neste caso, a equação 3.7 fica:

$$\sum_{k=1}^{p} a_k R[i-k] = -R[i], \qquad 1 \leq i \leq p \tag{3.33}$$

Os coeficientes R[i - k] formam a matriz de autocorrelação; por isso o método baseado na equação 3.7 é conhecido como método da autocorrelação. A matriz de autocorrelação é uma matriz simétrica onde todos os elementos de cada diagonal são iguais (matriz Toeplitz). Expandindo a equação 3.33 na forma matricial tem-se:

$$\begin{bmatrix} R_0 & R_1 & R_2 & \cdots & R_{p-1} \\ R_1 & R_0 & R_1 & \cdots & R_{p-2} \\ R_2 & R_1 & R_0 & \cdots & R_{p-3} \\ \vdots & \vdots & \vdots & & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_p \end{bmatrix} \tag{3.34}$$

Por causa das propriedades especiais da matriz de coeficientes é possível solucionar as equações de maneira mais eficiente pelo algoritmo recursivo de Durbin **[MAK 75]**. O método de Durbin segue o seguinte procedimento:

$$E_0 = R_0 \tag{3.35a}$$

$$k_i = \left[R_i - \sum_{j=1}^{i-1} a_j^{(i-1)} R_{i-j} \right] \Big/ E_{i-1} \tag{3.35b}$$

$$a_i^{(i)} = k_i \tag{3.35c}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \qquad 1 \le j \le i-1 \tag{3.35d}$$

$$E_i = (1-k_i^2)E_{i-1} \tag{3.35e}$$

onde $a_j^{(i)}$ representa a i-ésima iteração de a_j. O conjunto de equações (3.35a) a (3.35e) é resolvido recursivamente para i=1,2,...,p. A solução final é dada por:

$$a_j = a_j^{(p)}, \qquad 1 \le j \le p. \tag{3.35f}$$

Percebe-se que para obter a solução para um preditor de ordem **p**, são calculadas as soluções de todos preditores de ordem inferior a **p**. A solução da equação 3.34 é de pequena complexidade computacional, se comparada com o volume de cálculo necessário para a determinação dos elementos da matriz de autocorrelação ou de covariância. A determinação de cada um destes elementos requer uma quantidade de cálculo da ordem de (pN). A resolução da equação 3.34 pelo algoritmo das equações 3.35 requer cerca de $p^2 + O(p)$ operações, que geralmente é bastante inferior ao necessário para o cálculo das correlações, pois normalmente se trabalha com N>p.

A solução da equação 3.34 não é afetada se os coeficientes *R[i]* forem normalizados. Os coeficientes de autocorrelação normalizados *r[i]* são definidos como:

$$r[i] = \frac{R[i]}{R[0]} \tag{3.36}$$

Esta normalização é útil em aplicações cujas operações utilizam aritmética de ponto fixo, pois $|r[i]| \le 1$.

O valor E_i, calculado a cada iteração na equação 3.35e, é o erro mínimo total e deve diminuir ou permanecer o mesmo à medida que a ordem do preditor cresce **[MAK 75]**. Portanto:

$$0 \le E_i \le E_{i-1}, \qquad E_0 = R[0] \tag{3.37}$$

Quando os coeficientes de autocorrelação normalizados são usados, o erro mínimo total E_i também é dividido por *R[0]*.

Para o método da autocorrelação, **[WEI 92]** mostra que o ganho G é determinado por:

$$G^2 = R[0] + \sum_{k=1}^{p} \alpha_k R[k] \tag{3.38}$$

onde G^2 é a energia total da entrada (tanto para ruído branco como para impulsos).

Existe outro método para determinar os coeficientes α_k a partir da equação 3.35 chamado método da covariância. A definição deste método pode ser encontrada

em **[RAB 78] [ADA 96]**. Neste método, o erro E na equação 3.29 é minimizado para um intervalo finito. Neste caso, para $0 \leq n \leq N-1$, a equação 3.31 resulta em:

$$\sum_{k=1}^{p} \alpha_k \varphi_{ki} = -\varphi_{0i}, \qquad 1 \leq i \leq p \tag{3.39}$$

onde

$$\varphi_{ik} = \sum_{n=0}^{N-1} s[n-i]s[n-k] \tag{3.40}$$

é a covariância do sinal $s[n]$ para o intervalo de N amostras. Os coeficientes φ_{ki} formam a matriz de covariância. Esta matriz é simétrica ($\varphi_{ki} = \varphi_{ik}$), entretanto, os elementos de cada diagonal não são idênticos como na matriz de autocorrelação.

Calculados os coeficientes do filtro $H(z)$, é necessário avaliar se o filtro resultante é estável. Para um filtro deste tipo ser estável é necessário que os pólos, que são as raízes do denominador, estejam dentro de círculo de raio unitário. Para o método de autocorrelação, a estabilidade é garantida se R_i é calculado a partir de um sinal não-nulo, o que não é válido para o método da covariância **[MAK 75]**.

Mesmo no método da autocorrelação, a estabilidade de $H(z)$, conforme a equação 3.26, deve ser verificada, pois erros de arredondamento podem levar a uma solução onde os pólos se localizem fora do círculo unitário. Ao invés de calcular as raízes do denominador de $H(z)$, pode-se determinar a estabilidade através do algoritmo (equação 3.35), onde a condição $E_i > 0,\ 1 \leq i \leq p$, é necessária e suficiente para garantir a estabilidade de $H(z)$. Outra condição equivalente é $|k_i| < 1,\ 1 \leq i \leq p$. Quando o método de resolução da equação 3.31 não é o algoritmo de Durbin, a estabilidade só pode ser garantida por outros métodos de teste. Um método é calcular os coeficientes k_i pela equação 3.44 a seguir e então verificar a estabilidade.

Os coeficientes do filtro preditor α_k podem ser usados diretamente como representação da voz em sistemas de reconhecimento. Além disso, os coeficientes podem ser usados de diversas maneiras para representar as propriedades do sinal de voz.

No modelo proposto, o filtro H(z) representa a função de transferência do trato vocal. Portanto, a resposta em freqüência do trato é:

$$\left|H\left(e^{j\omega}\right)\right|^2 = \frac{G^2}{\left|1 + \sum_{k=1}^{p} \alpha_k e^{-jk\omega}\right|^2} \tag{3.41}$$

onde ω é a freqüência.

Uma forma simplificada da equação 3.41 é apresentada na equação 3.42.

$$\left|H\left(e^{j\omega}\right)\right|^2 = \frac{G^2}{\rho[0] + 2\sum_{i=1}^{p}\rho[i]\cos(i\omega)} \tag{3.42}$$

onde

$$\rho[i] = \sum_{k=0}^{p-i}\alpha_k\alpha_{k+i} \qquad \alpha_0 = 1, 0 \le i \le p \tag{3.43}$$

são os coeficientes de autocorrelação de α_k.

A figura 3.4 mostra um exemplo onde o espectro foi obtido usando a equação 3.41 para diversas ordens do preditor. Neste exemplo, o sinal de voz (fonema /a/) foi amostrado a 10 kHz. O método empregado foi o da autocorrelação e o sinal foi multiplicado pela janela de Hanning com $N = 200$.

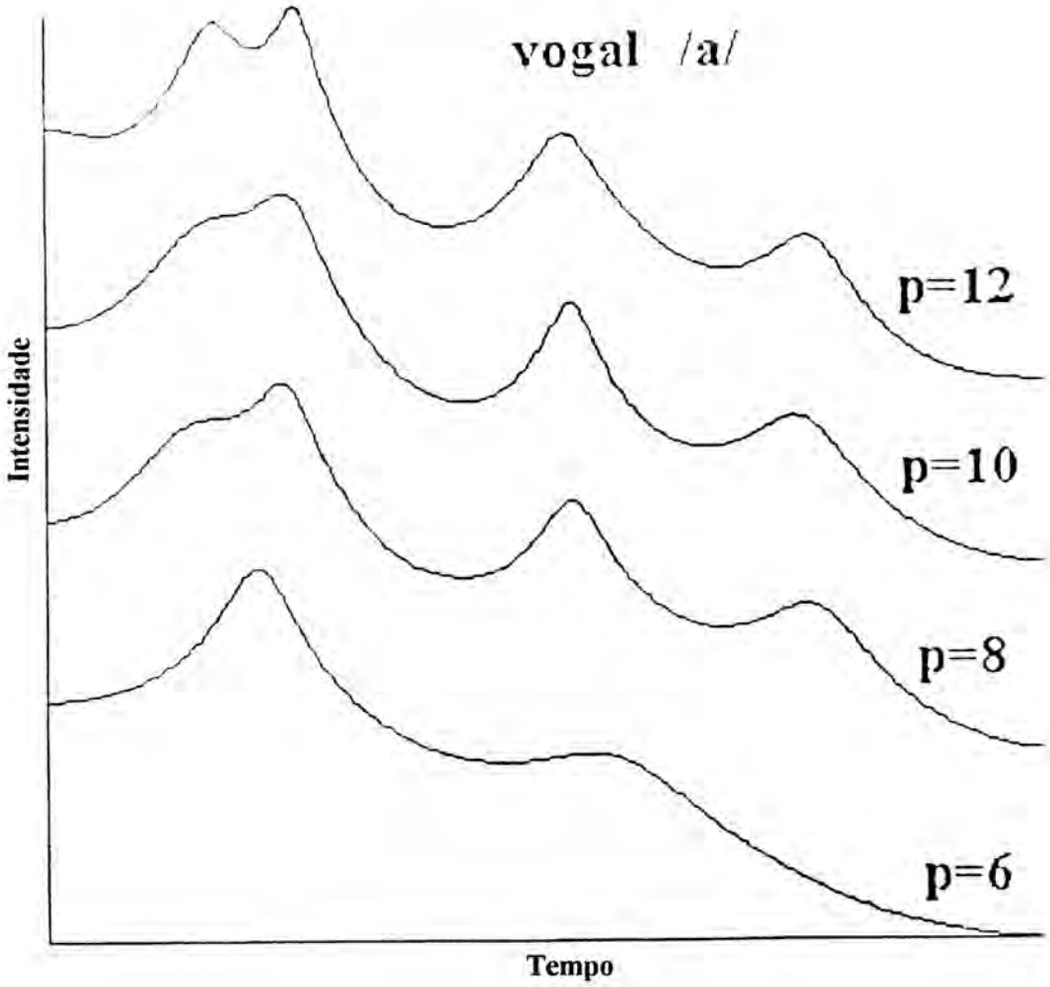

FIGURA 3.4 - Espectro do sinal de voz obtido através da função de transferência do trato vocal.

O espectro mostrado na figura 3.4 pode ser usado para determinar as freqüências formantes. Isto pode ser feito automaticamente por um algoritmo de detecção de picos.

Outro parâmetro que pode ser usado para representar as características da voz são os coeficientes de reflexão, também chamados coeficientes de correlação parcial. Os coeficientes de reflexão k_i são obtidos como um resultado do cálculo dos coeficientes α_k pelo algoritmo de Durbin(equação 3.35b). Estes também podem ser obtidos a partir dos coeficientes do filtro preditor através do seguinte algoritmo:

$$k_i = \alpha_i^{(i)} \tag{3.44a}$$

$$\alpha_j^{(i-1)} = \frac{\alpha_j^{(i)} - \alpha_i^{(i)} \alpha_{i-j}^{(i)}}{1 - k_i^2}, \qquad 1 \le j \le i-1 \tag{3.44b}$$

onde i=p, p-1, ..., 2, 1. A condição inicial é

$$\alpha_j^{(p)} = \alpha_j, \qquad 1 \le j \le p$$

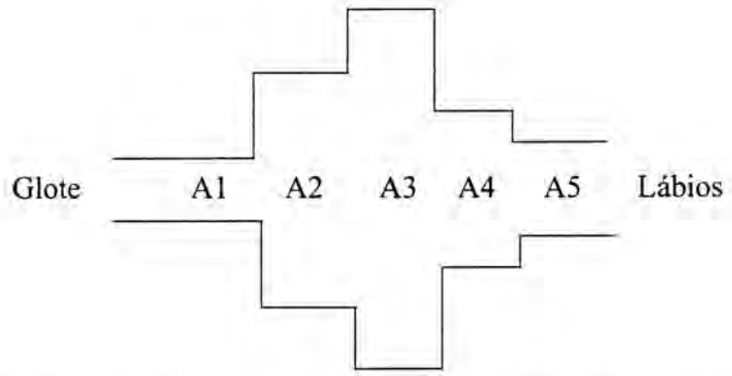

FIGURA 3.5 - Exemplo de modelo para produção de fala utilizando a concatenação de tubos.

Considerando-se o trato vocal como uma seqüência de tubos com diferentes áreas (como mostra a figura 3.5), k_i pode ser considerado como o coeficiente de reflexão entre duas seções. A relação entre as impedâncias acústicas de duas seções consecutivas é dada por **[MAK 75]**

$$\frac{Z_{i+1}}{Z_i} = \frac{1+k_i}{1-k_i}, \qquad 1 \le i \le p \tag{3.45}$$

Os coeficientes de reflexão também podem ser usados para estimar a área do trato vocal **[FUR 86]**.

Como já apresentado em 3.2.2, outra representação da voz de grande interesse para o processo de reconhecimento são os coeficientes cepstrais. Os coeficientes cepstrais que estão relacionados com a função de transferência do trato vocal são os primeiros coeficientes na escala de tempo (características espectrais de variação

lenta). Tais parâmetros podem ser diretamente obtidos a partir dos coeficientes de predição linear α_k pela fórmula de recorrência **[MAK 75].**

$$c[n] = \alpha_n - \sum_{m=1}^{n-1} \frac{m}{n} c[m]\alpha_{n-m}, \qquad 1 \le n \le p \tag{3.46}$$

4 Extração das Características

A identidade do locutor está correlacionada com as suas características fisiológicas e comportamentais **[MOR 91]**. Estas características existem no envelope espectral (características do trato vocal) e nas características supra-segmentais (características da fonte da fala) da voz. Mas, é impossível separar estes tipos de características, e muitas características da voz são difíceis de medir explicitamente.

Dentre as medidas do sinal, encontram-se as técnicas apresentadas no capítulo 3. Além destas tem-se a freqüência fundamental e as freqüências formantes, as quais são abordadas neste capítulo.

Estas medidas no problema do reconhecimento de locutor têm por objetivo discriminar efetivamente um locutor do outro.

4.1 Pré-processamento do sinal

Para um correto reconhecimento é necessário que se extraia características do sinal de voz, a fim de que tais características representem as informações necessárias para a solução do problema.

Por isso, antes de utilizar algum método apresentado no capítulo 3, realiza-se um processamento de filtragem e/ou seleção. Esta tarefa é chamada de pré-processamento do sinal, onde procura-se eliminar segmentos desnecessários, eliminar alguma interferência na amostra de voz (ruído de fundo, ruído do microfone, normalização do sinal).

Nas seções 4.1.1 e 4.1.2 são apresentados dois algoritmos que realizam a filtragem do sinal e eliminação de silêncio, respectivamente.

4.1.1 Pré-ênfase do Sinal

Segundo **[WIT 82]** **[MAR 76]**, há uma tendência espectral de aproximadamente -6dB/ oitava na fala radiada dos lábios como aumento da freqüência e -12db/oitava velocidade do volume glotal da forma de onda. Esta tendência deve ser eliminada para que o sinal torne-se mais puro.

Para um sistema digital, tais pré-ênfases podem ser implementadas como um circuito analógico, o qual precede o filtro e o digitalizador de pré-amostragem, ou como uma operação digital no sinal amostrado e quantizado. No primeiro caso, a característica é usualmente plana para um certo ponto, o que ocorre em algum lugar entre 100Hz e 1kHz (a exata posição não é crítica), em cujo ponto a ascensão de +6dB/oitava começa.

O efeito de ascensão de +6dB/oitava pode ser obtida pela diferenciação da entrada. A operação

$$y(n) = x(n) - ax(n-1) \tag{4.1}$$

é adequada, onde o parâmetro constante a é usualmente escolhido entre 0.9 e 1, e x(n) é o sinal amostrado.

4.1.2 Determinação Automática dos Limites da Palavra

A localização do início e fim da palavra pronunciada, ou detecção dos limites, é um problema diretamente associado ao reconhecimento de palavras isoladas.

Para que o processo de comparação tenha resultados eficientes, é necessário que a detecção dos limites seja capaz de localizar os diversos eventos acústicos presentes na palavra. Tal processo pode ser de fácil execução se o sinal de voz for amostrado com grande diferença em amplitude do ruído de fundo (boa qualidade do meio de aquisição e/ou ambientes não ruidosos).

Para uma aplicação em tempo real, um algoritmo de localização dos limites deve atender a requisitos como simplicidade e robustez, onde a detecção deve ocorrer de maneira síncrona com a pronúncia da palavra.

O tempo de processamento necessário para esta tarefa deve ser pequeno se comparado ao processo de reconhecimento do locutor. O algoritmo também deve adaptar-se às variações do ruído de fundo. O principal parâmetro do sinal de voz utilizado na detecção dos limites é a energia do sinal (seção 3.1.1).

No caso em que o ruído de fundo é desprezível, a tarefa de detecção é trivial, sendo suficiente determinar um patamar de energia acima do nível de ruído e comparar a energia do sinal com este patamar. O sinal é então classificado como silêncio ou sinal de voz se estiver, respectivamente, abaixo ou acima deste nível.

O início da palavra é considerado no primeiro ponto onde a energia ultrapassa o patamar especificado. Para a detecção do fim da palavra deve ser considerado que durante a pronúncia de certas palavras ocorrem períodos de silêncio. Por exemplo, quando no meio de uma palavra existem consoantes com sons plosivos, como /p/ e /t/, estas são precedidas de um período onde nenhum som é emitido. Portanto, é necessário especificar um período mínimo de silêncio entre duas palavras que seja maior que a duração do silêncio durante a pronúncia das mesmas. Na prática, um intervalo de 150ms é aceitável; se a pronúncia for muito lenta, porém este valor deve ser aumentado.

Quando a medida de energia não é suficiente para separar sons fricativos, como /s/ e /f/, do ruído de fundo, pode-se utilizar a medida da taxa de cruzamentos por zero.

[RAB 78] apresenta um algoritmo que utiliza a energia e a taxa de cruzamentos por zero. Neste algoritmo a primeira aproximação para os limites é feita através da energia. Dois patamares de energia são determinados e a detecção ocorre quando os dois níveis são ultrapassados, conforme ilustra a figura 4.1.

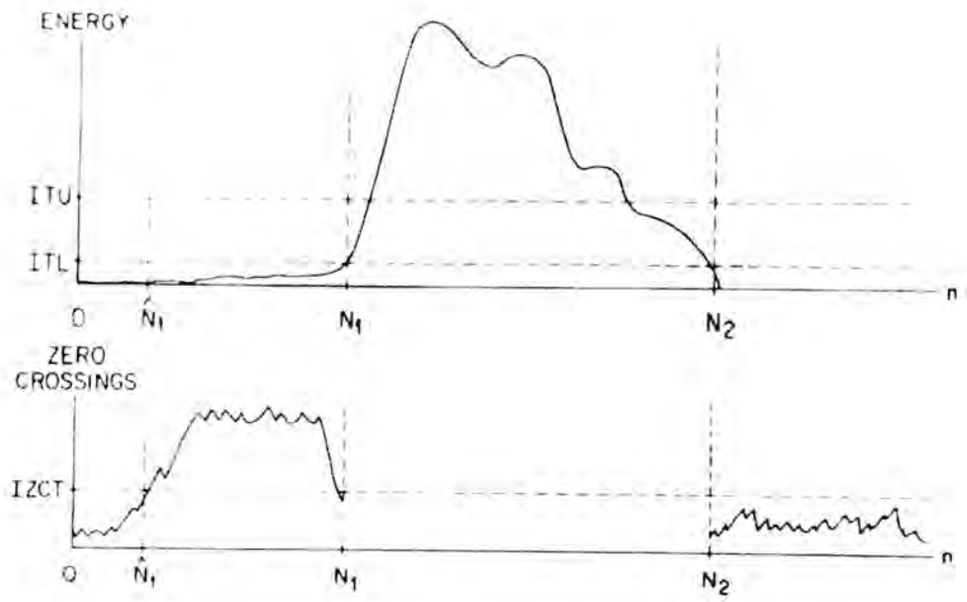

FIGURA 4.1 - Exemplo de detecção dos limites da palavra pela energia e taxa de cruzamento por zero **[RAB 78]**.

A partir dos pontos N_1 e N_2 detectados pela energia, é analisada a taxa de cruzamento por zero. Para o ponto inicial é feita uma pesquisa para trás (no tempo). Se a taxa de cruzamento por zero excede um limite (ZCT - *Zero Crossing Threshold*) três ou mais vezes, o início da palavra é escolhido no instante em que a taxa de cruzamento por zero ultrapassa o limite ZCT, no instante $\hat{N}_1$. Caso contrário, N_1 é escolhido. Procedimento semelhante é executado para o fim da palavra. Na figura 4.1 foram escolhidos os limites N_1 e N_2. Estes algoritmos estão descritos a seguir.

Variáveis e Funções

E(t):	medida de energia do segmento t;
TAM_SINAL:	número de amostras de sinal;
Tmin:	duração mínima do sinal;
Tint:	mínimo de silêncio entre as palavras;
PS:	patamar de silêncio.

Algoritmo Detecção_Início_de_Palavra

```
Tmin=550
t = 0
inicio=0
REPITA
        SE E(t) > PS
                ENTÃO
                        i=t
                        REPITA
                                t=t+1
                        ATE QUE (E[t] <= PS) OU (Tmin > t-i)
```

```
                    SE Tmin > t-i
                        ENTÃO
                            FIM REPITA
                    FIM SE
                FIM SE
ATÉ QUE t < TAM_SINAL
FIM Algoritmo


Algoritmo Detecção_Fim_de_Palavra

Tmin = 550
Tint = 2200
t = 0
inicio=0
REPITA
        SE E(t) <= PS
                ENTÃO
                        f=t
                        REPITA
                                i = t
                                REPITA
                                        t = t +1
                                ATE QUE ((E[t] > PS) E (t-i>= Tmin)) OU
                                ((E[t] <= PS) E (Tint >= t-f))
                        ATE QUE (Tint > t-f) OU (t-i>= Tmin)
                        SE Tint < t-f
                                ENTÃO
                                        FIM REPITA
                        FIM SE
        FIM SE
ATÉ QUE t < TAM_SINAL

FIM Algoritmo
```

4.2 Extração do Pitch

O período fundamental ou freqüência fundamental (*pitch*) refere-se à periodicidade do sinal para cada locutor. Este parâmetro é característico de cada pessoa, devido a fatores **[ADA 96]** como, por exemplo, fisiologia do aparelho vocal, idade, sexo. Para a extração do *pitch*, pode-se utilizar alguma das técnicas descritas a seguir.

1. Função de Autocorrelação;
2. Predição Linear;
3. Cepstro
4. Transformada de Fourier;

Dentre estas técnicas, a que é empregada neste trabalho é a obtida através da função de autocorrelação. Esta técnica é muito utilizada **[RAB 78] [SOR 95]**, pois a função de autocorrelação de um sinal periódico é periódica e tem o mesmo período que o sinal. Para isso, o processo consiste em obter o segundo maior pico da função, devido à característica de que o primeiro pico é sempre R[0] - energia.

Uma das maiores limitações da representação da autocorrelação é que a mesma detém muita informação do sinal de voz **[RAB 78]**. Esta limitação pode ser vista na figura 4.2, na qual existem muitas oscilações/picos (causados pelo trato vocal, o qual é responsável pela forma de cada período da onda de fala).

Os picos apresentados nas figuras 4.2a e 4.2b são maiores do que na figura 4.2c. Isto ocorre, neste caso, porque a janela é curta se comparada ao período de *pitch*, mas também é atribuída à mudança rápida das freqüências das formantes. No caso em que os picos da autocorrelação, devido à resposta do trato vocal, forem maiores que aqueles, devido à periodicidade da excitação vocal, o procedimento de obter o maior pico na função de autocorrelação falhará.

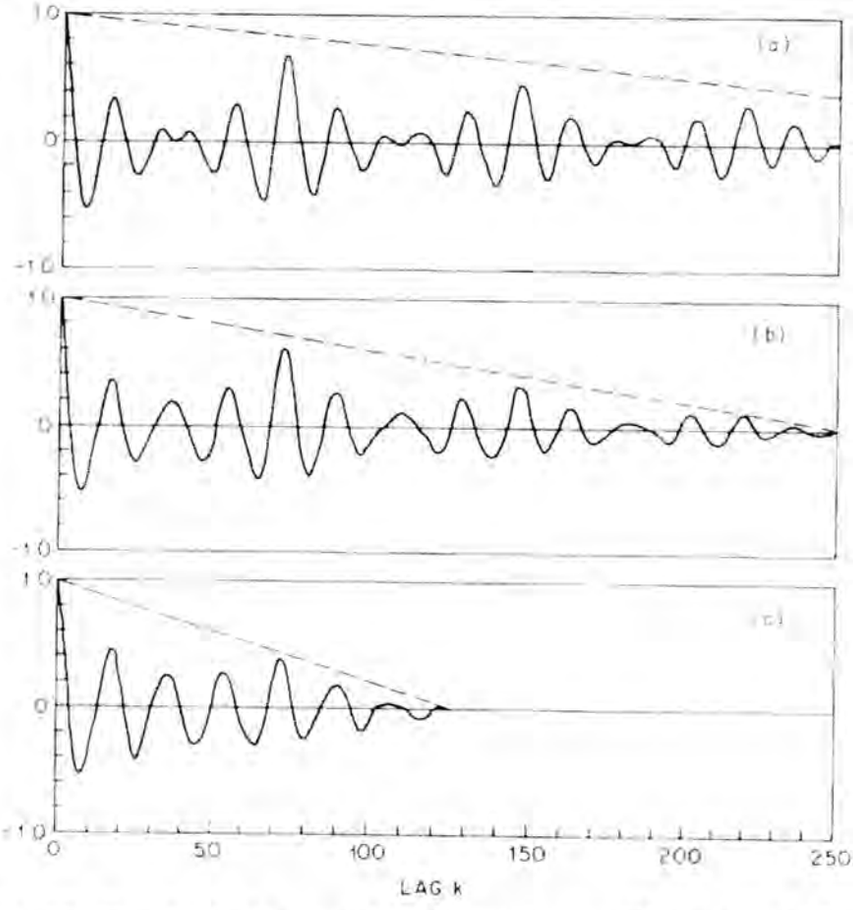

FIGURA 4.2 - Função de Autocorrelação[2] para sons vocálicos com (a) N = 401; (b) N = 251; e (c) N = 125.

[2] Foi usada uma janela retangular em todos os casos.

Para evitar este tipo de problema pode-se processar o sinal de voz a fim de que a periodicidade torne-se mais proeminente. Este processamento também envolverá a supressão das outras características, as quais não são relevantes para a aplicação. Esta abordagem permite o uso de algoritmos muito simples para detecção de *pitch*. As técnicas que executam este tipo de operação em um sinal têm como objetivo remover os efeitos da função de transferência do trato vocal, e, portanto, trazer cada harmônica para o mesmo nível de amplitude, como no caso de um trem de impulso. Existem várias técnicas para realizar este tipo de operação, mas neste trabalho será abordada uma técnica chamada *center clipping* **[RAB 78] [SOR 95a]**, a qual mostra ter vantagens neste tipo de aplicação.

No esquema proposto por Shondi **[MAG 95]**, o sinal de voz cortado no centro é obtido através de uma transformação linear

$$\begin{aligned} y(n) &= x(n) - C_L & x(n) \geq C_L \\ y(n) &= |x(n)| - C_L & x(n) \leq -C_L \\ y(n) &= 0 & \text{caso contrario} \end{aligned}$$

onde C_L é a constante de corte para este sistema, como mostra a figura 4.3.

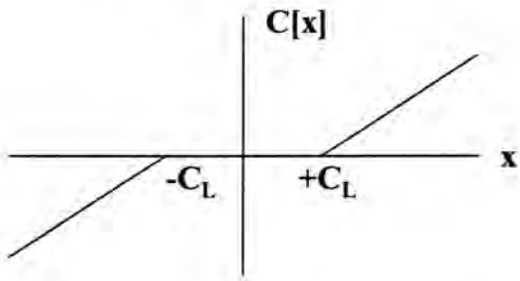

FIGURA 4.3 - Função *Center Clipping*

A operação de corte central é mostrada na figura 4.4 (b).. Para este segmento, a máxima amplitude, A_{max}, é encontrada e o nível de corte, C_L, é definido a uma porcentagem fixa de A_{max}. Na figura 4.4, pode ser visto que para as amostras acima de C_L, a saída do corte central é igual a entrada menos o nível de corte. Para as amostras abaixo do nível de corte a saída é zero.

No caso de níveis de corte muito elevados, poucos picos excederão o nível de corte e assim poucos pulsos aparecerão na saída e, portanto, poucos picos não relevantes aparecerão na função de autocorrelação. Isto é ilustrado na figura 4.4, a qual mostra a função de autocorrelação para o segmento de fala da figura 4.2. Com a diminuição dos níveis de corte, mais picos passam pela função de corte e assim a função de autocorrelação torna-se mais complexa.

A implicação deste exemplo é a mais clara indicação de periodicidade que é obtida para os maiores níveis de corte. Entretanto, é muito difícil utilizar um nível de corte muito grande, devido à variabilidade do sinal no período. Por isso, é definido um valor em função de uma porcentagem da máxima amplitude de todo o segmento. Por

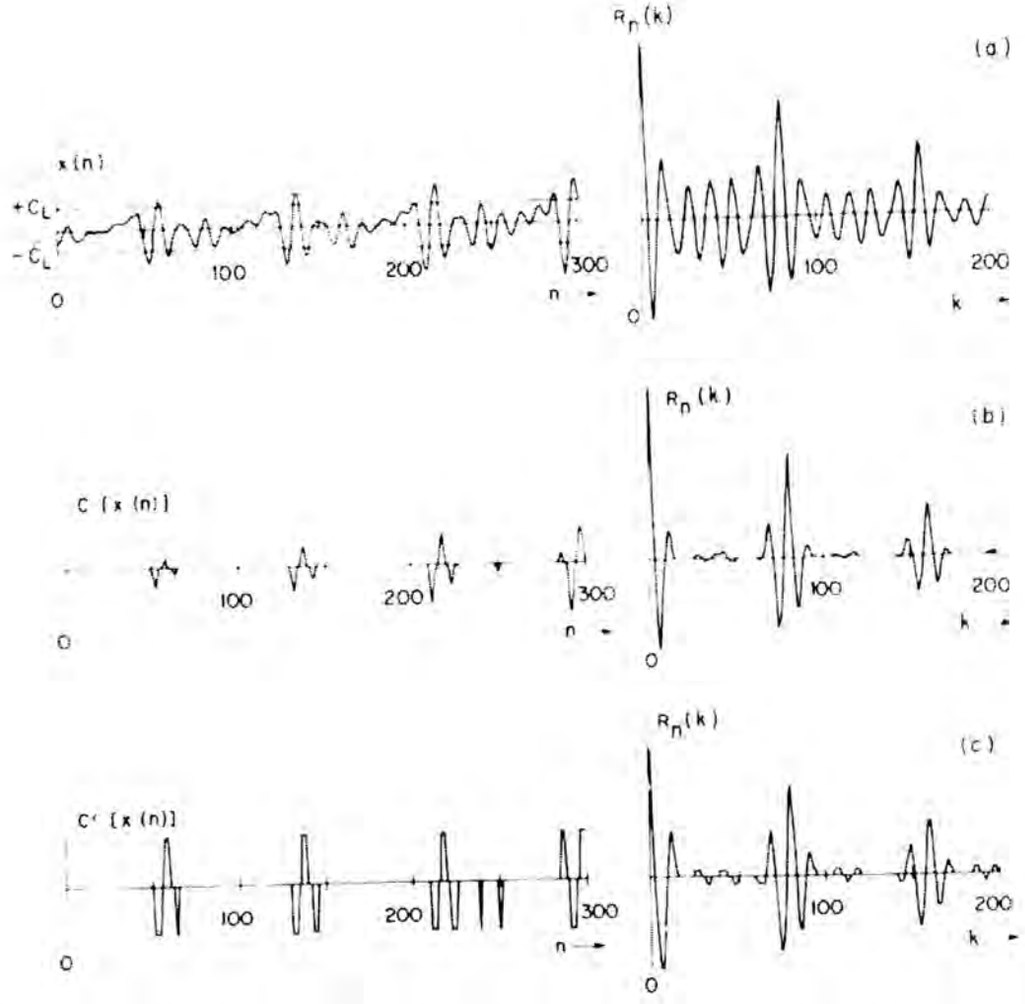

FIGURA 4.4 - Exemplo de formas de onda e função de correlação; (a) sem cortes; (b) center clipping; (c) 3-level center clipping.[3] [RAB 78]

esta razão, Shondi propôs um nível de corte de 30% da máxima amplitude. Um procedimento que pode-se utilizar para encontrar a constante de corte é dividir o segmento de fala em três e encontrar as máximas amplitudes do primeiro e últimos terços . A constante de corte é determinada tomando-se uma porcentagem do mínimo entre as duas amplitudes encontradas, como mostra a equação 4.2.

$$C_L = K * min(A1_{max}, A3_{max}) \qquad 0.6 <= K <= 0.8C \tag{4.2}$$

O problema de picos não relevantes na função de autocorrelação é em grande parte resolvido pelo corte central antes do cálculo da função de autocorrelação; mas, um problema ainda persiste no cálculo da função de autocorrelação, o qual é a quantidade de computação que é necessária. Uma pequena alteração na função de *center clipping* leva a uma simplificação na computação da função de autocorrelação

[3] Todas as funções de correlações foram normalizadas a 1.0.

sem degradação do sinal para detecção do *pitch*. Como mostra a figura 4.5, a saída do corte é +1 se $x(n) > C_L$ e -1 se $x(n) < -C_L$, caso contrário a saída é zero. Esta função é chamada de *3-level center clipping*. A figura 4.4 (c) mostra a saída do *3-level center clipping* para o segmento (a). Nota-se que os picos da periodicidade são enfatizados e os não relevantes são eliminados.

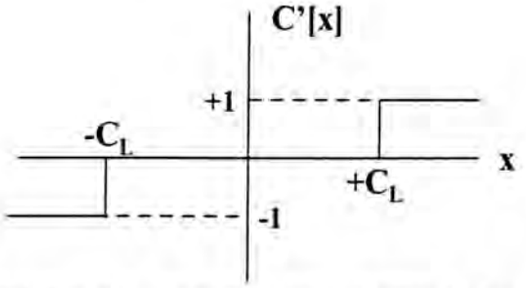

FIGURA 4.5 - Função 3-level Center Clipping

A computação da função de autocorrelação para um sinal processado pelo *3-level center clipping* é simples **[RAB 78] [SOR 95a]**.

A classificação do segmento de voz (vocálico ou não vocálico) pode ser determinada tomando-se o maior pico da função de autocorrelação, e comparando-o com um nível fixo, correspondente a 30% de R(0) de um segmento. Se a amplitude do maior pico é menor que o nível de decisão, então o segmento é classificado como não vocálico e o *pitch* é nulo; caso contrário, ele é vocálico (o *pitch* é definido como sendo o índice do segundo maior pico da autocorrelação).

4.3 Extração das Freqüências Formantes

As freqüências formantes são as freqüências ressonantes do tubo do trato vocal que produzem uma forma de onda para um determinado tipo de som. Estas freqüências dependem exclusivamente da forma e dimensão do trato vocal **[RAB 78] [SOR 95]**. Assim, as propriedades espectrais do sinal de voz variam com o tempo como o trato vocal varia, assumindo, então, que as freqüências formantes são um bom parâmetro para definição da identidade do locutor.

Para obter as freqüências formantes de um segmento de voz pode-se aplicar duas técnicas: Predição Linear **[SOR 95] [RAB 78]** e Transformada de Fourier **[ADA 96]**. A mais utilizada para estimar as formantes é a transformada de Fourier **[RUN 95]**.

Segundo **[RAB 78],** a parte baixa no tempo do cepstrum (o qual é extraído da Transformada de Fourier) corresponde primariamente ao trato vocal, pulso glotal e informação de radiação, enquanto que a parte alta no tempo corresponde à excitação, isto é, os picos no espectro correspondem essencialmente às freqüências formantes.

A partir disso, primeiramente, o segmento de sinal de voz foi filtrado (pré-ênfase) e foi feito um janelamento do tipo *Hamming*. Então, foi aplicada a FFT **[SOR 95] [RAB 78] [MOR 91] [EMB 91]** para obter o sinal no domínio freqüência. No domínio freqüência foram estimados os picos do segmento, os quais

representavam as freqüências formantes. Na figura 4.6, no primeiro gráfico (a), é mostrada uma forma de onda filtrada e feito janelamento, e no segundo gráfico (b) é mostrado o seu espectograma em freqüência onde foram sinalizadas com as barras 1, 2, 3, e 4, as quais representam as primeiras quatro formantes da locução em questão.

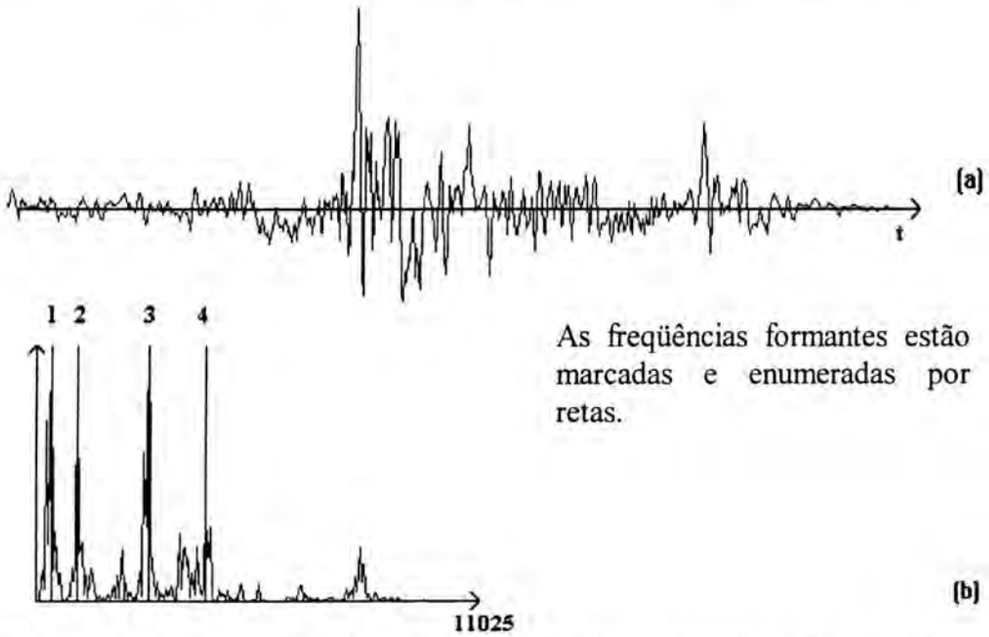

FIGURA 4.6 - Gráficos temporal (a) e espectral (b) característicos da contração entre os fonemas /ô/ e /ã/ do fim e início das palavras da locução "nono andar".

5 Classificação de Padrões

Nesta etapa, a abordagem utilizada deve ter capacidade de trabalhar com diferenças significantes na variabilidade do locutor, tal como as diferenças de pronúncia e sotaque. A complexidade desta etapa está relacionada ao tamanho da população, à taxa de fala, entre outras.

Nas seguintes seções serão apresentadas técnicas para classificação em problemas de reconhecimento de voz e locutor. As técnicas foram separadas em dois tópicos: redes neurais, onde são apresentados os modelos neurais utilizados, e métodos convencionais, onde são apresentados os métodos matemáticos convencionalmente utilizados.

5.1 Redes Neurais

A utilização do paradigma de Redes Neurais Artificiais tem crescido em processamento de fala **[MOR 91]** e tem sido somente considerada recentemente para o reconhecimento de locutor **[SOR 95a] [RUN 95]**. Nesta seção serão vistos alguns modelos e a sua utilização no processamento de voz.

5.1.1 Perceptron Multi-Camada

O modelo Perceptron Multi-camada (*Multi-Layer Perceptron* - MLPs) tem uma arquitetura comumente composta por uma camada de entrada, uma camada escondida e uma camada de saída. O seu treinamento é realizado através de um algoritmo iterativo gradiente descendente conhecido como *backpropagation* **[MOR 91]**. A figura 5.1 mostra um exemplo de arquitetura da rede MLP, que é formada por três neurônios na camada de entrada, seis neurônios na camada escondida e um neurônio na camada de saída.

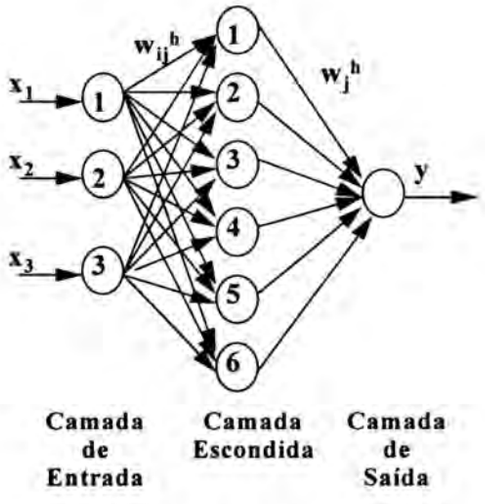

FIGURA 5.1 - *Perceptron Multi-Camada*

A maior desvantagem deste modelo de RNA é a determinação da arquitetura a ser utilizada no problema, já que a definição do número de nodos em cada camada e do número de camadas escondidas, entre outros, não é estabelecida de forma única. As formas utilizadas para se chegar a esta ótima arquitetura podem ser obtidas através de tentativa e erro, e algoritmos genéticos, entre outros.

Um problema que pode ser encontrado é a grande quantidade de locutores a serem treinados. Com um grande número de locutores a rede pode assumir que vários locutores podem ser identificados. Isto pode ser resolvido através da separação de grupos (por exemplo, pelo sexo **[BEN 91]**). O MLP pode ser aplicado ao reconhecimento de locutor **[OGL 90] [SOR 95]**.

5.1.2 Rede Neural com atraso temporal

Este modelo (*Time-Delay Neural Network* - TDNN) é um caso particular de MLP com conexões locais e restrições iguais entre alguns dos pesos. A sua arquitetura propicia uma interessante capacidade para análise de fala pois ela trata com sinal invariante no tempo e também variações na voz.

A arquitetura básica é composta por três camadas (entrada, escondida e saída), com conexões modificáveis. A camada escondida pode apresentar diversos níveis de camadas, isto é, pode-se ter mais de uma camada escondida, como mostra a figura 5.2. Cada nodo aplica uma função de transferência sigmoidal a N * (K + 1) entradas somadas e ponderadas, onde N é o número de características do vetor f(t), e K é o número de vetores com atraso no tempo.

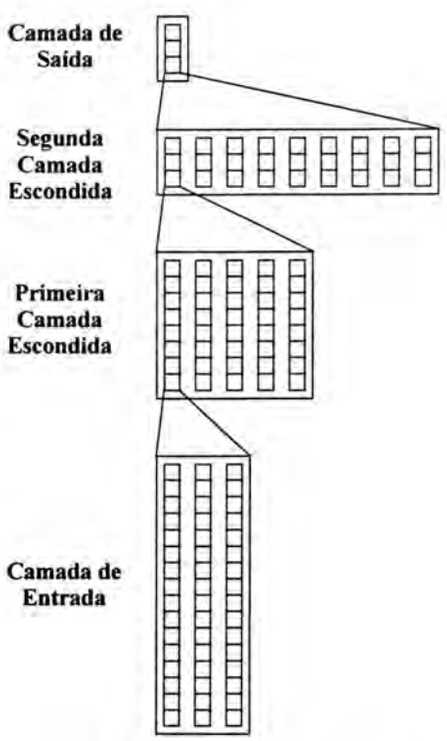

FIGURA 5.2 - Modelo de arquitetura da rede TDNN **[MAG 95]**.

No sistema de identificação do locutor independente do texto, **[BEN 91]** utiliza na camada de entrada 16 x 25 nodos os quais correspondem a 25 sucessivos quadros de tempo (aproximadamente 0.6 s) sobre o sinal. Esta entrada é um vetor de 16 características do sinal, de cada segmento, obtidas através da técnica LPC. A primeira camada escondida utiliza 12 extratores de características replicados 21 vezes, onde cada célula é conectada a 5 quadros consecutivos e é deslocado em um quadro para a direita da camada de entrada. A segunda camada escondida utiliza 10 extratores de características replicados 15 vezes, onde cada célula é conectada a 7 quadros consecutivos e é deslocado em um quadro para a direita da primeira camada escondida. A saída é totalmente conectada à última camada escondida.

O algoritmo utilizado para o treinamento deste modelo de rede neural é o *backpropagation*. Este modelo também é utilizado para classificação de fonemas **[WAI 89]**.

5.1.3 Quantização Vetorial Adaptativa

O princípio deste modelo (*Lerning Vector Quantization* - LVQ) é dividir um espaço vetorial em regiões discretas, denominadas classes, com um ou mais representantes por região **[KOH 88]**.

A arquitetura da rede LVQ é idêntica à rede de mapa de características **[KOH 82]**, exceto pelo procedimento de treinamento onde a rede LVQ é treinada com respostas "alvos". O nodo com vetor de pesos de menor distância do padrão é assumido para classificar o padrão. Se a classe da célula casa com o padrão, então o vetor de pesos é movido para mais perto do padrão, caso contrário é movido para longe da classe. Este modelo pode ser encontrado em aplicações como reconhecimento do locutor **[MAG 95]** e identificação do locutor **[BEN 91]**.

5.1.4 Rede Neural Artificial com aprendizado LMS (Least Mean-Square)

Algumas redes neurais utilizam como técnica de aprendizado o algoritmo LMS (Least Mean-Square) para minimizar o erro entre a saída obtida e a saída desejada. Os pesos são ajustados de acordo com a equação 5.1.

$$\Delta\omega_j(k) = \eta\delta(k)f_j(k), \tag{5.1}$$

com

$$\delta(k) = (x(k)^{alvo} - x(k)), \tag{5.2}$$

e

$$x(k) = F\left(\sum_{j=1}^{N}\omega_j(k)f_j(k) - \theta(k)\right) \tag{5.3}$$

onde $f_j(k)$ é a saída da iteração j no passo discreto k, $\delta(k)$ é a diferença entre a saída alvo e a saída real, e η é a taxa de aprendizagem. Tipicamente, a função de ativação $F(\cdot)$ é às vezes linear ou uma função degrau. Esta linha de iteração de atraso pode ser pensada como uma única camada e uma RNA com um único neurônio com uma função de decisão linear.

Bezerra **[BEZ 95]** implementou uma rede neural com o aprendizado LMS para o reconhecimento de locutor. A rede possuía uma arquitetura de 1120 entradas e 8 saídas e uma taxa de aprendizado menor que 0,01 ($\eta \leq 0{,}01$). Para treinamento, eram apresentadas 10 repetições do mesmo código de um dos locutores. Uma vez obtida a resposta da rede ao vetor de características da locução teste, a identificação do locutor era realizada utilizando-se o princípio da correlação máxima em função de uma matriz constante de locutores. O locutor que obtiver a maior correlação, da saída da rede com a matriz de locutores, que um determinado limiar (limiar = 4), o mesmo será escolhido como o locutor da declaração.

5.2 Métodos Convencionais

As seguintes seções apresentam três técnicas de classificação de padrões, os quais modelam a estrutura temporal da fala. Todos os três algoritmos são relativamente independentes das características selecionadas para a representação.

5.2.1 Programação dinâmica

No processo de comparação das características, o resultado a ser obtido é uma medida que represente a similaridade das características. Estas características podem ser compreendidas como coeficientes cepstrais, coeficientes LPC, formantes, entre outras.

A primeira vista, realizar uma comparação dos *frames* (intervalo finito de características) que correspondem exatamente no tempo não resultaria em uma medida correta devido ao problema que certos fatores afetam o sinal. Estes fatores podem ser definidos como: diferenciação não-linear no tempo (ritmo), freqüência (timbre), amplitude (intensidade), ambiente acústico e *stress* do locutor (alterações causados pelo cansaço físico ou doenças). Por isso, necessita-se alinhar as características do sinal a fim de que as características representem, aproximadamente, o mesmo intervalo de tempo.

Na figura 5.3 tem-se duas amostras para serem comparadas, as quais apresentam uma similaridade mas existe um deslocamento no tempo. Este deslocamento em uma comparação utilizando distância Euclidiana entre *frames* correspondentes no tempo comprometeria a qualidade do reconhecimento devido ao deslocamento entre as características a serem comparadas.

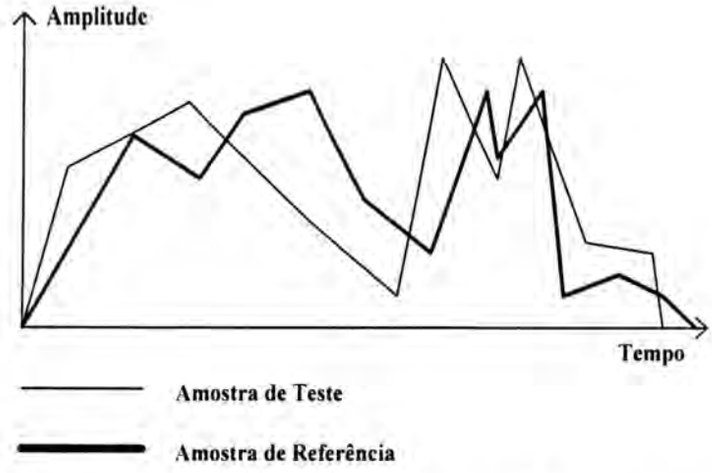

FIGURA 5.3 - Problema de alinhamento temporal de duas amostras de vozes.

Um método da programação dinâmica, chamado *Dynamic Time Warping* (DTW) **[MOR 90]**, busca por um alinhamento que minimize a distorção causada pelos efeitos da fala. Este alinhamento poderá ser medido através do cálculo da distância entre os *frames* a serem comparados. Esta distância pode ser definida como medida de similaridade. Na prática, as características são vetores multi-dimensionais e a distância entre ele é usualmente tomada como uma distância Euclidiana.

Existem diversas variações do algoritmo DTW, o qual usa diferentes métricas de distorção, caminhos permitidos, e procedimentos de buscas. A complexidade deste algoritmo é por volta de $O(N^2)$ com respeito ao número de vetores de características.

Na figura 5.4 é mostrado o alinhamento a ser realizado na amostra de teste para que a medida de similaridade represente realmente a relação entre as duas amostras.

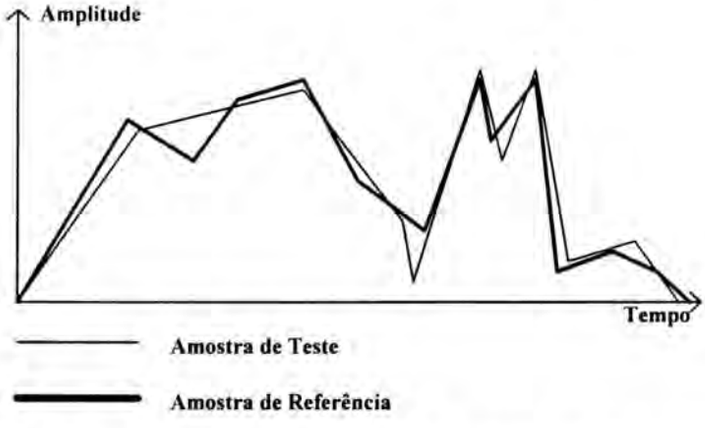

FIGURA 5.4 - Situação das amostras de sinal de voz após o alinhamento temporal gerado pelo DTW.

Basicamente o algoritmo DTW procura a amostra de entrada, ou candidata *X*, pelo caminho de alinhamento ótimo com cada amostra de referência *C* do sistema.

DTW é formulado como uma minimização sobre o tamanho da amostra usando uma medida de distorção local. Uma simples medida das somas dos quadrados pode calcular a matriz de distâncias "locais" $d_C(i,j)$ para o padrão de referência c, no ponto (i,j) de acordo com a equação 5.4.

$$d_c(i,j) = \sum_{k=0}^{N-1} \left(X(i,k) - C_c(j,k) \right)^2 \tag{5.4}$$

A amostra de entrada, X(i,k), consiste de I vetores de características, indexados em i, com N características indexadas em k. A c-ésima amostra de referência, C_c(j,k), similarmente consiste de J_c vetores de características indexados em j. O custo acumulativo, g(i,j), associado com o espaço de busca pode ser definido, usando programação dinâmica, utilizando a equação 5.5.

$$g(i,j) = \min \begin{vmatrix} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i,j-1) + d(i,j) \end{vmatrix} \tag{5.5}$$

Em cada intervalo discreto no tempo i, um custo final cumulativo é obtido. Este custo é freqüentemente dividido pelo comprimento do caminho para obter um custo médio por transição, ou vetor de característica, G, chamado valor de saída. A medida de similaridade é calculada pela equação 5.6.

$$G = g(I,J)/(\mathrm{LP}) \tag{5.6}$$

Onde LP representa o tamanho do caminho. O tamanho do caminho pode ser calculado pela:

- distância dos pontos no eixo x;
- distância dos pontos no eixo y;
- soma das distâncias ao longo do eixo x e eixo y.

Na figura 5.5 é mostrado um exemplo no qual a amostra discreta "*steam*" no eixo x, é comparada a uma amostra de referência para "*steam*" no eixo y. Neste exemplo o tamanho da amostra candidata é I = 11, a amostra de referência é J = 9, e o número de características em cada vetor é N = 2. Cada característica representa a energia aproximada das altas e baixas freqüências (em uma escala de 0 a 5) em algum intervalo discreto no tempo. Neste exemplo, a medida de distorção é a soma do valor absoluto da diferença entre cada elemento nos vetores de características da amostra candidata e referência.

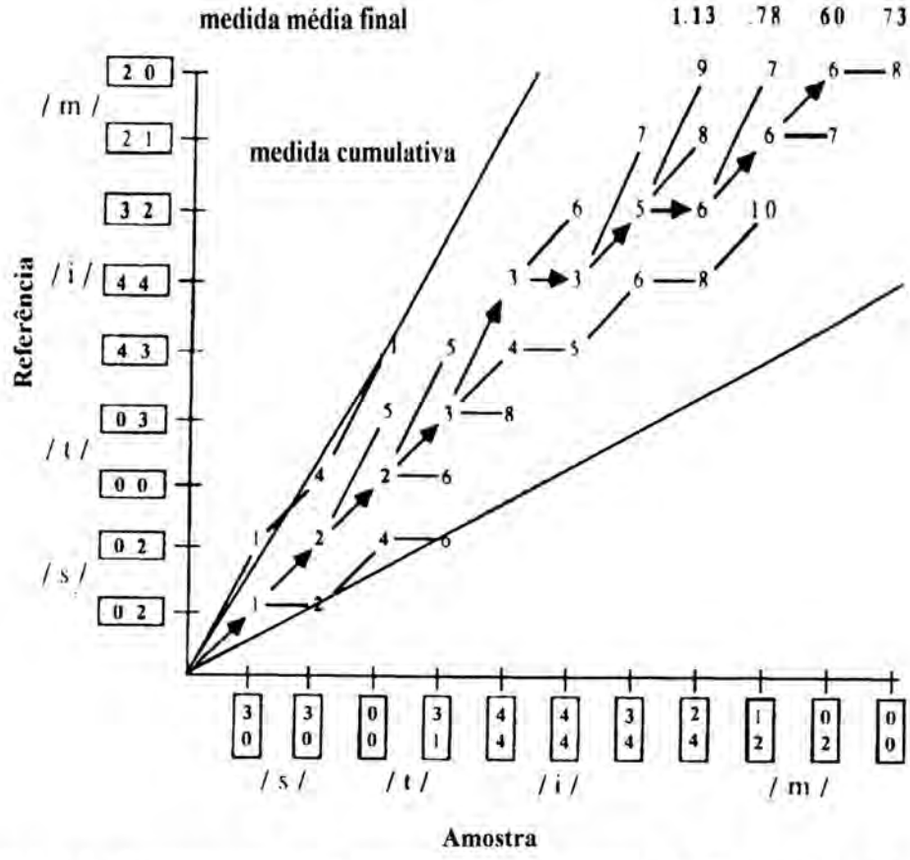

FIGURA 5.5 - Exemplo de Alinhamento temporal usando DTW entre duas amostras de vozes (palavra "*steam*")

Na figura 5.5 vários valores cumulativos são mostrados, juntamente com os caminhos de alinhamento potenciais. O caminho com o menor valor é indicado com flechas. Na parte de baixo da figura 5.5, os valores cumulativos são inicialmente definidos a distância local *d(i,0)*. Os valores cumulativos fora das extremidades do gráfico são definidos com valores infinitos. A equação 5.5 é então usada para calcular o valor cumulativo da esquerda para a direita, determinado o melhor ponto predecessor para cada caminho.

Em um sistema de reconhecimento de palavras isoladas, DTW começa e termina (ou perto do) nos pontos limites de cada amostra.

Segundo **[MOR 91]**, o algoritmo DTW pode levar em conta variações de locutores como o sotaque e a pronúncia se estas variações estiverem presentes na palavra de referência. Nos sistemas de reconhecimento de voz é muito usual o treinamento de diversas vezes a mesma palavra para definir um melhor limiar de distância. Segundo **[MOR 91]**, para um reconhecimento que seja dependente do locutor é suficiente de 2 a 5 amostras por palavra.

Os sistemas de reconhecimento de locutor, que utilizam o algoritmo DTW, apresentam a estrutura mostrada na figura 5.6. Neste sistema a base de dados é formada por amostras de palavras previamente selecionadas para o reconhecimento.

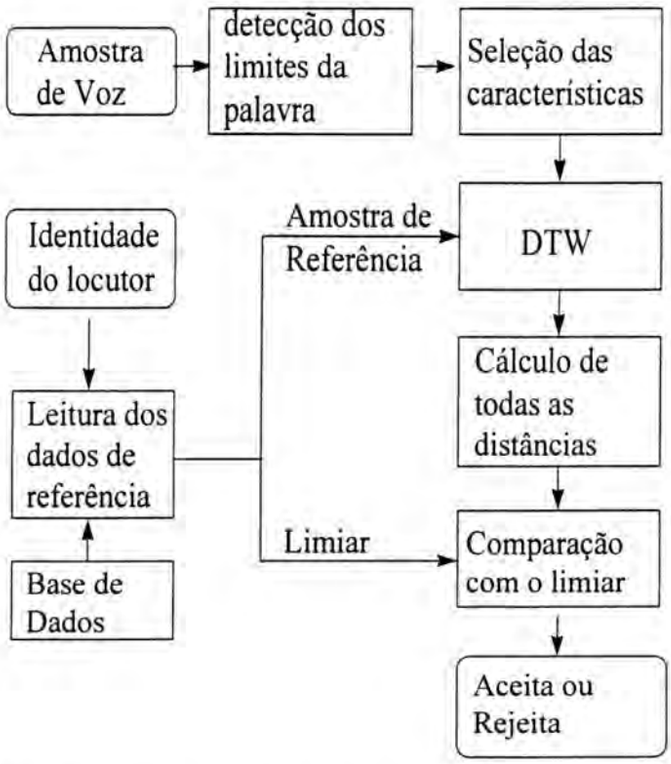

FIGURA 5.6 - Estrutura típica de um sistema baseado no algoritmo DTW.

Na fase de reconhecimento é declarada a identidade do locutor e os dados (vetor de características e limiar) referentes a este locutor são buscados na base de dados. Em seguida é feita a aquisição da amostra de voz e a extração das características do sinal. Com os vetores de características é realizado o algoritmo DTW com os vetores de referência para se obter uma medida de similaridade geral. Esta medida de similaridade é comparada com o limiar calculado para aquele locutor para decidir se aceita ou rejeita a identidade do locutor.

5.2.2 Quantização Vetorial

A Quantização Vetorial (*Vector Quantization* - VQ) é uma técnica utilizada para a redução de dados, que ainda mantém as informações necessárias para caracterizar diferentes sons.

A Quantização Vetorial trabalha um vetor de características o qual foi extraído de um sinal de voz. Esta técnica depende da criação de um "ótimo" *codebook* tal que a distorção média na substituição de qualquer vetor de características pelo codebook mais próximo é minimizada. Este *codebook* consiste de um pequeno número de vetores de características representativos os quais são suficientes para caracterizar um locutor,. por exemplo.

Para um *codebook* de tamanho Q indexado em q, o objetivo é selecionar o conjunto de vetores codebook, $\bar{c}_q$ tal que a distorção média entre vetores de treinamento é minimizada. Mais formalmente, para um conjunto finito de I vetores, $\bar{v}_i$ a equação 5.7 mostrará a seleção dos *codebooks*.

$$\|D_Q\| = \min_{\vec{c}_q}\left\{(1/I)\sum_{i=1}^{I}\min_{1\le q\le Q}\left[d(\vec{c}_q,\vec{v}_i)\right]\right\} \tag{5.7}$$

Na equação 5.7, $d(\vec{c}_q,\vec{v}_i)$ é uma medida de distorção a qual calcula a distância entre dois vetores. Exemplos de medidas de distorções incluem erro quadrático médio, erro quadrado médio ponderado, distorção preditiva linear. O *codebook* é derivado de uma amostra de treinamento estatisticamente representativa no qual o número de vetores de características I >> Q.

Um método para determinar os *codebooks* é solucionar recursivamente para $\vec{c}_q$ com Q incrementado em potências de dois **[MOR 91]**. Inicialmente Q = 2 e os dois melhores *codebooks* são selecionados os quais satisfazem a equação 5.7. O dado é então particionado de acordo com o *codebook* selecionado e o processo é repetido.

Um outro algoritmo para o projeto do *codebook* usa o algoritmo de agrupamento *k-means* **[MAK 85]**. Este algoritmo particiona o espaço de características em *Q* regiões e usa os centróides das regiões como os *codebooks*. O algoritmo começa pela seleção de *Q* vetores arbitrários como centróides. O dado de treinamento é então comparado (usando uma medida de distorção) para cada centróide a fim de determinar a região na qual ele pertence. Os centróides das regiões são então recalculados e o processo é repetido até os centróides permanecerem fixos.

O algoritmo LBG para geração de codebook **[PIR 90]** é uma generalização do algoritmo *k-means*, e procede por passos iterativos: primeiro o centróide de toda a base de dados de voz é calculada; então uma perturbação é aplicada ao vetor centróide para obter dois novos vetores, os quais são usados para os vetores de entrada de acordo com o princípio da mínima distância. Os dois centróides são iterativamente ajustados até um dado limiar de distorção mínima ser alcançado: neste estágio um "ótimo" codebook de duas dimensões é obtida. Cada centróide é novamente perturbado, e um codebook de 4 níveis é calculado. O procedimento continua até que um codebook de uma desejada dimensão é gerado.

A vantagem de VQ é que ele permite aos estágios subsequentes do reconhecedor serem muito menos complexos. Estes estágios necessitam somente obter variações entre os Q vetores, antes do que todos os possíveis vetores. Os relacionamentos entre estes Q vetores pode ser pré-calculado, economizando tempo de processamento. Estes relacionamentos podem incluir medidas de distorções como as descritas nesta seção, ou a probabilidade que o vetor de codebook $\vec{c}_{q_i}$, será seguido por um outro $\vec{c}_{q_j}$.

A figura 5.7 mostra um método usando um *codebook* para vetores de características consistindo de características instantâneas e transacionais calculadas para os coeficientes cepstrais e freqüência fundamental (Matsui em **[FUR 94]**). Desde que a freqüência fundamental não pode ser extraída de um som não-vocálico **[RAB 75]**, existe um *codebook* para som vocálico e outro para não-vocálico para cada locutor.

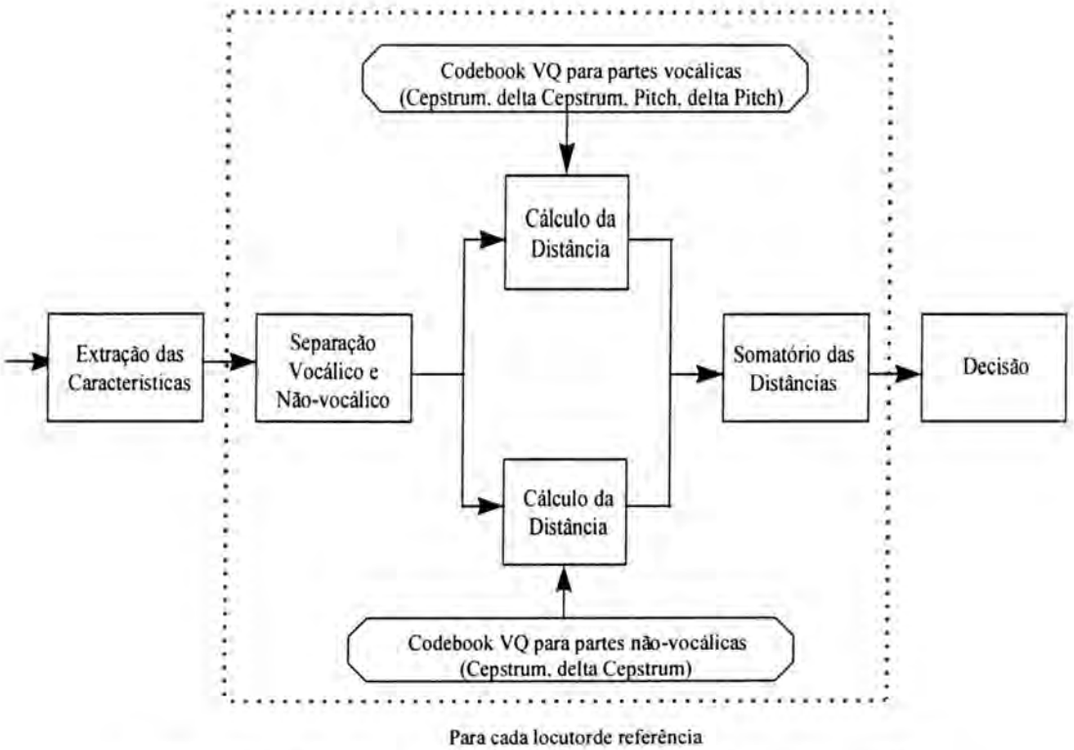

FIGURA 5.7 - Uma estrutura de um método baseado em VQ

5.2.3 Cadeias de Markov

A técnica das Cadeias de Markov (*Hidden Markov Models* - HMM) foi apresentada em um artigo por Baum **[MOR 91]**, que propôs este modelo como um método estatístico de estimação de funções probabilísticas de uma cadeia de Markov. Essencialmente, Hidden Markov Models (HMM) é um método para modelagem de sistemas com discretos "processos" de curto espaço de tempo e transições entre eles.

Um HMM pode ser definido como uma máquina de estados finita onde as transições entre os estados são dependentes da ocorrência de algum "símbolo". Associado com cada transição de estado há uma distribuição de probabilidade de saída a qual descreve a probabilidade com o que um símbolo ocorrerá durante a transição, e uma probabilidade de transição indicando a probabilidade desta transição

Portanto, o HMM é um processo Markov porque a probabilidade de estar em um estado particular no tempo $t + 1$, dada a seqüência de estados antes do tempo t, depende somente do estado no tempo t. Uma ilustração de um HMM típico é mostrado na figura 5.8, onde as setas representam as transições (probabilidades) entre os estados.

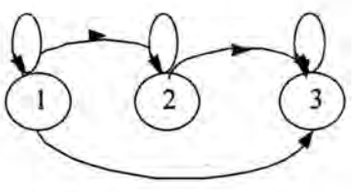

FIGURA 5.8 - Representação de um HMM.

O modelo HMM descreve um processo estocástico, o qual produz uma seqüência de eventos ou símbolos observados. O HMM é chamado um modelo de Markov escondido porque há um processo estocástico elementar que não é observável, mas afeta a seqüência observada dos eventos. A seqüência de observações pode ser descrita usando a notação:

$$O_1^T = O_1, O_2, \ldots, O_t, \ldots, O_T,$$

onde o processo foi observado para os T passos discretos no tempo do tempo t = 1 a T. As observações podem ser qualquer um dos K símbolos, v_k.. No reconhecimento da fala, estes símbolos podem ser *codebooks* VQ calculadas em intervalos regulares, ou fonemas.

Na figura 5.8, se o estado 1 representasse as letras "A" e "a", o estado 2 a letra "a", o estado 3 a letra "B", a sentença "AaaaB" seria produzida pelas seguintes seqüências de estados:

- $\rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 3$,
- $\rightarrow 1 \rightarrow 1 \rightarrow 1 \rightarrow 2 \rightarrow 3$,
- $\rightarrow 1 \rightarrow 1 \rightarrow 2 \rightarrow 2 \rightarrow 3$,

o que mostra porque chama-se modelo escondido de Markov, pois não se sabe explicitamente qual das três seqüências seria usada para formar a sentença desejada..

Um algoritmo conhecido como *forward/backward* (ou Baum-Welch) encontra um conjunto de probabilidades de transições de estados e distribuições de saída de símbolos para cada HMM. Este algoritmo gradiente descendente usa dados de treinamento para iterativamente refina um conjunto inicial (possivelmente aleatório) de parâmetros do modelo, tal que o HMM gera os padrões do conjunto de treinamento.

Depois deste estágio inicial de treinamento, uma palavra ou uma sentença a ser reconhecida é falada, e as medições da fala são feitas para reduzir a amostra de voz em uma seqüência de símbolos. No caso de reconhecimento de palavras isoladas, o algoritmo *forward* calcula a probabilidade de que cada modelo de palavra produz uma seqüência observada de símbolos - o modelo com a maior probabilidade representa a palavra reconhecida.

6 Reconhecimento de Locutor

O reconhecimento de locutor (RL) é dividido em duas tarefas denominadas identificação de locutor (identificar o locutor em um grupo de locutores) e verificação de locutor (verificar se o locutor é mesmo quem diz ser).

Um modelo genérico utilizado para o reconhecimento de locutor é mostrado na figura 6.1. Nesse modelo são extraídas as características do sinal de voz e enviadas para um classificador, o qual realizará a decisão final de identificação ou verificação. do locutor.

Além disso, os sistemas de reconhecimento de locutor podem ser dependentes ou independentes do texto. Sistemas de RL dependentes do texto requerem que o locutor pronuncie uma frase ou uma dada senha pré-determinada, enquanto que o sistema independente do texto não requer a exigência do caso anterior.

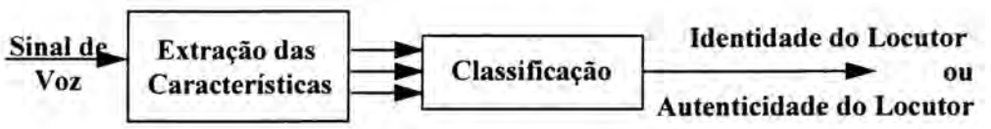

FIGURA 6.1 - Representação geral do problema de reconhecimento de locutor

6.1 Verificação de Locutor

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

Para realizar esta decisão, um conjunto de características projetadas para reter a informação essencial sobre a identidade do locutor é extraída de uma ou mais amostras (pronúncias) de voz do locutor. Feito isso, tais características são comparadas (freqüentemente é realizada alguma medida de comparação altamente não-linear) a um conjunto de padrões de referência.

Assim, para a verificação de locutor, somente uma simples comparação entre o conjunto (ou conjuntos) de medidas e o padrão de referência se faz necessário para obter a decisão final de aceitar ou rejeitar a identidade afirmada. Geralmente, a medida de distância entre as medidas dadas e a distribuição de referência armazenada é computada. Baseado nos custos relativos de fazer dois possíveis tipos de erros (isto é, verificar um impostor, ou rejeitar um locutor verdadeiro) um apropriado *threshold* (limiar) é definido na função de distância.

6.2 Identificação de Locutor

O problema da identificação de locutor difere significativamente do problema de verificação de locutor. Neste caso, o sistema é requisitado a fazer uma identificação absoluta entre os N locutores na população de usuários. Assim, em vez de uma única comparação entre um conjunto de medidas e um padrão de referência armazenado, N comparações completas são necessárias. A regra de decisão para tais sistemas é essencialmente da forma:

$$\textit{escolha locutor } i \textit{ tal que } p_i(x) > p_j(x),$$
$$\textit{onde } j = 1,2,...N,\ j \neq i$$

isto é, escolher o locutor com a mínima probabilidade absoluta de erro. Neste caso parece plausível **[MOR 91]** que, como a população de usuários é muito grande, a probabilidade de erro deve tender a um desde que um número infinito de distribuições não pode permanecer distinta em um espaço de parâmetros finito - isto é, torna-se gradualmente provável que dois ou mais locutores no conjunto terão distribuições de medidas extremamente próximas uma da outra. Sob estas circunstâncias, uma identificação fidedigna do locutor torna-se impossível.

Este tipo de reconhecimento pode ocorrer de duas formas:

- **conjunto-aberto**: o locutor pode não estar entre a população. Para isso, costuma-se utilizar um limiar como na verificação do locutor para determinar se um locutor está fora do conjunto.
- **conjunto-fechado**: sabe-se *a priori* que o locutor é um membro da população.

6.3 Parâmetros Identificadores de Locutor

As diferenças da fala, segundo **[DOD 71]**, estão ligadas tanto aos aspectos fisiológicos quanto aos aspectos comportamentais do locutor. As diferenças anatômicas se relacionam a variações no tamanho e na forma dos órgãos do trato vocal, como mostra a figura 6.2, onde a voz é apresentada como sendo resultado do ar expelido do pulmões através do trato vocal, o que mostra claramente que a voz depende da anatomia de cada locutor. Estas diferenças são refletidas no *pitch*, na magnitude espectral, nas freqüências das formantes, e na nasalidade do som, entre outras.

As diferenças comportamentais apresentadas na fala, por serem mais subjetivas, são mais difíceis de especificar. Segundo Wolf em **[PRA 95]**, estas resultam de diferenças nos padrões de comandos neurais para separar os articuladores, sendo estes comandos aprendidos individualmente por cada locutor, o que resulta em variações na dinâmica do trato vocal. Isto afeta a taxa de transição das formantes e a coarticulação. Além disso, pode-se observar diferenças de dialeto, estilo de fala, o estado emocional e a forma de pronúncia individual.

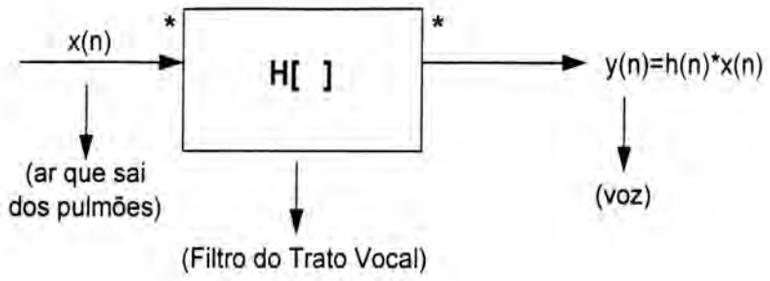

FIGURA 6.2 - A voz como resultado da combinação do ar com o trato vocal

Para que o processo de identificação do locutor seja bem sucedido, é de fundamental importância, segundo **[PRA 95]**, que o conjunto de parâmetros escolhidos para caracterizar cada indivíduo seja o mais apropriado possível. Wolf mencionado em **[PRA 95]** afirma que o conjunto ideal de características da fala deve conter parâmetros que satisfaçam os seguintes critérios:

- Representem eficientemente as informações particulares do locutor;
- Sejam facilmente medidos e extraídos;
- Sejam estáveis, ou seja, de pouca variação;
- Ocorram de forma natural e com freqüência na fala;
- Não sejam suscetíveis a imitações;
- Sejam resistentes a ruídos externos, de tal forma que não sejam mascarados pela presença do ruído no sinal de voz.

As características apresentadas por Wolf mencionado em **[PRA 95]** definem que os parâmetros ideais da fala devem expressar as características inerentes a cada indivíduo, além de destacarem as diferenças entre os diversos locutores. Certamente, o conjunto ideal de parâmetros deve ser capaz de representar de forma não ambígua os padrões relacionados a cada pessoa. Outra vantagem deste conjunto de parâmetros está relacionada ao processo de identificação do locutor, pois este processo consiste em comparar um padrão qualquer aos padrões já conhecidos pelo sistema de reconhecimento.

Os sistemas de reconhecimento de locutor buscam altas taxas de reconhecimento (aproximadamente 97% no mínimo **[RAB 75]** **[MOR 91]**) para que o mesmo possa ser utilizado em ambientes reais de segurança. A margem de erro tanto no reconhecimento do verdadeiro locutor e do falso locutor definirá a qualidade do sistema. Isto é, não poderá acontecer a aceitação de um falso locutor assim como a rejeição de um locutor verdadeiro.

7 Implementação utilizando Redes Neurais Artificiais

Neste capítulo serão apresentadas as técnicas utilizadas e informações relativas ao trabalho prático efetuado no problema de reconhecimento de locutor utilizando redes neurais artificiais.

7.1 Base de Dados

A base de dados utilizada na implementação a ser abordada neste capítulo consiste de comandos para elevadores, no qual são amostradas frases pré-definidas:

- primeiro andar;
- segundo andar;
- terceiro andar;
- quarto andar;
- quinto andar;
- sexto andar;
- sétimo andar;
- oitavo andar;
- nono andar;
- décimo andar;

Estas frases foram amostradas sem alguma restrição de pausa entre as palavras. Foram amostrado em um total de 10 locutores masculinos com idade entre 18 e 24 anos. No total foram obtidos 100 amostras de vozes.

As locuções utilizadas neste trabalho foram amostradas a 22 kHz (22050Hz) com codificação em 16 bits, e microfone no modo mono. A aquisição dos dados foi através de uma placa de som (*Sound Blaster 16*) em um microcomputador.

As amostras de vozes foram pré-processadas a fim de extrair os segmentos de voz correspondente ao fim da primeira palavra e o início da segunda palavra. Estes segmentos podem ser encontrados utilizando o algoritmo de detecção de limites mas alterando-se a variável que controla o intervalo de tempo sem fala (Tmin).

7.2 Pré-processamento do sinal

Após a aquisição do sinal de voz, o sinal passou por um filtro de pré-ênfase, que aumenta a amplitude do sinal de voz nas freqüências mais altas. Este filtro foi implementado segundo a função de transferência mostrada na equação 4.1, com a=0.95.

Depois da filtragem, foi realizada a detecção de limites, utilizando o algoritmo de Rabiner **[RAB 78]** apresentado na seção 4.1.2.

7.3 Extração das Características

O sinal foi dividido em segmentos de 10ms com sobreposição (*overlaping*) de 5ms para cobrir as alterações do trato vocal **[RAB 78]**.

Os parâmetros utilizados neste trabalho são:

1. *pitch* (freqüência fundamental);

2. 4 primeiras freqüências formantes;

3. 20 coeficientes cepstrais extraídos do algoritmo LPC.

Na implementação do algoritmo de detecção do *pitch*, foi aplicada uma janela retangular para cada segmento a fim de não alterar os componentes característicos do segmento vocal. Após o janelamento, foi aplicada a técnica de *center clipping* para eliminar as amostras não relevantes para o cálculo de autocorrelação. A constante de corte foi calculada em função de 70% do maior pico (exceto R[0]).

Em seguida ao corte, foi calculada a função de autocorrelação. Para facilitar a detecção de picos foi realizado mais um *center clipping* (constante de corte em 40%) para eliminar as variações do trato vocal.

Estas fases são mostradas na figura 7.1, onde percebe-se que a saída da autocorrelação de um segmento periódico é de fácil localização da periodicidade (*pitch*).

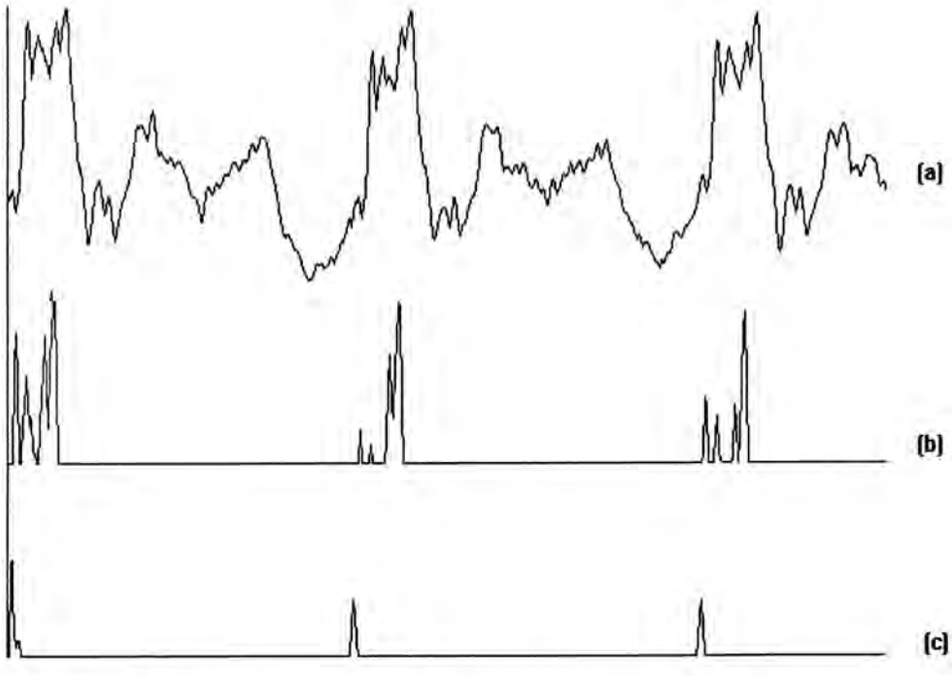

FIGURA 7.1 - Gráficos do processo de extração de *pitch*; (a) forma de onda original (b) forma de onda com *center clipping*; (c) autocorrelação do segmento com *center clipping*.

Na implementação do algoritmo de detecção das freqüências formantes, foi aplicada uma janela de *Hamming*, para atenuar o processo de corte e sobreposição dos segmentos de fala. Após este processo, foi realizada a busca dos maiores picos de freqüências do espectro do sinal, baseado no algoritmo proposto por **[RAB 78]**.

Após o processo de extração das características, prosseguiu-se para a fase que é responsável pela comparação das características dos locutores, como será apresentado na seção 7.4.

Após passar o sinal por uma janela de Hamming, foram calculados 12 coeficientes LPC deste segmento de voz. Destes 12 coeficientes foram calculados 20 coeficientes cepstrais utilizando a equação 3.46.

7.4 Classificação do Locutor

O problema do reconhecimento de locutor, como visto no capítulo 6, divide-se em dois tipos: identificação e verificação de locutor. Neste trabalho, o sistema implementado realiza a identificação do locutor, onde nenhuma identidade é declarada e o sistema se encarrega de reconhecer a identidade do locutor.

Após extraídos os parâmetros característicos do locutor, é realizada uma classificação do locutor em um conjunto de N locutores. Para realizar este processo, foi utilizado, neste trabalho, um modelo de Rede Neural Artificial (RNA) **[MOR 91]**.

As RNAs obtêm um desempenho desejável para problemas que envolvem dados discretizados. Sua aplicação na área de reconhecimento de voz é de grande utilização **[MOR 91] [SOR 95]**. Os modelos de RNAs mais utilizados já foram apresentados na seção 5.1.

Dentre os modelos discutidos nas seções 5.1.1 a 5.1.3, neste trabalho foi utilizado o *Multi Layer Perceptron* (MLP), devido à sua facilidade de implementação, desempenho e taxa de aprendizado **[MOR 91].**

7.4.1 Arquitetura da Rede Neural MLP

Em uma primeira abordagem foi construída uma rede MLP com 45 entradas para as características obtidas do sinal de voz. As características referentes a este modelo são o *pitch* e as quatro primeiras formantes. Estas características eram extraídas de 9 segmentos do sinal de voz, totalizando as 45 entradas.

Na segunda abordagem foram utilizados três tipos de características, isto é, além do *pitch* e das quatro freqüências formantes, foram utilizados os 20 coeficientes cepstrais. Assim, no que diz respeito à camada de entrada, foram utilizadas 225 entradas.

Todos os padrões de entrada foram normalizados dentro do intervalo [-1, +1]. Esta normalização foi realizada por *feature* **[SOR 95]**, isto é, atribuindo à máxima magnitude o valor +1 e, -1 no caso contrário. Esta normalização tem por objetivo proporcionar uma convergência mais rápida.

Para cada locutor foi treinada uma rede neural específica, isto é, na fase de treinamento a rede neural seria treinada para reconhecer somente um locutor

específico e rejeitaria os demais. A taxa de aprendizado para o algoritmo *backpropagation* foi 0,015.

A nível de treinamento, a saída desejada na rede neural seria +1 para o locutor que a rede deve reconhecer e -1 para os demais locutores. Portanto, a camada de saída da rede neural tem somente um neurônio.

No caso do locutor não pertencer à população de N locutores , o mesmo não seria reconhecido em nenhuma das redes neurais.

7.5 Treinamento e Testes da Rede Neural MLP

Para o treinamento foram utilizadas 1 camada oculta com 5, 10 ou 15 neurônios e 1 camada de saída com 2 neurônios para a primeira e segunda abordagem.

O processo de treinamento e testes foi dividido em três experimentações:

1. modelo de rede neural utilizando como padrão a ser reconhecido duas características: *pitch* e freqüências formantes;
2. modelo de rede neural utilizando como padrão a ser reconhecido três características: *pitch*, freqüências formantese coeficientes cepstrais;
3. modelo de rede neural utilizando como padrão a ser reconhecido três características descritas no experimento dois mas sobre toda a amostra.

Para o treinamento e validação da Rede Neural MLP, as amostras foram divididas em dois grupos de 5 amostras para cada fase, como mostra a tabela 7.1

TABELA 7.A - Grupos de amostras de vozes para a fase de treinamento e testes.

Fase de Treinamento	**Fase de Testes**
primeiro andar	segundo andar
terceiro andar	quarto andar
quinto andar	sexto andar
sétimo andar	oitavo andar
nono andar	décimo andar

O treinamento foi realizado individualmente para cada locutor, isto é, existe um modelo de rede neural para cada locutor. Para cada modelo existem 10 amostras de voz do locutor verdadeiro e 90 amostras para serem utilizadas como locutores falsos. Sendo que para cada fase foram utilizadas amostras diferentes, isto é, na fase de treinamento foram utilizadas 5 amostras do locutor verdadeiro e 45 amostras como locutor falso, e na fase de teste eram utilizadas as demais amostras.

Para a demonstração dos resultados obtidos no sistema de reconhecimento foram avaliados dois aspectos:

- **Taxa de acerto**: percentual representativo de acertos do locutor verdadeiro. Esta taxa é em função do número de amostras a serem reconhecidas (neste trabalho foram consideradas 5 amostras para cada fase);

- **Taxa de rejeição**: percentual representativo de não reconhecimentos de locutores falsos. Esta taxa é em função do número de amostras a não serem reconhecidas (neste trabalho foram consideradas 45 amostras para cada fase);

O erro é calculado em função da distância Euclidiana entre a saída desejada da rede e a retornada pela rede, normalizada pelo número de amostras.

7.5.1 Experimentação com duas características

Na fase de treinamento, foram realizadas 1500 épocas para atingir um erro médio de $2 * 10^{-4}$. Este erro mostra a grande separação das características do locutor no espaço de características para o treinamento.

Na fase de treinamento foram realizadas avaliações das taxas de acerto e rejeição do locutor , conforme mostra a tabela 7.2.

TABELA 7.B - Taxas de reconhecimento da rede neural.

	Fase de Treinamento		**Fase de Testes**	
Camada Escondida	**Taxa de Acerto**	**Taxa de Rejeição**	**Taxa de Acerto**	**Taxa de Rejeição**
5 neurônios	100%	100%	58%	89%
10 neurônios	100%	100%	62%	92%

De acordo com a tabela 7.2, percebe-se que, para aquele conjunto de padrões de treinamento, a rede se comportou satisfatoriamente para o reconhecimento do locutor.

Com os valores obtidos na fase de treinamento, partiu-se para a fase de testes. Nesta fase, foram fornecidos para a entrada da rede as cinco locuções que não tinham sido utilizadas na fase de treinamento. Como mostra a tabela 7.2, a taxa de reconhecimento não é muito confiável para o problema proposto.

7.5.2 Experimentação com três características

Nesta experimentação foi mantido o processo de treinamento com 5 locutores como padrão de treinamento foram utilizadas amostras de 5 locuções para cada locutor. Foram utilizadas 1000 iterações independente do erro.

A tabela 7.3 apresenta os resultados obtidos utilizando 10 neurônios para a camada escondida e 2 neurônios para a camada de saída.

TABELA 7.C - Taxas de reconhecimento do modelo MLP de três camadas, 225X10X2.

Locutor	Fase de Treinamento		Fase de Testes		Erro
	Taxa de Acerto	Taxa de Rejeição	Taxa de Acerto	Taxa de Rejeição	
1	100%	100%	80%	98%	0.0014
2	100%	100%	80%	92%	0.0902
3	100%	100%	60%	93%	0.0033
4	100%	100%	60%	93%	0.0006
5	100%	100%	60%	96%	0.2689
6	100%	100%	100%	96%	0.0092
7	100%	100%	100%	92%	0.0010
8	100%	100%	100%	93%	0.0015
9	100%	100%	20%	93%	0.0017
10	100%	100%	80%	93%	0.7149
Totais	**100%**	**100%**	**74%**	**94%**	

Na tabela 7.3 observa-se que alguns locutores obtiveram altos índices de reconhecimento, o que pode ser resultado de uma boa amostragem ou uma grande distância das suas características, no espaço de características, dos demais locutores. A taxa de rejeição apresenta um boa percentagem para o sistema pois o mesmo garante que não haverá, na grande maioria reconhecimento de locutores falsos.

Como mostra a tabela 7.4, a taxa de reconhecimento não houve alterações, mas a taxa de rejeição houve uma pequena alteração. Isto mostra, que com uma arquitetura de 10 neurônios na camada oculta e dois neurônios na camada de saída são suficientes para esta abordagem. Além disto, a necessidade de processamento seria maior pois tem-se 50% a mais de neurônios na camada oculta, o que não seria justificável já que as taxas foram as mesmas. No que diz respeito à taxa de reconhecimento, a taxa de 75% ainda não demonstra a total eficiência do sistema.

TABELA 7.D - Taxas de reconhecimento do o modelo MLP de três camadas, 225X15X2.

Locutor	Fase de Treinamento		Fase de Testes		Erro
	Taxa de Acerto	Taxa de Rejeição	Taxa de Acerto	Taxa de Rejeição	
1	100%	100%	100%	100%	0.0003
2	100%	100%	100%	98%	0.0012
3	100%	100%	80%	88%	0.0017
4	100%	100%	40%	96%	0.0016
5	100%	100%	60%	93%	0.0004
6	100%	100%	20%	100%	0.0015
7	100%	100%	80%	98%	0.0027
8	100%	100%	100%	100%	0.0013
9	100%	100%	60%	93%	0.1437
10	100%	100%	80%	87%	0.0014
Totais	**100%**	**100%**	**74%**	**95%**	

Além destes foram realizados testes utilizando uma arquitetura com 15 neurônios na camada oculta e 3 neurônios na cama de saída, mas não houve alterações significativas nas taxas de reconhecimento.

7.5.3 Experimentação com três características sobre toda a amostra

Para a elaboração desta experimentação foram utilizadas algumas conclusões sobre as duas experimentações anteriores. Os parâmetros alterados nesta experimentação serão mostrados nesta seção além dos resultados obtidos.

Nesta experimentação, a base de dados é diferente. Foi utilizada apenas uma palavra como amostra: *computer*. Esta palavra contém muitos sons vocálicos, o que torna bem representativa as características do locutor. As amostras foram adquiridas utilizando o mesmo *hardware* mas com a taxa de amostragem de 11 kHz. Segundo a teoria de Nyquist **[RAB 75]**, com esta taxa de amostragem consegue-se obter freqüências até 5,5 kHz, o que já é suficiente para as características pois as principais freqüências encontram-se concentradas em freqüências até 4 kHz **[RAB 75]**.

Foram amostrados apenas 4 locutores (3 masculinos e 1 feminino), sendo que para cada locutor há 10 repetições da mesma palavra. Assim, produzindo uma base de dados de 40 locuções da palavra *computer*.

Foi observado que nos testes anteriores houve uma saturação no treinamento devido ao grande número de locutores que seriam considerados falsos. Por isso, nesta

experimentação foram realizados treinamentos onde o número de amostras dos locutores verdadeiros eram iguais as dos locutores falsos.

A tabela 7.5 mostra as taxas de reconhecimento para esta experimentação. Foi utilizada uma arquitetura de 10 neurônios na camada escondida e 2 neurônios na camada de saída.

TABELA 7.E - Taxas de Reconhecimento da rede MLP sobre a palavra inteira.

Locutor	Fase de Treinamento		Fase de Testes	
	Taxa de Acerto	**Taxa de Rejeição**	**Taxa de Acerto**	**Taxa de Rejeição**
1	100%	100%	100%	100%
2	100%	100%	50%	100%
3	100%	80%	100%	75%
4	100%	100%	100%	100%
Totais	**100%**	**95%**	**88%**	**94%**

Na tabela 7.5 pode-se perceber que houve uma melhora na taxa de reconhecimento. Este aumento pode-se atribuir a fatores como:

- maior número de sons vocálicos;
- menor número de locutores;
- distribuição de amostras nos treinamento da rede.

8 Implementação Utilizando Métodos Convencionais

Neste capítulo será apresentada a implementação do problema de reconhecimento de locutor, a qual utilizou-se de métodos matemáticos convencionais para a identificação dos locutores.

8.1 Base de dados

Neste experimento foi utilizado como entrada amostras de vozes da base de dados SR4X do *toolkit* do CSLU (*Center for Spoken Language Understanding*) do Oregon Graduate Institute. Esta base de dados é composta por 36 locutores onde cada um fala 11 palavras pré-determinadas 6 vezes. As palavras desta base são :

⇒ startrek	⇒ nebula	⇒ sungeeta
⇒ supernova	⇒ processing	⇒ computer
⇒ tektronix	⇒ singularity	⇒ 71523
⇒ generation	⇒ abracadabra	

Estas locuções foram gravadas utilizando 4 diferentes canais: telefone comercial, telefone residencial, telefone com microfone de carbono, e um telefone viva-voz. Estas amostras foram adquiridas com uma taxa de amostragem de 10 kHz e tamanho de palavra de 16 bits.

Neste trabalho foram utilizadas somente palavras que foram extraídas do conjunto de palavras gravadas por um telefone comercial para o treinamento. Da base SR4X foram escolhidas somente palavras que satisfizessem 95% de boa qualidade de aquisição.

No total foram utilizados 216 locuções para o conjunto de locutores verdadeiros e 7560 locuções para o conjunto dos impostores.

8.2 Característica do sinal utilizado

Há muitos modos no qual o sinal original pode ser modificado, tal que quando o resultado é reproduzido, ele ainda é prontamente reconhecido pelo ouvido humano como o original. Isto aumenta a questão de quais aspectos do sinal realmente carregam a informação necessária para o reconhecimento do locutor.

Um dos primeiros passos envolvidos no reconhecimento de locutor é remover a distorção causada pelo meio físico, no qual foi realizada a amostragem do sinal. O processo de remoção de distorção nem sempre consegue extrair a distorção do sinal.

Após a remoção da distorção do sinal, foram extraídos os intervalos de sinais, os quais continham silêncio. Esta remoção é necessário para reduzir a quantidade de

processamento e principalmente não prejudicar a própria fase de classificação do sinal.

Com o sinal puro, pode-se extrair realmente as características do sinal para a classificação. Diferente da implementação proposta no capítulo 6, o reconhecimento realizado neste sistema somente utilizará como característica do locutor os coeficientes cepstrais do sinal (20 coeficientes), mais a energia para cada segmento **[RAB 75]**. Estes coeficientes cepstrais foram obtidos através do cálculo do LPC. Estes parâmetros foram obtidos utilizando janelas *Hamming* com intervalo de amostras de 20ms com deslocamentos de 10ms.

Ainda nesta fase foi realizada uma seleção dos vetores mais "característicos" extraídos do sinal de voz. A partir disso, para cada conjunto de vetores foram calculados a média e desvio padrão da energia de todos vetores. Com estes dados, foram extraídos os conjuntos de vetores como aqueles que energia não satisfaz a condição: $E >= m - d$, onde E é a energia, m é a média e d é o desvio padrão. Os vetores que satisfazem a condição imposta podem ser definidos como aqueles que apresentam uma maior concentração de energia, isto é, caracterizam o som vocálico. A figura 8.1 apresenta um gráfico, onde a parte hachurada representa os vetores retirados do processamento.

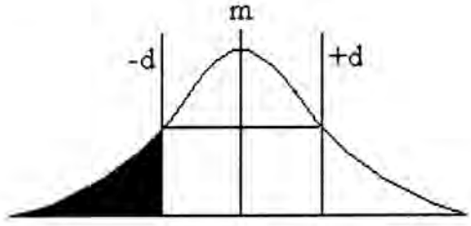

FIGURA 8.1 - A parte hachurada da curva representa os vetores a serem descartados.

8.3 Classificação dos locutores

Nesta fase do reconhecimento foram utilizadas duas técnicas convencionais para classificar o locutor:

- Dynamic Time Warping (DTW);
- Vector Quantization.

As técnicas utilizadas fornecerão valores que serão utilizados pelo módulo de decisão para definir a autenticidade do locutor.

Estas duas técnicas trabalharam as características dos locutores diferentemente uma da outra. A descrição da implementação de cada uma está descrita nas seções a seguir.

8.3.1 Programação Dinâmica

Nesta fase o conjunto de vetores do locutor a ser identificado é analisado com um conjunto de vetores de referência usando o algoritmo DTW. Entretanto, devido ao problema da diferença do número de vetores a serem comparados, foi utilizado um número fixo de vetores. A forma de fixar o número de vetores não era simplesmente descartar os mesmos, mas realizar uma média sobre os vetores.

Ao extrair os coeficientes cepstrais dos segmentos de um sinal, obtinha-se um número irregular de segmentos devido á taxa de fala. Deste conjunto de segementos foram extraídos somente *n* segmentos, utilizando-se da média de vetores para comprimir (no caso em que número de segmentos seja maior que o desejado) e a duplicação de vetores para expandir (no caso em que o número de segmentos seja menor que o desejado). Esta abordagem tenta resolver a diferença no alinhamento das locuções para o cálculo do DTW. É possível haver uma queda na representação dos vetores mas que pode ser considerada menor do que o erro causado pela diferença no alinhamento.

Na fase de treinamento, este algoritmo para extração de um número fixo de segmentos é aplicado a 3 locuções para cada locutor e palavra. Para definir qual das três locuções seria o padrão de referência foi realizada uma validação cruzada, isto é, eram calculadas as distâncias entre cada locução e as demais. A locução escolhida era a que obtivesse a menor soma das distâncias das demais locuções.

Na fase de reconhecimento as características a serem reconhecidas eram analisadas em função das características de referência obtidas na fase de treinamento. Esta análise era calculada através do uso do algoritmo DTW.

8.3.2 Quantização Vetorial

Com o propósito de reduzir a capacidade de armazenagem, e processamento consequentemente, enquanto mantendo a rejeição de locutores falsos versus a aceitação de locutores verdadeiros, neste trabalho foi utilizado um *codebook* generalizado para descrever o espaço de estados.

No cálculo dos *codebooks* foram utilizados 128 células de Voronoi, o qual é mais fácil de treinar, computacionalmente mais eficiente, e também usa significantemente menos memória **[PIR 90]**.

Na fase de treinamento, eram concatenados os vetores de características de pelo menos 3 amostras da palavra. Este conjunto de vetores concatenado era utilizado para calcular os *codebooks*. Os *codebooks* calculados eram então gravados para utilização posterior.

Na fase de reconhecimento os *codebooks* gravados eram analisados a fim de obter a menor distorção (distância entre o centróide VQ e o vetor a ser reconhecido).

8.4 Testes e validação

Para a validação foi projetado um sistema, como mostra a figura 8.2, onde o locutor falaria a senha e o sistema identificaria o mesmo.

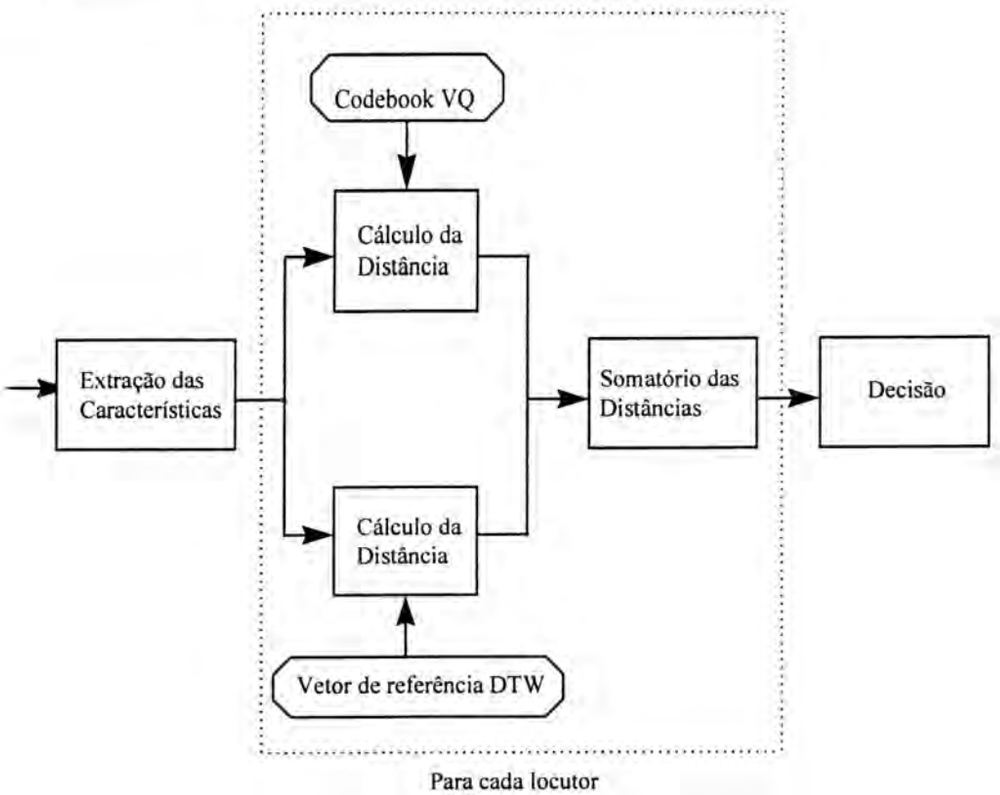

FIGURA 8.2 - Estrutura do sistema de reconhecimento de locutor utilizando VQ e DTW.

O primeiro passo foi gerar uma base de dados para treinamento, na qual seriam gravados os vetores de referência para o cálculo do DTW e *codebooks* de todas as amostras da base de dados. Esta abordagem é sugerida por **[PIR 90]** para que se obtenha uma grande amostragem estatística.

No segundo passo foram calculadas as distâncias extraídas do DTW e a distorção extraída do VQ. Para verificar a efetividade do sistema de identificação do locutor foi utilizada a curva de características de operação do receptor (**ROC**) utilizada pela psicofísica **[FUR 94]**. A curva ROC é obtida pela atribuição de duas probabilidades, a probabilidade de uma aceitação correta e a probabilidade de uma aceitação incorreta, para os eixos verticais e horizontais respectivamente, e variando o limiar de decisão. No caso da figura 8.3 o limiar é facilmente encontrado pois as curvas não se encontram.

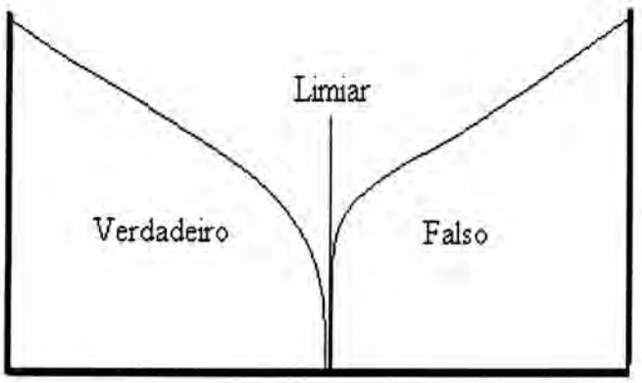

FIGURA 8.3 - Curva ROC ideal para reconhecimento de locutor.

Na figura 8.4 o limiar já não consegue separar perfeitamente as curvas probabilísticas de reconhecimento. Para isto, deve-se escolher um limiar que tente minimizar o problema do reconhecimento do locutor falso e o não reconhecimento do locutor verdadeiro.

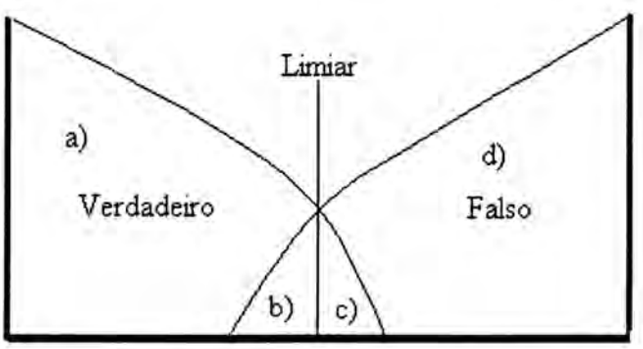

FIGURA 8.4 - Curva ROC com limiar não abrangente. Neste caso o reconhecimento pode ocorrer das seguintes maneiras: a) locutor verdadeiro reconhecido; b) locutor falso reconhecido; c) locutor verdadeiro não reconhecido; d) locutor falso não reconhecido.

Com estas curvas calculadas para todas as palavras e locutores da base de dados, foi calculado um limiar que mantivesse uma taxa pequena de erro. O valor para este limiar seria de 0.45 com uma taxa de acerto de 95%, onde a probabilidade do reconhecimento de um impostor será de 5% e a probabilidade de rejeição de um locutor verdadeiro será de 4.8%.

Antes de se obter este limiar para este conjunto de amostras, foi observado que o limiar para cada palavra diferenciava-se de uma para outra superficialmente. Esta alteração foi significante no que diz respeito às taxas de reconhecimento. A tabela 8.1 mostra alguns dos limiares e suas respectivas taxas de reconhecimento para o vocabulário utilizado.

A tabela 8.1 mostra que as taxas de reconhecimento estão ligadas à escolha da palavra, isto é, uma locução que contenha diferentes características fonéticas para uma melhor seleção do locutor. Com isto, pode-se definir que o limiar estará relacionado à palavra.

TABELA 8.A- Limiares e suas respectivas taxas de reconhecimento.

Palavra	Limiar	Falso Aceito	Falso Rejeitado
Startrek	0.430	6.5	6.7
Supernova	0.450	3.7	3.6
Tektronix	0.450	5.1	5.4
Generation	0.465	4.6	4.7
Nebula	0.505	2.8	3.9
Processing	0.485	4.2	4.4
Singularity	0.455	3.2	4.5
71523	0.620	10.7	9.4
Abracadabra	0.430	3.7	4.3
Sungeeta	0.445	4.2	4.2
Computer	0.475	4.6	4.6

A partir desta idéia foi necessário a atribuição de um limiar para cada locutor ou locução a fim de melhorar o desempenho do reconhecimento. Com este princípio foi observado que a taxa de reconhecimento atingiu níveis satisfatórios de reconhecimento, isto é, taxas que atingiam 96% de reconhecimento.

9 Conclusões

Este trabalho de dissertação foi idealizado e implementado visando contribuir para o aperfeiçoamento do sistema de reconhecimento de locutor, bem como mostrar os fundamentos teóricos e testes práticos.

Analisando a implementação realizada e apresentada no capítulo 7, pode-se realizar diversas considerações.

Na fase de treinamento foi verificado que a taxa de reconhecimento e rejeição para 10 locutores satisfaz as necessidades do sistema. Isto mostrou que o modelo *Multi-Layer Perceptron* é uma técnica eficaz para o mapeamento de problemas complexos. A arquitetura da rede pode ser estendida a fim de obter-se melhores taxas de reconhecimento e em ambientes ruidosos. Nesta fase verificou-se a necessidade de uma grande amostragem para as diversas configurações do locutor, isto é, as locuções realizadas em ambientes ruidosos, com *stress* físico ou emocional, amplitude, taxa de fala, entre outros.

Um caminho que pode ser utilizado na solução da complexidade do problema: é a divisão de tarefas através do uso de modelos específicos em cada tarefa, pelo fato de que o aparelho auditivo humano é composto por vários tipos de neurônios com diferentes graus de conexões **[KOH 82]**. Por isso, pode-se pensar em uma solução de separar os locutores em função de uma determinada característica para posterior de reconhecimento. Isto é, pode-se primeiro definir dois grupos que se diferenciam pelo *pitch* (mulheres e crianças em um grupo e homens em outro), para depois realizar o reconhecimento em um grupo menor ainda, por exemplo **[HWA 93]**.

O processo de reconhecimento de locutor necessita de uma alta precisão (em torno de 97% de acerto) pois a aplicação do mesmo será em atividades determinísticas que obrigam a certeza do resultado. Como o reconhecimento foi realizado com independência de texto e considerável número de locutores, a taxa de acerto não apresentou bons resultados na fase de validação.

A baixa taxa de reconhecimento apresentada na seção 7.3.3 pode ser resultado de uma série de fatores, importantes na configuração da solução do problema. Tais fatores pode ser descritos como:

- distância das características dos locutores no espaço de características: as características que foram utilizadas mostram que não são suficientes para definir uma caracterização bem definida do locutor;

- independência do texto: como as amostras foram obtidas de diferentes declarações, pode ter ocorrido que o mapeamento não foi abrangente suficientemente para a solução do problema proposto neste trabalho. De uma certa maneira, foi utilizada uma amostra de cada locução (número do andar mais a palavra "andar").

- nível de ruído do fundo: devido à presença de ruído nas amostras realizadas, foi comprometida a qualidade do sistema. Por isso, pode-se aplicar filtros para amenizar os efeitos causados pelos ruídos.

- número de amostras utilizadas para o treinamento.

Em alguns casos, o reconhecimento de locutor atingiu taxas de 100% devido a sua correta definição no espaço de características do locutor. Por isso, reforça-se a idéia de que a amostragem na fase de treinamento tem que ser bastante representativa para cada locutor.

Observou-se que as taxas de rejeição eram muito altas. Isto deve-se ao fato que para cada grupo de 50 amostras, somente 5 eram verdadeiras e 45 eram falsas, o que prejudicava no aprendizado devido à saturação no aprendizado da RNA.

Foi observado pela experimentação sobre uma palavra pré-definida, as taxas de reconhecimento atingiram valores como 88% devido a fatores como:

- dependência do texto: somente havia uma única frase;

- pequeno número de locutores: o espaço de características por ser menor facilitou a distribuição das características dos locutores neste espaço;

- número de amostras para treinamento: a utilização de um número similar de amostras tanto para locutor verdadeiro como para locutor impostor proporcionou um melhor aprendizado da RNA.

Deve-se ressaltar que, para um padrão de entrada pequeno utilizado neste trabalho, as taxas de acerto e rejeição não comprometem esta abordagem, visto que os trabalhos desenvolvidos para este problema utilizam na ordem de 500 a 2000 elementos característicos **[SOR 95a] [RUN 95]**.

Em função do capítulo 8 observa-se pelas taxas de reconhecimento, em comparação às obtidas no capítulo 7, que as técnicas convencionais para classificação predominam sobre as RNAs. Observa-se este fato na maioria dos sistemas de reconhecimento de locutor são baseados em técnicas convencionais de classificação.

Um outro fato observado foi que as técnicas convencionais não têm o problema de saturação, o qual percebeu-se nas RNAs. Além de que os padrões de treinamento não passavam de 3 amostras.

O único problema nas técnicas convencionais abordadas neste trabalho é encontrar um limiar com o objetivo de maximizar o reconhecimento de locutores verdadeiros e minimizar o reconhecimento de impostores. Este objetivo pode ser considerado universal entre as técnicas de classificação pois baseiam-se na qualidade da extração das características que diferenciam um locutor do outro.

Em trabalhos futuros, pode-se utilizar locuções mais ricas em sons vocálicos, aumentar a amostragem de características para que aumente a taxa de reconhecimento.

Pensa-se em reduzir a taxa de amostragem para auxiliar no cálculo dos parâmetros característicos, sem perder a qualidade do sinal.

Do ponto de vista da interface homem com a máquina, é importante considerar como os usuários devem ser avisados e como os erros de reconhecimento devem ser manuseados. Estudos de técnicas que automaticamente extraem os períodos de fala de cada pessoa separadamente de um diálogo, envolvendo mais do que duas pessoas, recentemente aparecem como uma extensão da tecnologia do reconhecimento do locutor.

Além disto tudo, com os conhecimentos em processamento de sinais adquiridos com o desenvolvimento deste trabalho, pode-se ainda realizar outras pesquisas como sistemas de verificação de locutores e até possivelmente trabalhar em problemas como reconhecimento de voz.

Em trabalhos futuros podem ser pesquisados novos modelos de RNAs que realizem a classificação das características. Pode-se pensar também em novos modelos que baseiam-se na hibridação de classificadores convencionais com RNAs.

É muito importante investigar parâmetros característicos que sejam estáveis sobre o tempo, insensíveis à variação do locutor, incluindo a taxa de fala, e robustos contra variações na qualidade da fala, devido a causas como resfriados e imitações. É também importante desenvolver um método para eliminar o problema da distorção causados pelos meios e ambientes de aquisição.

Enfim, o reconhecimento de locutor necessita de mais pesquisas para que se possa obter, em poucas amostragens, a identidade do locutor com alto grau de precisão para aplicações que envolvem segurança de acesso.

Bibliografia

[ADA 94] ADAMI, A. G. **Implementação de um método de Busca Heurística e Descrição de uma Interface Vocal.** Caxias do Sul: DI/UCS, 1994. Trabalho de Conclusão.

[ADA 96] ADAMI, A. G. **Estudo sobre o processamento de sinal de voz aplicado ao reconhecimento de locutor:** trabalho individual. Porto Alegre: CPGCC/UFRGS, 1994.

[ASC 86] ASCHEKENASY, E.; WEISS, M.R. **Multichannel advanced speech enhancement development.** New York: Rome Air Development Center, Griffs AFB, 1986. (Technical Report RADC-TR-86244).

[ATA 68] ATAL, B. S. **Automatic speaker recognition based on *pitch* contours**. New York: Polytechnic Institute,1968. Tese de Pós-Doutorado.

[ATA 84] ATAL, B. S; SCHRODER, M. R.. Stochastic coding of speech at very low bit rates. In: INTERNATIONAL CONFERENCE ON COMMUNICATIONS, 1984, Amsterdam. **Proceedings...** Amsterdam: [s. n.], 1984. p. 1610-1613.

[BAK 89] BAKER, J. K. A second-generation large vocabulary system. **Speech Technology**, [S. l.], v. 4, p. 20-24, Apr. 1989.

[BEN 91] BENNANNI, Y.; GALINNARI, P. A Modular Connectionist Architecture for Text-Independent Talker Identification. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORK, 1991, Seattle. **Proceedings...** Seatle: [s. n.], 1991.

[BEZ 95] BEZERRA, M. R.; SANTOS, S. C. B.; PRADO, P. P. L., Reconhecimento Automático de Locutor utilizando técnicas de Redes Neurais. In: SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES, 13., Campinas, 1995. **Anais...** Campinas: SID,1995.

[BRI 71] BRICKER, P. D. *et al.* Statistical techniques for talker identification **Bell Systems Technical Journal,** New York, v. 50, p. 1427-1454, Apr. 1971.

[DOD 71] DODDINGTON, G. R. A method of speaker recognition, **Journal Acoustics Society America**, [S. l.], v. 49, p. 139(A), Jan. 1971.

[DOR 93] DORNELES, Ricardo V. **Um Estudo sobre Processamento Adaptativo de Sinais Utilizando Redes Neurais**. Porto Alegre: CPGCC/UGRGS, 1993. Dissertação de Mestrado.

[EMB 91] EMBREE, Paul M.; KIMBLE, Bruce. **C Language Algorithms for Digital Signal Processing**. Englewood Cliffs: Prentice Hall, 1991.

[EPH 90] EPHRAIM, Y. A minimum mean square error approach for speech enhancement. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 1990, New Mexico. **Proceedings...** Albuquerque: [s. n.], 1990, p. 829-832.

[FUR 73] FURUI, S.; ITAKURA, F. Talker recognition by statistical features of speech sounds. **Electronics Communication**, [S. l.], v. 56-A, p. 62-71, 1973.

[FUR 81] FURUI, S. Cepstral analysis technique for automatic speaker verification. **IEEE Transactions on Acoustics, Speech, and Signal Processing,** New York, v. 29, p. 254-272, Apr. 1981.

[FUR 89] FURUI, S. **Digital speech processing, synthesis and recognition.** New York: Marcel Dekker, INC., 1989.

[FUR 94] FURUI, S. An overview of speaker recognition technology. In: WORKSHOP ON AUTOMATIC SPEAKER RECOGNITION, IDENTIFICATION AND VERIFICATION, 1994, **Proceedings...** [S. l.: s. n.], 1994.

[GRI 87] GRIFFIN, D. W. **Multi-Band Excitation Vocoder**.[S. l.]: MIT, 1987. Dissertação de Pós-Doutorado.

[HIG 93] HIGGINS, A. L. *et al.*. Voice identification using nearest-neighbor distance measure. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, New York, v. 2, p. 375-378, 1993.

[HWA 93] HWANG, M. Y.; HUANG, X. Shared-Distribution hidden Markov models for speech recognition. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, New York, v. 1, n. 4, 1993.

[JOS 89] JOSEPH, R. Large vocabulary voice-to-text systems for medical reporting. **Speech Technology**, [S. l.], v. 4, p. 49-51, Apr. 1989.

[KOH 82] KOHONEN, T. Self-organized formation of topologically correct feature maps. **Biological Cybernetics,** [S. l.], v. 43, p. 59-69, 1982.

[KOH 88] KOHONEN, T. **Learning Vector Quantization**. Berlin: Springer-Verlag, 1988.

[LEE 89] LEE, K. **Automatic speech recognition:** the development of the SPHINX system. Massachusets: Kluwer Academics Press, 1989.

[LEE 95] LEE, H. **SimNet Neural Network.** Disponível por WWW no endereço http://www.cs.umr.edu/~hlee/papers/voice (1995).

[LUF 94] LUFT, Joel A. **Reconhecimento Automático de Voz para Palavras Isoladas e Independente do Locutor**. Porto Alegre: PPGEMM/UFRGS, 1994. Dissertação de Mestrado.

[LUM 73] LUMMIS, C. Speaker verification by computer using speech intensity for temporal registration. **IEEE Trans. on Audio e Eletroacoustics**, New York, v. 21, p. 80-89, Jan. 1973.

[MAG 95] MAGNI, A. B.; CABRAL, E. Redes Neurais Artificiais e Informações de Excitação no Reconhecimento Automático do Locutor.In: CONGRESSO BRASILEIRO DE REDES NEURAIS, 2., São Carlos, 1995. **Anais...** São Carlos: [s. n.], 1995.

[MAK 75] MAKHOUL, J. Linear Prediction: a Tutorial Review. **Proceedings of the IEEE**, New York, v. 63, n. 1, p. 561-580, 1975.

[MAK 85] MAKHOUL, J. *et al.* Vetor quantization in speech coding. **Proceedings of the IEEE**, New York, v. 73, n. 11, p. 1551-1589, 1985.

[MAR 76] MARKEL, J. D.; GRAY, A. H. **Linear Prediction of Speech.** New York: Springer-Verlag, 1976.

[MEI 89] MEISEL, W. S.; FORTUNATO, M. P.; MICHALEK, W. D. A phonetically-based speech recognition system. **Speech Technology**, [S. l.], v. 4, p. 44-48, Apr. 1989.

[MOR 90] MORGAN, D. P. The application of dynamic programming to connected-speech recognition. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, New York, v. 7, n. 3, p. 6-25, July 1990.

[MOR 91] MORGAN, D. P.; SCOFIELD, C. L. **Neural Networks and Speech Processing**. Massachusets: Kluwer Academics, 1991.

[OGL 90] OGLESBY, J.; MASON J. S. Optimization of neural models for speaker identification. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, Albuquerque, Apr. 1990.

[OPP 75] OPPENHEIM, A. V.; SCHAFER, R. W. **Digital Signal Processing.** Englewood Cliffs: Prentice-Hall, 1975.

[PRA 95] PRADO, Maria C. **Uma Abordagem Neural à Identificação Homomórfica de Locutor**. Brasília: CPGCC/UB, 1995. Dissertação de Mestrado.

[PIR 90] PIRANI, G. **Advanced Algorithms and Architectures for Speech Understanding**. New York: Springer-Verlag, 1990.

[PRU 63] PRUZANSKY, S. Pattern-Matching procedure for automatic talker recognition. **Journal Acoustics Society America**, [S. l.], v. 35, p. 354-358, Mar. 1963.

[RAB 78] RABINER, L.R.; SCHAFER, R. W. **Digital Processing of Speech Signals.** Englewood Cliffs: Prentice-Hall, 1978.

[REY 75] REYNOLDS, D. A.; ROSE R. C. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, New York, v. 3, n. 1, p. 72-83, Jan. 1975.

[RUN 95] RUNSTEIN, F; VIOLARO, F. NUNES, H. F. Uso de Diferentes Parâmetros de Entrada em um Sistema de Reconhecimento de Fala Baseado em Redes Neurais. In: SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES, 13., 1995, Campinas. **Anais...** Campinas: [s. n.], 1995.

[SOR 95] SORIA, R. A. B.; CABRAL, E. Practical Speaker Identification Using Artificial Neural Networks. In: CONGRESSO BRASILEIRO DE REDES NEURAIS, 2., Curitiba. **Anais...** Curitiba: [s. n.], 1995.

[SOR 95a] SORIA, R. A. B.; CABRAL, E. Reconhecimento Automático de Locutor com Pré-Processamento Clássico e Classificação por Redes Neurais Artificiais. In: SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES, 13., 1995, Campinas. **Anais...** [S. l.: s. n.], 1995.

[TIS 91] TISHBY. Z. On the application of mixture AR hidden Markov models to text independent speaker recognition. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, New York, v. 39, p. 563-569, Mar. 1991.

[WAI 89] WAIBEL, A. *et al.* Phoneme recognition using time-delay neural networks, **IEEE Transactions on Acoustics, Speech, and Signal Processing**, New York, v. 37, p. 328-339, Mar. 1989.

[WEI 92] WEIHMANN, T. **Processamento digital de sinais aplicado à transmissão de voz**. Porto Alegre: PPGEMM/UFRGS, 1992. Dissertação de Mestrado.

[WIT 82] WITTEN, I. H. **Principles of Computer Speech**. London: Academic Press, 1982.

[ZIS 89] ZISSMAN, M. A. et al. Speech-state-adaptive simulation of co-channel talker iterference suppression. In: INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 1989, Scotland. **Proceedings...** Glasgow: [s. n.], 1989. p. 361-364.

Informática  UFRGS

CURSO DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Sistema de Reconhecimento de Locutor Utilizando Redes Neurais Artificiais.

por

André Gustavo Adami

Dissertação apresentada aos Senhores:

Prof. Dr. Euvaldo Ferreira Cabral Júnior (USP)

Prof. Dr. Altamiro Amadeu Suzim

Prof. Dr. Paulo Martins Engel

Vista e permitida a impressão.
Porto Alegre, 04/06/97.

Prof. Dr. Dante Augusto Couto Barone,
Orientador.

Profa. Carla Maria Dal Sasso Freitas
Coordenadora Substituta do Curso de
Pós-Graduação em Ciência da Computação
Instituto de Informática - UFRGS