UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

Reconhecimento de Voz para Comandos de Direcionamento por meio de Redes Neurais

por

JOÃO FRANCISCO VALIATI

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

Prof. Dr. Paulo Martins Engel
Orientador

Porto Alegre, Novembro de 2000.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Valiati, João Francisco

Reconhecimento de Voz para Comandos de Direcionamento por Meio de Redes Neurais / por João Francisco Valiati – Porto Alegre : PPGC da UFRGS, 2000.

128f. : il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2000. Orientador: Engel, Paulo Martins.

1. Reconhecimento de Voz. 2. Processamento de Sinais Digitais. 3. Redes Neurais. I. Engel, Paulo Martins. II. Título.

Agradecimentos

MUITO OBRIGADO, a todos que de uma forma ou de outra contribuíram para a realização deste trabalho, em especial:

- Ao meu orientador Prof. Paulo Engel, por seus ensinamentos, comentários, e opiniões no desenvolvimento deste trabalho e principalmente pela confiança que demonstrou em relação a minha capacidade;

- À minha família pelo carinho e todo apoio imensurável que me desejaram no decorrer deste trabalho;

- Ao colega Adriano Petry, pelo tempo dedicado à resolução de dúvidas pertinentes, por dicas e conselhos relativas ao processo de reconhecimento de voz e ao material cedido, os quais serviram como guia ao meu trabalho;

- Ao Prof. Luigi Carro, pelas suas aulas e empréstimo da base de dados produzida no Laboratório de Processamento de Sinais e Imagens do Departamento de Engenharia Elétrica;

- À colega Claúdia Gemelli pelo material concedido para pesquisa;

- À Suelaine Diniz pelo empréstimo de material e pela ajuda na resolução de dúvidas, que mesmo sem nos conhecermos não mediu esforços para responder aos meus e-mail;

- A todos os usuários que forneceram as suas vozes para compor as bases de dados, pois sem eles meu trabalho não seria possível;

- Aos meus amigos pelo apoio e votos de sucesso sempre desejados;

- Aos meus colegas Eduardo Appel, Claúdia Rizzi, Thais Cattani e Valesca Almeida, pela troca de idéias e realização de trabalhos disciplinares que contribuíram para nosso crescimento;

- À UFRGS, ao Instituto de Informática e ao PPGC pelos meios fornecidos para realização deste curso de Mestrado;

- Aos funcionários e ao pessoal da biblioteca pelo atendimento e profissionalismo demonstrado;

- Ao CNPq pela concessão da bolsa de estudos, pois sem ela a realização deste curso seria difícil.

Sumário

Lista de Abreviaturas..... 7

Lista de Figuras..... 8

Lista de Tabelas..... 11

Resumo..... 12

Abstract..... 13

1 Introdução..... 14
1.1 Abordagens ao Reconhecimento de Fala..... 15
1.2 Aplicações do Reconhecimento de Fala..... 17
1.2.1 Interfaces em Sistemas de Computação..... 17
1.2.2 Educação..... 17
1.2.3 Auxílio à Deficientes..... 18
1.2.4 Uso do Computador..... 18
1.2.5 Controle de Ambientes..... 18
1.2.6 Telecomunicações..... 18
1.2.7 Auxílio à Tarefas Profissionais..... 19
1.2.8 Aplicações e Produtos ao Consumidor..... 20
1.3 Outros Trabalhos Realizados..... 21
1.3.1 Reconhecimento de Palavras Isoladas Dependentes do Locutor..... 21
1.3.2 Reconhecimento de Comandos de Vídeo Cassete..... 21
1.4 Posicionamento e Justificativa..... 21
1.5 Objetivos do Trabalho..... 22
1.6 Estrutura da Dissertação..... 23

2 Processamento dos Sinais de Fala..... 24
2.1 O Processamento Natural..... 24
2.1.1 A Produção da Fala..... 24
2.1.2 A Percepção da Fala..... 28
2.2 Modelos para a Produção da Fala..... 29
2.3 Os Métodos para Processamento do Som..... 30
2.3.1 Aquisição do Sinal..... 31
2.3.2 Conversão Analógico/Digital..... 31
2.3.3 Amostragem..... 31
2.3.4 Quantização..... 32
2.3.5 Normalização..... 33
2.3.6 Janelas..... 33
2.3.7 Estimação da Energia..... 35
2.3.8 Taxa de Cruzamento por Zero..... 36
2.3.9 Análise Cepstral..... 38

3 Redes Neurais..... 41
3.1 Motivação..... 41
3.2 Inspiração Biológica..... 42
3.2.1 Potencial de Ação..... 42
3.3 O que são as RNAs ?..... 43
3.4 O Surgimento das RNAs..... 43
3.5 Neurônio Artificial..... 44
3.6 Funções de Ativação..... 45
3.7 Aprendizado..... 46
3.7.1 Aprendizado Supervisionado..... 47
3.7.2 Aprendizado Não-Supervisionado..... 47
3.8 Caracterização Geral das RNAs..... 48
3.9 A Rede *Backpropagation*..... 49
3.9.1 Arquitetura..... 49
3.9.2 Funcionamento..... 49
3.9.3 Regra de Aprendizado..... 50
3.9.4 O Fator da Convergência..... 51
3.9.5 Os Parâmetros Aprendizado e Momento..... 52
3.9.6 O Algoritmo *Backpropagation*..... 53
3.10 Teoria da Ressonância Adaptativa..... 53
3.10.1 *Fuzzy ART*..... 54
3.10.2 Codificação Complementar..... 56
3.10.3 Aprendizado em *Fuzzy ART*..... 56
3.10.4 *Fuzzy ARTMAP*..... 58
3.10.5 O Aprendizado de *Fuzzy ARTMAP*..... 59

4 Ferramentas e Métodos..... 61
4.1 Captura da Fala para Geração dos Padrões de Treinamento..... 62
4.2 Pré-processamento do Sinal Capturado..... 65
4.2.1 A Normalização..... 65
4.2.2 Determinação dos Limiares..... 65
4.3 Processamento para a Obtenção de Padrões..... 69
4.3.1 Determinação das Janelas..... 69
4.3.2 Extração dos Coeficientes Cepstrais..... 71
4.4 Apresentação dos Padrões as Rede Neural..... 72
4.4.1 A Rede *Backpropagation*..... 73
4.4.2 A Rede *Fuzzy ARTMAP*..... 74
4.3.3 Testes com o Simulink..... 75

5 Resultados Obtidos..... 77
5.1 Resultados da Base de Comandos de Direcionamento..... 77
5.1.1 Comandos de Direcionamento Aplicados à *Backpropagation*..... 77
5.1.2 Comandos de Direcionamento Aplicados à *Fuzzy ARTMAP*..... 85
5.2 Resultados de Dados Parciais da Base do Projeto Revox..... 86
5.2.1 Comandos da Base do REVOX Aplicados à BP..... 86
5.2.2 Comandos da Base do REVOX Aplicados à *Fuzzy ARTMAP*..... 87
5.3 Resultados dos Dados da Base do LaPSI..... 88
5.3.1 Comandos da Base do LaPSI Aplicados à BP..... 89
5.3.2 Comandos da Base do LaPSI Aplicados à *Fuzzy ARTMAP*..... 92

5.4 Comparações e Comentários sobre os Resultados Obtidos..... 93

6 Conclusão e Considerações Finais..... 95
6.1 Trabalhos Futuros..... 96

Anexo 1 – Resultados da Saída da Rede R2..... 98

Anexo 2 – Tabela de Resultados..... 123

Bibliografia..... 124

Lista de Abreviaturas

A/D	Analógico/Digital
ANN	Artificial Neural Networks
ART	Adaptive Resonance Theory
DTW	Dynamic Time Warping
eq	Equação
EMQ	Erro Médio Quadrado
HMM	Hidden Markov Models
Hz	Hertz
KHz	Kilohertz
LaPSI	Laboratório de Processamento de Sinais e Imagens
LPC	Coeficientes de Predição Linear
LMS	Least Mean-Square
LTM	Long Term Memory
GHz	Gigahertz
MCP	McCulloch & Pitts
MLP	Multi-Layer Perceptron
MSVQ	Quantização Vetorial Multisecção
ms	Milisegundos
mV	Milivolts
ns	Nanosegundos
PC	Personal Computer
PSHNN	Parallel Self-Organizing Hierarquical Neural Network
RAL	Reconhecimento Automático do Locutor
RAV	Reconhecimento Automático de Voz
RNAs	Redes Neurais Artificiais
RN	Rede Neural
UFRGS	Universidade Federal do Rio Grande do Sul
WTA	Winner Take All

Lista de Figuras

FIGURA 2.1 – Representação da glote fechada e aberta..... 26
FIGURA 2.2 - Aparelho Fonador..... 27
FIGURA 2.3 - Sistema Auditivo..... 28
FIGURA 2.4 - Modelo simples para geração da voz..... 29
FIGURA 2.5 - Modelo fonte/filtro..... 30
FIGURA 2.6 - Conversão analógico/digital..... 31
FIGURA 2.7 - (a)Função contínua (b)Função amostrada..... 32
FIGURA 2.8 - Tipos de janelas..... 34
FIGURA 2.9 - Evolução da energia do comando direita..... 36
FIGURA 2.10 - Evolução da taxa de cruzamento por zero do comando direita..... 37
FIGURA 2.11 - Comparação de multiplicações requeridas pelo cálculo direto e o requisitado pela FFT..... 39
FIGURA 2.12 - Processo para obtenção dos cepstros..... 40
FIGURA 3.1 - Neurônio Biológico..... 42
FIGURA 3.2 - Neurônio Artificial de McCulloch e Pitts..... 44
FIGURA 3.3 - Funções de Ativação..... 45
FIGURA 3.4 – Aprendizado Supervisionado..... 47
FIGURA 3.5 – Aprendizado Não-Supervisionado..... 47
FIGURA 3.6 - Rede *Feedforward* com 4 camadas..... 48
FIGURA 3.7 - Rede Recorrente de uma camada..... 49
FIGURA 3.8 - Arquitetura de *Fuzzy ART*..... 55
FIGURA 3.9 - Arquitetura da *Fuzzy ARTMAP*..... 58
FIGURA 4.1 - Modelo do sistema proposto e implementado para o reconhecimento de comandos falados..... 61
FIGURA 4.2 - Tela de configuração do Creative WaveStudio..... 63
FIGURA 4.3 - Tela para determinação dos meios e volumes utilizados..... 63
FIGURA 4.4 - Tela principal do Creative WaveStudio..... 64
FIGURA 4.5 - Tela para determinação e gravação de novas amostras..... 64
FIGURA 4.6 - Diagrama em bloco do algoritmo de determinação dos *endpoints*..... 66
FIGURA 4.7 - Diagrama para a estimação do *endpoint* inicial baseado na energia..... 67
FIGURA 4.8 - Diagrama para a estimação do *endpoint* final baseado na energia..... 68
FIGURA 4.9 - Determinação dos *endpoints* de uma locução do comando direita..... 69
FIGURA 4.10 - Comando direita com 0,988 segundos de duração, aplicadas 6 janelas adaptativas com 76% de superposição e 4950 amostras por janela..... 70
FIGURA 4.11 - Comando direita com 0,908 segundos de duração, aplicadas 6 janelas adaptativas com 76% de superposição e 4550 amostras por janela..... 71
FIGURA 4.12 - Arquitetura da rede neural utilizada..... 73
FIGURA 4.13 - Módulo para captura e reconhecimento de comandos..... 76
FIGURA 5.1 - Gráfico dos resultados obtidos para o comando **direita** relativos à base de treinamento sobre as redes BP..... 78

FIGURA 5.2 - Gráfico dos resultados obtidos para o comando **esquerda** relativos à base de treinamento sobre as redes BP..... 79
FIGURA 5.3 - Gráfico dos resultados obtidos para o comando **siga** relativos à base de treinamento sobre as redes BP..... 79
FIGURA 5.4 - Gráfico dos resultados obtidos para o comando **pare** relativos à base de treinamento sobre as redes BP..... 80
FIGURA 5.5 - Gráfico dos resultados obtidos para o comando **recue** relativos à base de treinamento sobre as redes BP..... 80
FIGURA 5.6 - Gráfico de desempenho geral das rede neurais BP a todos os comandos de voz que fizeram parte da base de treinamento... 81
FIGURA 5.7 - Gráfico dos resultados obtidos para o comando **direita** aos novos dados desconhecidos apresentados as redes BP..... 82
FIGURA 5.8 - Gráfico dos resultados obtidos para o comando **esquerda** aos novos dados desconhecidos apresentados as redes BP..... 82
FIGURA 5.9 - Gráfico dos resultados obtidos para o comando **siga** aos novos dados desconhecidos apresentados as redes BP..... 83
FIGURA 5.10 - Gráfico dos resultados obtidos para o comando **pare** aos novos dados desconhecidos apresentados as redes BP... 83
FIGURA 5.11 - Gráfico dos resultados obtidos para o comando **recue** aos novos dados desconhecidos apresentados as redes BP... 84
FIGURA 5.12 - Gráfico de desempenho geral das rede neurais BP a todos os comandos de voz provenientes do grupo de amostras desconhecidas..... 84
FIGURA 5.13 - Gráfico de resultados obtidos com a rede *Fuzzy ARTMAP* sobre comandos de voz provenientes das amostras de usuários desconhecidos da base de comandos de direcionamento..... 85
FIGURA 5.14 - Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes das amostras de treinamento da base de comandos do REVOX..... 87
FIGURA 5.15 - Gráfico de resultados obtidos com a rede BP sobre comandos de voz de usuários desconhecidos provenientes da base de comandos do REVOX..... 87
FIGURA 5.16 - Gráfico de resultados obtidos com a rede *Fuzzy ARTMAP* sobre comandos de voz de usuários desconhecidos provenientes da base de comandos do REVOX..... 88
FIGURA 5.17 - Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes do conjunto de amostras utilizadas no treinamento do conjunto da base de comandos do LaPSI..... 89
FIGURA 5.18 - Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes da metade do conjunto de amostras utilizadas no treinamento da base de comandos do LaPSI..... 90
FIGURA 5.19 - Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes da metade de amostras desconhecidas do conjunto da base de comandos do LaPSI 90
FIGURA 5.20 - Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes do treinamento de 70% do conjunto da base de comandos do LaPSI..... 91

FIGURA 5.21 - Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes de 30% de amostras desconhecidas do conjunto da base de comandos do LaPSI 91
FIGURA 5.22 - Gráfico de resultados obtidos com a rede *Fuzzy ARTMAP* sobre comandos de voz provenientes da metade de amostras desconhecidas do conjunto da base de comandos do LaPSI..... 92
FIGURA 5.23 - Gráfico de resultados obtidos com a rede *Fuzzy ARTMAP* sobre comandos de voz provenientes de 30% de amostras desconhecidas do conjunto da base de comandos do LaPSI 93

Lista de Tabelas

TABELA 5.1 – Resultados alcançados no treinamento das redes neurais BP com diferentes parâmetros..... 77
TABELA A2 – Taxas de reconhecimento das redes neurais sobre as bases de dados..... 123

Resumo

Este trabalho relata o desenvolvimento de uma aplicação capaz de reconhecer um vocabulário restrito de comandos de direcionamento pronunciados de forma isolada e independentes do locutor.

Os métodos utilizados para efetivar o reconhecimento foram: técnicas clássicas de processamento de sinais e redes neurais artificiais. No processamento de sinais visou-se o pré-processamento das amostras para obtenção dos coeficientes cepstrais. Enquanto que para o treinamento e classificação foram utilizadas duas redes neurais distintas, as redes: *Backpropagation* e *Fuzzy ARTMAP*.

Diversas amostras foram coletadas de diferentes usuários no sentido de compor um banco de dados flexível para o aprendizado das redes neurais, que garantisse uma representação satisfatória da grande variabilidade que apresentam as pronúncias entre as vozes dos usuários.

Com a aplicação de tais técnicas, o reconhecimento demostrou-se eficaz, distinguindo cada um dos comandos com bons índices de acerto, uma vez que o sistema é independente do locutor.

Palavras-chave: Processamento de Sinais, Redes Neurais Artificiais, Reconhecimento de Voz, Reconhecimento de Palavras Isoladas, Independência de Locutores.

TITLE: “Speech Recognition for Direction Commands with Neural Networks”

Abstract

This work reports the development of an application able to recognize a restrict vocabulary of direction commands pronounced in isolated form and speaker independent.

The methods used to perform the recognition were: classic signal processing techniques and artificial neural networks. The classic signal processing techniques are responsible for the pre-processing of the samples to extract the cepstral coefficients. In the training and classification tasks we used two neural models: Backpropagation and Fuzzy ARTMAP.

Several samples were collected from different users to compose a flexible database to train the Neural Networks, guaranteeing a satisfactory representation of the large variability showed by the users voices.

The recognition proved to be effective, discriminating each command with a good rate of recognition, confirming that this system is speaker independent.

Keywords: Signal Processing, Artificial Neural Networks, Speech Recognition, Isolated Word Recognition, Speaker Independent.

1 Introdução

A constante evolução tecnológica faz do reconhecimento de fala um campo de estudos fascinante e ao mesmo tempo desafiador, uma vez que é bastante grande a gama de aplicações em que a voz tem o papel de agilizar e facilitar a realização de tarefas cotidianas, buscando extrair da fala as informações relevantes para a realização do reconhecimento. E ao mesmo tempo, buscam-se sempre metodologias inovadoras que permitam um reconhecimento mais ágil e preciso para fazer com que a máquina compreenda a linguagem falada.

Um antigo desejo do homem foi sempre poder controlar suas máquinas por meio da fala. Talvez o início disto tudo seja decorrente de séculos passados e que se prolonga até os dias atuais, embora numa minoria, com o uso de animais que servem de auxílio em tarefas, principalmente em atividades agrícolas, onde o uso da palavra domina seus atos[VAL 99]. Com o advento da tecnologia que muito evoluiu, as máquinas predominam em quase todos os cenários, desde a conquista espacial até o uso residencial, sendo assim nada melhor do que dotar tais equipamentos com a capacidade de percepção e compreensão da voz humana, que é a forma mais simples, natural e eficaz do ser humano expressar seus pensamentos, e desta forma humanizar mais a comunicação homem-máquina.

As pesquisas relacionadas ao reconhecimento de fala iniciaram em meados deste século, ainda sem a utilização do computador, por meio de experimentos mecânicos que simulavam a produção sonora humana[WIT 82]. Inicialmente, o reconhecimento da fala foi muito difundido e essa tecnologia foi considerada como de fácil dominação. À medida que eram descobertas as diferenças de freqüências e intensidade da voz de uma pessoa para outra vislumbrou-se também que tal tarefa não se tratava de um problema de fácil resolução.

Atualmente, os estudos pertinentes ao Reconhecimento Automático de Voz (RAV) evoluíram muito graças aos avanços tecnológicos como o computador e as placas de som, as quais tornaram possível a conversão do som em dados digitais. As técnicas de processamento de sinais permitem a extração de características que realmente mereçam destaque, pois atuam no sentido de fornecer não somente a informação de interesse ao processamento de determinada amostra de som como também ocasionar uma redução considerável na quantidade de informações a serem processadas. Tais informações serão responsáveis pela produção de padrões a serem identificados numa comparação entre determinada referência previamente registrada e a apresentação de uma nova amostra para teste, onde o papel fundamental do sistema será validar ou não determinada amostra, dependendo do tipo e funcionalidade do sistema de reconhecimento em questão[TRI 94][FUR 89].

Dentre os métodos para comparação entre padrões existem vários, podendo-se destacar: Variação Temporal Dinâmica (*Dynamic Time Warping* -DTW), Modelos Ocultos de Markov (*Hidden Markov Models* -HMM) e Redes Neurais Artificiais (*Artificial Neural Networks* -ANN). Todos já aplicados em experimentos que pesquisam esta tecnologia e podem ser utilizados individualmente ou em conjunto no processo de reconhecimento da fala[DEL 93][RAB 78].

O uso das Redes Neurais sobre o reconhecimento da fala ocorreu na década de 80, pelo fato das mesmas possuírem uma capacidade de processamento paralelo e distribuído, visando a redução nas baixas taxas de confiabilidade que os sistemas existentes na época obtinham com a utilização de outros métodos. Sua utilização também buscava fundamentos por ser uma técnica que possui sua origem proveniente da Inteligência Artificial, a qual visa investigar e extrair os conhecimentos relativos ao cérebro humano, estando assim relacionada com a produção da fala que é um processo cognitivo e inteligente exclusivo da espécie humana.

E desta forma com a utilização de técnicas clássicas de processamento digital de sinais e Redes Neurais Artificias é que este trabalho visa a validação de tais técnicas na efetivação do reconhecimento de determinados comandos falados independentes do locutor.

1.1 Abordagens ao Reconhecimento de Fala

O estudo da fala está dividido em 3 áreas principais: análise, síntese e reconhecimento, estando o reconhecimento subdividido no Reconhecimento Automático do Locutor (RAL) e no RAV.

O RAL objetiva a discriminação de indivíduos tanto pela identificação como pela verificação de locuções, que por meio da extração de características distintas que cada locutor apresenta, tornam este tipo de sistemas aplicável a usos forenses ou a controles de acesso a determinados sistemas ou áreas restritas.

O RAV, por sua vez, deve compreender automaticamente uma elocução, que pode ser tratada sob 3 aspectos: o reconhecimento de palavras isoladas, palavras conectadas e fala contínua. No reconhecimento de palavras isoladas é estabelecido que deve existir um intervalo de silêncio, de mais de 250 ms, separando cada palavra. Isto é diferente do reconhecimento de palavras conectadas, em que não é necessário haver pausas forçadas entre palavras, mas uma pronúncia clara de cada locução, o que ocasiona, geralmente, um intervalo de silêncio na ordem de 50 ms. O reconhecimento de fala contínua está relacionado a uma conversação natural, onde certas palavras estão naturalmente ligadas a outras produzindo um único som, tornando a realização deste tipo de reconhecimento bastante complexa.

A maioria das aplicações realiza o reconhecimento de palavras isoladas, e sua utilização destina-se para seleção de menus ou em resposta a tipos de questões “sim/não”. Quando se trabalha com o reconhecimento de palavras isoladas é necessário considerar o tempo de resposta, o qual se refere ao período de tempo que o sistema gasta do final da pronúncia de uma palavra até a resposta do reconhecimento.

Este tempo de resposta geralmente se alonga um pouco visto que deve ocorrer um certo período de silêncio até que seja declarado o final da palavra e a fase de reconhecimento inicie. A maioria das aplicações deve tolerar um tempo de resposta não superior a 500 ms para palavras isoladas.

Outra abordagem do RAV refere-se à dependência ou não do locutor no reconhecimento. Sistemas dependentes do locutor são capazes de reconhecer a fala de

um único locutor ao qual foi treinado. Os usuários deste tipo de sistema devem sempre treiná-lo antes de usá-lo. Este treinamento envolve alguns fatores relativos:

- a escolha de um vocabulário;

- cada sistema é treinado a fim de reconhecer a fala de um determinado locutor e seu vocabulário específico;

- é necessário que cada usuário repita o vocabulário de palavras ordenadas diversas vezes para ser criado um padrão;

- o sistema deve ser treinado em um ambiente apropriado.

Após o usuário concluir a fase de treinamento, o sistema habilitará o usuário ao seu vocabulário específico.

Entre as vantagens de sistemas dependentes do locutor, destacam-se:

- a flexibilidade do vocabulário: o usuário é capaz de adicionar ou alterar palavras para um vocabulário simples, bastando-lhe um treinamento prévio;

- o tamanho do vocabulário: o sistema pode suportar um vocabulário extenso.

As desvantagens observadas em sistemas deste tipo são:

- a limitação do número de usuários;

- a necessidade de uma grande quantidade de processamento antes que o sistema venha a ser utilizado e o fato de certas aplicações poderem englobar muitos usuários, pode ocasionar uma elevada capacidade de armazenamento do sistema para os parâmetros dos diversos locutores;

- a fase de treinamento é desagradável ao usuário;

- devido a fatores como: cansaço, doença, posição e tipo do microfone, a voz do mesmo locutor varia.

Os sistemas de reconhecimento independentes do locutor são aqueles que realizam a identificação de uma entrada falada sem necessitar de um treinamento prévio, devendo ser capazes de reconhecer diferentes tipos de vozes para um vocabulário específico. A independência de locutor é um requerimento indispensável em aplicações para redes telefônicas, por exemplo, pois a identidade do locutor geralmente não é conhecida.

Um sistema de reconhecimento de voz independente do locutor deve fornecer alta exatidão no reconhecimento, independente do sexo, dialeto e outras características relevantes ao locutor.

Os desenvolvedores usualmente criam vocabulários para o reconhecimento independente do locutor por meio de um grande banco de dados de fala. Em geral,

quanto maior e mais representativo for o banco de dados mais riqueza terá o vocabulário. A rede pública de telefones é um exemplo de um ambiente desafiante aos sistemas RAV pela gama de variedades que a compõe.

É estimado, por exemplo, que para representar adequadamente o reconhecimento de voz em uma rede telefônica são necessários, aproximadamente, em média 1000 locutores, os quais devem prover condições de atender a grande variação de pronúncias. Isto, assumindo-se um vocabulário composto por dígitos de “0” a “9” e poucas palavras de controle.

A inexistência da fase de treinamento por usuários e a minimização da memória de armazenamento, onde um único conjunto de padrões acomoda múltiplos usuários são tidos como algumas das vantagens encontradas nestes sistemas[SHA 82].

Um tópico relevante é que a taxa de erro de um sistema de reconhecimento independente do locutor é de três a cinco vezes superior ao de um dependente do locutor.

1.2 Aplicações do Reconhecimento de Fala

1.2.1 Interfaces em Sistemas de Computação

As aplicações deste tipo utilizam o reconhecimento de fala como uma interface para sistemas de computador, assim como para aplicativos que executam sobre esses sistemas. A maior parte destes sistemas utiliza a fala para a manipulação de menus, janelas, caixas de diálogo, palhetas e outros componentes.

Em sua maioria as interfaces são projetadas para controle do ambiente Microsoft® Windows e seus aplicativos. É possível a criação de vocabulário para menus e outros componentes do ambiente Windows. Alguns aplicativos requerem entrada de palavras isoladas enquanto outros oferecem entrada de fala contínua.

1.2.2 Educação

Sistemas deste tipo, na sua maioria, estão relacionados a simulações que realizam com o usuário buscando alguma semelhança com uma futura situação prática que o aluno poderá vir a enfrentar dependendo da atividade que venha a exercer. Seja com técnicas para operações do mercado financeiro, tais como venda e compra de ações na bolsa, ou para sistemas que simulem o controle de tráfego aéreo configurados para representar situações corriqueiras que possam ocorrer, onde o operador deve fornecer informações para o piloto simulado ajustar velocidade, altitude e direção para preparar o pouso.

Também é interessante apresentar um sistema de aprendizado, citado por Rosenblum apud [MAR 96], que permite aos estudantes observar a representação gráfica de como os fonemas específicos que eles produzem estão sendo pronunciados, o que auxilia no ensino de adultos e crianças em sessões de fonoaudiologia, auxiliando por meio da visualização de imagens a maneira correta da produção de determinados sons.

1.2.3 Auxílio a Deficientes

As pessoas com deficiências usam o reconhecimento de fala para realizar tarefas que os tornem mais independentes em sua vida pessoal. Um dos principais problemas encontrados no desenvolvimento de tais sistemas é que cada grupo de pessoas com deficiências possui requerimentos e interesses únicos de sua deficiência, que variam desde pessoas paraplégicas até portadores da Síndrome do Esforço Repetitivo. E que, embora a diversidade de seus problemas, possuem o objetivo comum de independência pessoal[SCO 94].

O reconhecimento de fala surge como uma importante ferramenta no auxílio a pessoas com deficiências físicas, sendo que a maioria das aplicações desenvolvidas são destinadas para pessoas com impedimento motor ou visual, mas também atendendo indivíduos com problemas de fala e audição[HAR 89].

1.2.4 Uso do Computador

Sistemas de ditado são amplamente utilizados em reconhecimento de fala, tanto para uso pessoal como para profissional, uma vez que seu uso não requer a utilização das mãos, adequando-se a pessoas com impedimentos motor e também para auxiliar cegos em determinadas aplicações. A ligação da síntese com o reconhecimento da fala é muito importante para que pessoas com problemas visuais possam verificar a entrada de dados reconhecida.

As aplicações que fazem uso do computador tornam-se acessíveis, uma vez que os comandos estão todos adaptados ao reconhecimento da voz, fazendo do computador um meio de lazer e trabalho a estas pessoas.

1.2.5 Controle de Ambientes

O controle de cadeiras de roda, camas de hospital, luzes e demais itens do ambiente pessoal são muito úteis para ajudar em tarefas desta natureza por meio da voz, tornando o ambiente adequado às necessidades de cada usuário.

Tais sistemas são formados por um pequeno e prático vocabulário de comandos projetados para usuários portadores de grandes ou temporárias deficiências físicas, sua atenção está mais voltada para controle de equipamentos hospitalares como camas e outros objetos que fazem parte deste ambiente.

1.2.6 Telecomunicações

Tais tarefas estão relacionadas a operação de serviços, como por exemplo, a substituição da antiga operadora que atendia à solicitação de chamadas do consumidor para determinada localidade pela requisição de chamada do usuário diretamente ao sistema de reconhecimento da fala.

A requisição de serviços que representam um campo bastante rentável nas telecomunicações, incluindo serviços de discagem pela voz, direcionamento de chamadas, chamada por cartão e a maioria dos serviços pagos. Para estes serviços serem atrativos eles devem fornecer descrição sucinta dos benefícios e serem fáceis de usar.

Dentre os vários serviços fornecidos destaca-se a utilização de páginas amarelas, onde o usuário requisita o nome de uma empresa, o produto ou o nome promocional e o sistema se encarrega de fornecer o número, como também já possibilita a realização da ligação direta com a empresa. Outro serviço permite que sejam realizadas ligações de qualquer telefone, bastando ao usuário informar um número de identificação, previamente, fornecido pela operadora, possibilitando que tenha acesso às suas ligações que estão em sua secretária eletrônica, como também realize ligações de sua própria linha telefônica, não sendo necessário para isto saber todo o número da pessoa com quem deseja entrar em contato, mas somente fornecer um apelido previamente cadastrado.

1.2.7 Auxílio à Tarefas Profissionais

Entre as aplicações que auxiliam em tarefas profissionais destacam-se:

- área de saúde: o registro de pacientes hospitalares por sistemas de ditado de dados cadastrais e sistemas para a produção de relatórios de diagnóstico utilizados em algumas especialidades médicas como radiologia, patologia e medicina de emergência, conforme Kurzweil apud [MAR 96]. Cada qual com um vocabulário próprio, conforme os termos técnicos de suas áreas. Outra aplicação de interesse é o protótipo de um robô ativado pela voz, responsável por agendar e entregar a medicação ao paciente nos horários estabelecidos, assim como alertar a central em casos de emergência. Ele responde à entrada de comandos e controles, altamente estruturados, pronunciados pelo paciente;

- área jurídica: acesso a bases de dados on-line, onde textos referentes a determinada requisição oral do termo de consulta ao sistema busca artigos que já são projetados no editor de textos, propiciando agilidade na composição de textos jurídicos;

- em serviços bancários e cartões de crédito: essencialmente ligados à utilização do telefone, servindo no atendimento ao cliente. A fala utilizada em serviços do tipo *home-banking* oferece diversas vantagens aos clientes, uma vez que está 24 horas disponível, não existem filas demoradas, o cliente pode checar saldos e extratos, transferir fundos entre duas contas, realizar aplicações, e requisitar informações. Em aplicações que envolvem o uso de cartões de crédito a preocupação está mais relacionada com a verificação do usuário do que com o reconhecimento, uma vez que fraudes freqüentemente são detectadas, limitando o número de operações com cartões de crédito a consultas como: limite do cartão e datas de pagamento.

1.2.8 Aplicações e Produtos ao Consumidor

O desenvolvimento de produtos e aplicações ao consumidor com tecnologia de reconhecimento de fala é bastante novo e desafiante; a definição dos produtos é muito ampla, assim como a definição da maneira correta para a voz ser aplicada. A variedade de plataformas é grande, possibilitando ao sistema de reconhecimento estar dentro de uma máquina de médio porte, ou até mesmo embutido em um *chip*.

Algumas aplicações já desenvolvidas dão conta de atender:

- eletrodomésticos: vídeos cassete, TVs, refrigeradores, máquinas de lavar e secar e microondas;

- viagens: sistemas para traduções simultâneas, máquinas fotográficas, quiosques de informações e navegação de carros;

- diversão: brinquedos interativos, quiosques interativos, jogos de computador;

- serviços: quiosques, *drive-thru* para alimentos e bancos.

Uma das áreas que teve grande desenvolvimento foi a de atendimento aos clientes, tanto de restaurantes *fast-food*, como o serviço bancário. Outro campo com grande potencial são os quiosques, que trazem informações das mais variadas às pessoas que os consultam, englobando desde informações turísticas em determinados lugares como a *Disney World* à demonstração de automóveis aos consumidores.

Para a implantação da tecnologia de reconhecimento de fala em aparelhos eletro-eletrônicos é importante saber qual será o benefício que o mesmo trará. Por exemplo, os consumidores esperam que um controle por meio da voz de seus vídeos cassete, possa incorporar benefícios como poder programar o equipamento para gravar, o que é uma atividade um pouco complexa para pessoas que não possuem muita experiência.

Outra aplicação onde a fala já começa a aparecer é na criação de brinquedos como veículos e bonecas, possibilitando as crianças interagir mais com seus brinquedos e desenvolver novas atividades de recreação. Também, na produção de jogos para computadores, o reconhecimento de fala ainda é algo novo, pois sua implantação neste tipo de jogos é bastante difícil uma vez que a música e os demais sons relativos ao próprio ambiente do jogo tornam-se desafios para o reconhecimento[MAR 96].

A quantidade de aplicações e produtos onde a voz pode estar presente torna-se bastante ampla conforme a imaginação dos criadores. Avanços nos algoritmos para reconhecimento, melhor qualidade dos microfones e redução dos custos tecnológicos são fatores que contribuem para a expansão do reconhecimento de fala.

1.3 Outros Trabalhos Realizados

1.3.1 Reconhecimento de Palavras Isoladas Dependentes do Locutor

Em 1995, Tafner propôs em seu trabalho de mestrado o reconhecimento de palavras faladas de forma isolada por um determinado locutor.

Para a realização de tal experimento foi feita a extração de dados das amostras, visando uma redução na quantidade de informações. Com isto foi possível: a eliminação do ciclo negativo do sinal amostrado, a detecção da forma de onda (envoltória), a mediação do sinal amostrados e a normalização do sinal mediado. Com a efetivação destas etapas os dados foram apresentados para a rede neural de Kohonen, que foi treinada e testada, obtendo índices de reconhecimento na faixa de 80% a 90%[TAF 95].

O trabalho de Tafner se caracteriza por utilizar uma outra forma de abordagem na extração das características significativas das amostras, visando a redução do volume de dados. E por utilizar uma rede neural com aprendizado não-supervisionado.

1.3.2 Reconhecimento de Comandos de Vídeo Cassete

Um trabalho proposto por Diniz, visou comparar e avaliar a utilização de três diferentes modelos de redes neurais aplicadas ao reconhecimento de comandos de voz independentes do locutor relativos ao controle de um vídeo cassete[DIN 97].

Para a construção da base de dados foram utilizados somente locutores do sexo masculino, responsáveis por fornecer amostras sonoras em um ambiente sem isolamento acústico, captados por uma placa *Sound Blaster* convencional conectada ao computador.

De cada sinal amostrado foi realizada a extração de quarenta características: 12 *mel-cepstrum*, 12 *delta cepstrum*, 1 *delta* energia, 1 *delta log* energia, 12 *cepstrum*, 1 taxa de cruzamento por zero e 1 relação taxa de cruzamento por zero entre cada metade do sinal, e uma variante do Discriminante de Fischer.

As características extraídas foram aplicadas aos modelos neurais: Perceptron de Múltiplas Camadas (MLP), Rede de Funções de Base Radial (*Radial Basis*) e Rede Neural Hierárquica Paralelamente Auto-organizával (*Parallel Self-Organizing Hierarquical Neural Network* -PSHNN). As taxas de reconhecimento obtidas estão na ordem de 95%, 98% e 96%, respectivamente para cada modelo.

1.4 Posicionamento e Justificativa

Diante dos atuais avanços tecnológicos, principalmente na área de *hardware*, os quais possibilitam o surgimento de novos sistemas computacionais capazes de tornar acessível a comunicação homem-máquina com a utilização da voz, o número de pesquisas relativas ao reconhecimento de fala cresce a cada instante.

Outro ponto importante é a grande quantidade de possíveis aplicações onde o reconhecimento de fala pode atuar, seja no controle de sistemas pessoais ou na

manipulação de equipamentos e utensílios eletrônicos, visando a agilidade e praticidade na realização de tarefas.

Desta forma, buscou-se o desenvolvimento de um sistema capaz de efetivar o reconhecimento de palavras faladas de forma isolada e independente do locutor, relacionadas a comandos de direcionamento, através do uso de técnicas de processamento de sinais e redes neurais artificiais.

O sistema proposto foi implementado com a utilização do software Matlab 5.3, o qual implementa uma linguagem de programação interpretada. A escolha de sua utilização deve-se a fatores como facilidades de prototipação e de desenvolvimento de modelos matemáticos. O Matlab é uma ferramenta amplamente utilizada no meio científico, possuindo funções dedicadas ao referido trabalho, tais como processamento de sinais e modelagem de redes neurais.

1.5 Objetivos do Trabalho

A principal meta deste trabalho é a construção de um sistema que reconheça um vocabulário restrito, composto pelos comandos: **direita**, **esquerda**, **siga**, **pare** e **recue**; os quais podem ser destinados para o direcionamento de um veículo autônomo, com o uso de palavras faladas de forma isolada e independente do locutor, onde a utilização de redes neurais será de fundamental importância na tarefa de reconhecimento.

Como objetivos mais específicos são relatados:

- A realização da aquisição do som através de uma interface analógica-digital (A/D) para compor a base de dados do sistema;

- A melhoria da representatividade do sinal digital convertido em relação ao sinal original, por meio de técnicas de pré-processamento e processamento de sinais digitais;

- A extração das informações relevantes do sinal digital que representem o sinal original, sem perda das principais características próprias de determinada amostra sonora;

- A utilização das redes neurais *Backpropagation* e *Fuzzy ARTMAP* para o treinamento e aprendizado das redes, como também para realizar a tarefa de comparação de padrões: os de teste com os de referência;

- O desenvolvimento de um sistema computacional, unindo ambas as técnicas de processamento de sinais com as de redes neurais artificiais, que se demostre capaz de efetuar o reconhecimento com eficiência;

- Aplicação de outras bases de dados para validar os métodos propostos através dos resultados obtidos.

1.6 Estrutura da Dissertação

Está monografia apresenta no segundo capítulo o processamento geral da fala, sua produção e percepção pelo ser humano, assim como todo o processo matemático e computacional para a aquisição das principais informações que caracterizam a fala. O terceiro capítulo descreve as redes neurais, sua definição, origem, funcionamento e principais características, com destaque para a descrição do *Backpropagation* e *Fuzzy ARTMAP*, ambos modelos neurais utilizados neste trabalho. O quarto capítulo destina-se ao relato da aplicação desenvolvida, desde a forma como as amostras foram adquiridas até o treinamento e reconhecimento pela rede neural. No quinto capítulo são apresentados os resultados obtidos pelo reconhecimento, tanto para os comandos que compõem a base de treinamento como também para novas amostras de usuários desconhecidos ao sistema, também são aplicados testes com outras bases de dados sonoros e os resultados comparados. O sexto e último capítulo relata as conclusões e considerações finais do trabalho.

2 Processamento dos Sinais da Fala

2.1 O Processamento Natural

A importância do estudo da produção e percepção da fala é fundamental para uma melhor compreensão deste magnífico processo, assim como tornar possível a criação de modelos que auxiliem no seu conhecimento.

A voz é uma característica que só os humanos possuem, e se baseia na produção de sons articulados, formados pela inteligência originando uma linguagem que é a fonte para a fala. A voz não é meramente a emissão de uma onda sonora, este som comunica alguma coisa, não somente no conteúdo da mensagem que está sendo transmitida, mas também diz respeito ao estado físico e emocional do locutor, que visa provocar, freqüentemente, inconscientemente, uma reação no ouvinte.

De um lado, está o cérebro humano que produz uma seqüência de comandos motores que controlam diversos músculos do aparelho fonador para produzir o som desejado. De outro lado, a mensagem é captada pelo sistema auditivo e convertida em uma série de pulsos neurológicos que fornecem subsídios para a própria produção da fala, uma vez que o próprio locutor também é um ouvinte de sua voz, o que lhe oferece um maior controle e monitoramento de seus órgãos vocais.

A produção da fala surge com a concepção de uma idéia ou pensamento que o locutor quer transmitir ao ouvinte, a mensagem a ser transmitida é convertida em uma estrutura lingüística pela escolha de palavras e termos apropriados, segundo as regras particulares de cada linguagem, no sentido de compor frases que representem a idéia original. Ainda é necessário adicionar a esta informação as características globais ou locais, tal como entonação ou ênfase que expressam aspectos importantes da comunicação pelo tom emocional que acompanha a locução, manifestando a vontade que o locutor quer que esta informação represente, com o intuito de produzir uma reação no ouvinte[DEL 93].

Todo este processo é tão automático e espontâneo que sua atuação passa despercebida, permitindo uma perfeita produção e interpretação destas ondas sonoras, mesmo em ambientes compostos por diversos sons diferentes ou semelhantes que são distinguidos com grande agilidade.

2.1.1 A Produção da Fala

A produção da fala é possível graças ao aparelho fonador, o qual não é, exclusivamente, responsável pela produção da voz, mas também por funções vitais relacionadas a respiração e alimentação. Basicamente, os órgãos que o constituem são: a boca, os pulmões, a traquéia e a laringe, que são controlados pelo sistema nervoso que os adapta para moldar os sopros por eles emitidos através dos lábios, língua, dentes e palato produzindo a fala.

A “máquina humana” que hoje fala não nasceu sabendo falar, antes de estar pronta para funcionar teve que ir aprendendo a reconhecer sons distintos e fazer uso da memória para reproduzi-los.

O recém-nascido tem pouca ou nenhuma experiência sonora, mas aos poucos dias de vida já é capaz de associar o tom, a altura fundamental do gerador glótico da mãe, a intensidade e duração de outros sons, como sensações placentárias e/ou desagradáveis que acontecem ao mesmo tempo. Claro que ainda não é capaz de compreender o significado de uma palavra, mas pode identificar sons distintos de carinho, ameaça, alegria ou susto. Por isso, ainda que se diga a ele: “não gosto de você nem um pouco”, quando a modulação do tom fundamental corresponde a uma sensação de alegria por vê-lo, a criança sorrirá, especialmente se quem está a lhe falar é uma voz familiar. A articulação glótica é o que se aprende primeiramente e este aprendizado continua durante toda a vida, de tal forma que quando adultos é possível distinguir variações muito sutis de ironia, dor, alegria, tristeza, medo ou vergonha. A freqüência chega independentemente do contexto da frase, pela freqüência fundamental, amplitude, duração e as variações da freqüência dos impulsos glóticos[ROC 87].

O processo de produção da fala no homem é realizado de forma espontânea e é bastante complexo se visto passo a passo. Logo abaixo são citados os principais componentes do aparelho fonador, constituído pelos seguintes órgãos:

- Pulmões: funcionam como dois foles, produzindo a corrente de ar;

- Brônquios e Traquéia: são os canais que conduzem a corrente de ar até a laringe;

- Laringe: é o órgão mais importante da fonação e está situada na parte superior da traquéia. Na laringe estão localizados a glote, a epiglote (válvula elástica que tapa a glote durante a deglutição) e as cordas vocais;

- Glote: é uma pequena abertura de forma triangular localizada na laringe, na altura do pomo-de-adão. Com a chegada do fluxo de ar vindo dos pulmões, a glote pode abrir-se ou fechar-se, bastando que os bordos das cordas vocais se afastem ou se aproximem. Se a glote se abrir, o ar passa livremente, sem fazer vibrar as cordas vocais: assim, o fonema produzido é surdo. Ao contrário se a glote se fechar, o fluxo de ar força a passagem, fazendo vibrar as cordas vocais, produzindo um fonema sonoro. A glote pode ser melhor compreendida através da fig. 2.1;

- Cordas Vocais: são duas pregas musculares, que são distendidas horizontalmente diante da glote devido a sua elasticidade;

- Faringe: cavidade ligeiramente afunilada, entre a boca e a parte superior do esôfago, conduz o ar para a boca e fossas nasais;

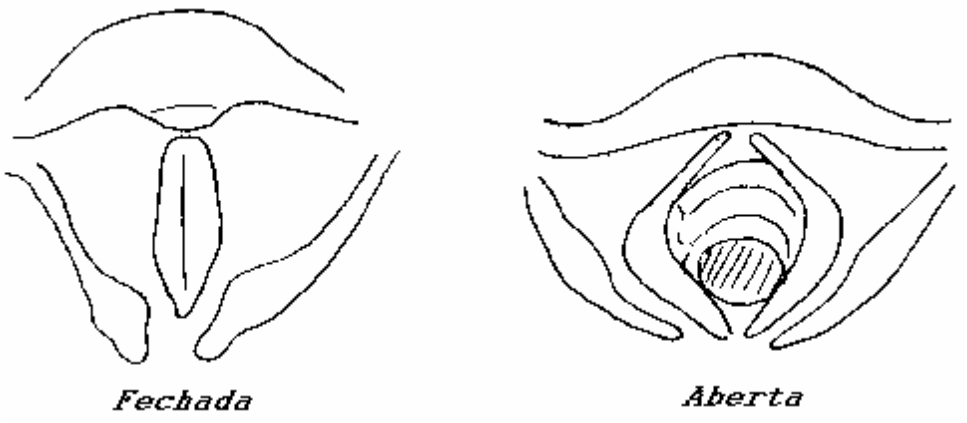


FIGURA 2.1 - Representação da glote fechada e aberta

- Úvula: é o apêndice flexível do véu palatino, também conhecido vulgarmente como "campainha". Tem a função de fiscalizar a passagem do ar. Levantando-se contra a parede posterior da faringe intercepta a passagem de ar para as fossas nasais forçando o ar escoar-se pela boca, gerando o que chamamos de fonema oral. Abaixando-se a úvula, a corrente de ar escapa em parte para as fossas nasais, produzindo um fonema nasal;

- Boca e órgãos anexos: pode-se dizer que os fonemas nascem na laringe e se completam na boca. Isto acontece graças ao concurso das arcadas dentárias, dos alvéolos, do palato duro (céu da boca) e do palato mole (ou véu palatino) e sobretudo, à atividade da língua, lábios e das bochechas, os quais se movimentam para modificar a corrente sonora e moldar os fonemas;

- Fossas Nasais: cavidades situadas no maxilar superior, funcionam como caixa de ressonância dos fonemas nasais.

Embora a maior parte das ondas sonoras seja originária da boca o som também emana das narinas, garganta e bochechas. O funcionamento do mecanismo de produção de fala descrito acima e visualizado na fig. 2.2, ocorre quando a pressão do ar contido nos pulmões segue pela traquéia até a laringe e chega a glote, que é o orifício entre as cordas vocais (fig. 2.1). Geralmente, a glote está aberta, durante a respiração, permitindo fluxo livre de ar. No momento em que as cordas vocais são estiradas, a glote é temporariamente fechada, obstruindo a passagem de ar, modulando o ar em pulsos discretos[FUR 89]. O trato vocal que inicia na glote e termina nos lábios é um tubo acústico não uniforme e variável[RAB 75a], que atua como um tubo ressonante com a finalidade de filtrar os sons da fonte, ou seja, o conjunto de pulsos produzidos. Um dos principais componentes do trato vocal são os articuladores, compostos por partes móveis, como: maxilar, véu palatino, língua e lábios. O véu palatino possui um papel importante, pois atua abrindo ou fechando a cavidade nasal, controlando o som produzido pelas narinas.

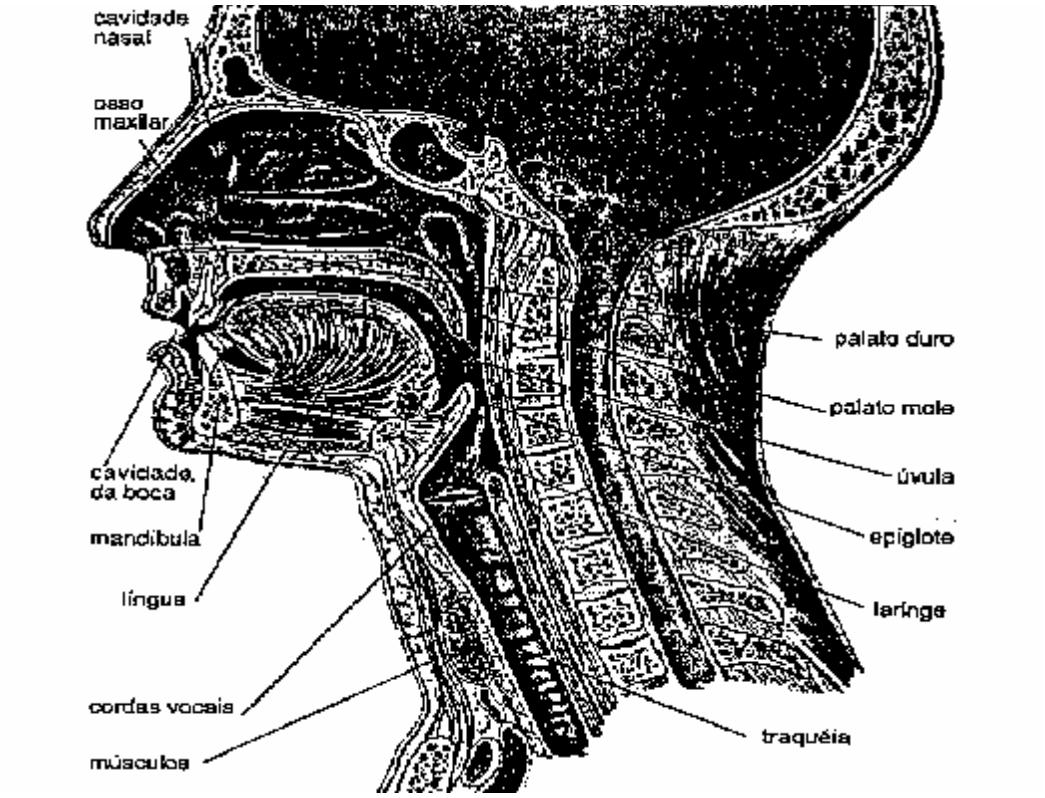


FIGURA 2.2 - Aparelho Fonador

O movimento de oscilação das cordas vocais ocorre em uma freqüência fundamental, *pitch*, a qual varia durante a fala com relação a magnitude das variações de pressão sobre a glote, da tensão nas cordas vocais e da massa das bordas vibrantes, gerando a entonação da pronúncia. As freqüências produzidas, podem variar na faixa de 80 a 150 Hz para homens, 150 a 250 Hz nas mulheres e superior a 250 Hz para crianças. Apesar dos valores apresentados, a linguagem articulada abrange sons com freqüências variantes de 250 Hz a 500 Hz [FOL 68]. As freqüências ressonantes são chamadas de formantes e são conhecidas como F_1,F_2,F_3...F_n, onde as três primeiras são as mais importantes, fornecendo a identidade da vogal.

A onda acústica gerada é irradiada para fora do trato vocal, pelos lábios e narinas, que produzem sons que podem ser classificados como vocálicos e não-vocálicos. Os sons vocálicos são produzidos quando a passagem de ar através do trato vocal se dá de forma contínua e sem turbulência, como é o caso das vogais, onde o som é produzido em função das freqüências de ressonância do trato vocal naquele instante. Já os sons não-vocálicos são gerados quando o trato vocal impõe resistência à passagem do ar e estão associados a determinadas consoantes classificadas como fricativas e explosivas[LUF 91]. Abaixo, listam-se as principais características dos tipos de sons (fonemas):

- Fricativos não-vocálicos: /f/, /s/ e /ch/, o som da fonte é um ruído produzido por uma corrente de ar vinda dos pulmões em um estreitamento da cavidade bucal. Para o /f/ este estreitamento ocorre entre o lábio inferior e os dentes superiores; no /s/ entre a ponta da língua e a borda dos dentes; e para o /ch/ entre a língua e o palato duro;

- Fricativos vocálicos: /v/ e /z/ possuem duas fontes: a vibração das cordas vocais e a turbulência do ar;

- Explosivos não-vocálicos: /p/, /t/ e /k/, na produção destes sons a passagem de ar através da cavidade bucal é interrompida por um curto espaço de tempo, acumulando a pressão do ar que é repentinamente liberada, produzindo um som curto e explosivo. A oclusão no /p/ ocorre entre os lábios, no /t/ entre a ponta da língua e as bordas dos dentes, e no /k/ entre a parte posterior da língua e o véu palatino;

- Explosivos vocálicos: /b/,/d/ e /g/, também ocorre uma abertura repentina da cavidade bucal, mas diferem-se das não-vocálicas pelo fato de as cordas vocais vibrarem durante a oclusão[HAR 83].

2.1.2 A Percepção da Fala

A produção da fala não tem valor algum se o ouvido e o cérebro do ouvinte não forem capazes de decodificar suas características acústicas e compreender a mensagem falada. O ouvido não atua como um telefone receptor, mas realiza uma considerável quantidade de análises no sinal de entrada, antes de enviar a mensagem, parcialmente, processada para o cérebro. O cérebro age como um incrível processador de padrões poderosíssimo, decifrando a mensagem em tempo real.

O ouvido está dividido em três seções: ouvido externo, médio e interno. E é visualizado na fig. 2.3.

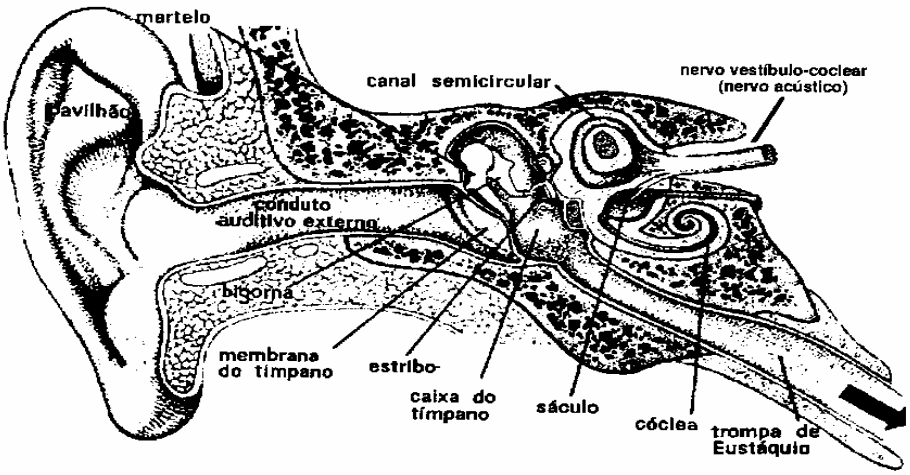


FIGURA 2.3 - Sistema Auditivo

O ouvido externo é a parte visível, é composto pelo pavilhão auditivo e o canal auditivo externo. Sua função é captar o som, servindo como uma corneta acústica que conduz a onda sonora até o tímpano. O tímpano é uma membrana que vibra em resposta aos sons que entram da mesma forma que o diafragma de um microfone.

O ouvido médio transmite e amplifica mecanicamente as pressões sonoras captadas pelo ouvido externo até o interno, e é formado por três pequenos ossos: o

martelo, a bigorna e o estribo, responsáveis por transferir a energia vibracional do tímpano ao ouvido interno através da janela oval.

O ouvido interno, o qual consiste de um sistema de cavidades preenchidas com fluído, possui a cóclea, que é uma estrutura espiral que converte as vibrações sonoras (mecânicas) em impulsos nervosos (elétricos), os quais são captados por cerca de 40 mil fibras nervosas, cada uma sintonizada em uma faixa de freqüência. O ouvido humano pode diferenciar cerca de 1400 freqüências diferentes, sendo que em acústica são consideráveis audíveis as freqüências na faixa de 20 Hz a 20 KHz.

Existem várias teorias que procuram modelar o processo de percepção da fala e que são classificadas como: passivas, a percepção é considerada como um processo independente da produção de fala; ou ativas, onde a percepção e produção são processos que interagem[FAN 86].

A teoria acústica, como uma teoria passiva, considera a fala como uma resposta do filtro do trato vocal a uma fonte sonora; as unidades de som que compõem a fala são mapeadas pelo cérebro em características distintas que serão, posteriormente, processadas para serem decodificadas em fonemas, sílabas e palavras; o cérebro considera várias características do sinal ao mesmo tempo e tenta encontrar um padrão que reconheça.

Entre as teorias ativas, destaca-se o modelo de análise-por-síntese, que propõe que o ouvinte percebe a fala gerando hipóteses sobre o que virá a seguir, baseado no que já foi escutado e nas regras de geração da fala, comparando estas hipóteses internas com o sinal de entrada. Existe um consenso entre pesquisadores que, ambos mecanismos, ativo e passivo, são utilizados no processo de percepção da fala, mesmo não se sabendo a extensão de cada um.

2.2 Modelos para Produção da Fala

Um modelo simples para a geração da voz é baseado no mecanismo humano de produção de fala, como descrito anteriormente. Aqui ele é representado matematicamente, por uma fonte geradora de impulsos, representada por *g(t),* e separada da articulação *h(t)* que pela convolução da excitação gera o som irradiado[FUR 89]. Este modelo é representado na fig. 2.4.

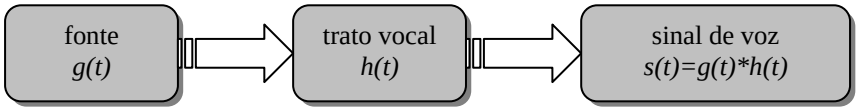


FIGURA 2.4 - Modelo simples para geração da voz (o símbolo ‘*’ representa a convolução)

O modelo apresentado anteriormente pode ser aperfeiçoado, considerando que a fonte possa ser modelada de duas formas: uma para geração dos sons vocálicos e outra para os sons não-vocálicos. O novo modelo é conhecido como modelo digital fonte/filtro, onde uma fonte de sons vocálicos é produzida por um gerador de pulsos e o intervalo entre os pulsos corresponde ao período fundamental (*pitch*), os quais são filtrados por um modelo de pulso glotal *G(z)*. Os sons não-vocálicos são construídos a partir de um gerador de ruído randômico. Estas fontes são amplificadas e aplicadas a um filtro digital variável no tempo que utiliza parâmetros relativos ao trato vocal. Tais características variam lentamente com o tempo, entretanto podem ser consideradas constantes durante intervalos de aproximadamente 10 ms[RAB 75a]. Este modelo pressupõe que as amostras relativas a onda de voz saiam de um filtro digital variável no tempo, que é responsável por aproximar as propriedades de transmissão do trato vocal e as propriedades espectrais da forma do pulso glotal[DIN 97]. O modelo é apresentado na fig. 2.5.

A finalidade da utilização de modelos que representem a produção de fala é criar uma ligação com o sistema humano que produz a fala, tornando possível o estudo e uma melhor compreensão deste mecanismo no sentido de dar suporte à extração da informação significativa do sinal sonoro.

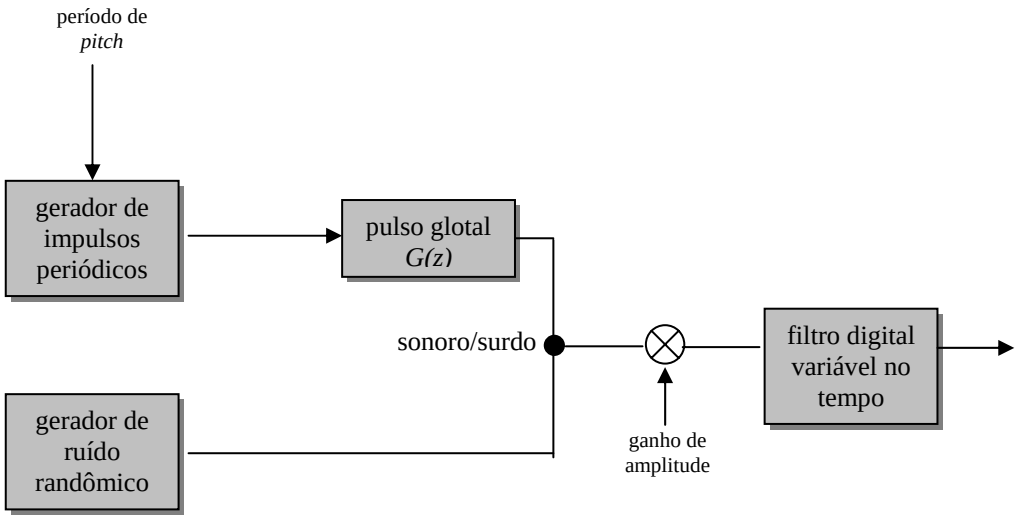


FIGURA 2.5 - Modelo fonte/filtro

2.3 Os Métodos para o Processamento do Som

Os métodos descritos a seguir visam extrair do sinal sonoro as principais características que representem o sinal original. O conjunto destas técnicas aplicadas sobre a amostra de fala ocasiona uma redução significativa na quantidade de informações a serem processadas, oferecendo parâmetros seguros para o treinamento e aprendizado de uma rede neural.

2.3.1 Aquisição do Sinal

A aquisição de uma amostra de fala, com o uso de um computador, ocorre através de um microfone e de uma placa de som. Ambos mecanismos citados são indispensáveis, o primeiro é responsável pela captura do som irradiado em determinado ambiente e pela sua conversão em um sinal elétrico, conduzindo-o a uma placa de som, que por sua vez transforma o impulso elétrico em dados discretizados, possibilitando a sua manipulação pelo computador.

Também é necessário que o computador possua algum software que permita o manuseio de informações sonoras, ou melhor, o registro e reprodução das mesmas, tornando possível o armazenamento deste tipo de informação.

2.3.2 Conversão Analógico/Digital

A conversão A/D permite que os sons presentes na natureza e aqueles que os humanos produzem sejam discretizados, no sentido de serem processados em um computador. É necessário que eles estejam na forma digital, sendo necessário que sejam transformados inicialmente em sinais elétricos, que por sua vez são gerados por um transdutor acústico-elétrico, um microfone. Convertido o som em eletricidade, pode-se transformá-lo em informação digital, através da tensão elétrica em determinado tempo qualquer, correspondente a amplitude de onda. Com o sinal digitalizado é possível que sistemas computacionais o manipulem. O processo de conversão do sinal analógico em informação digital é visualizado na fig. 2.6.

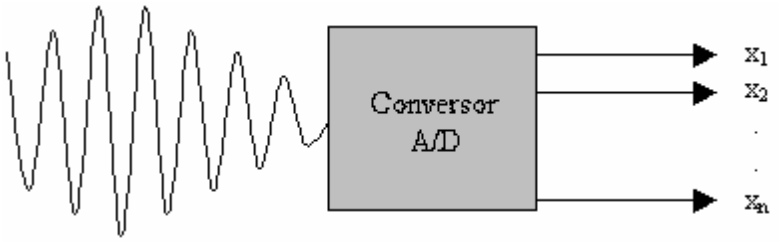
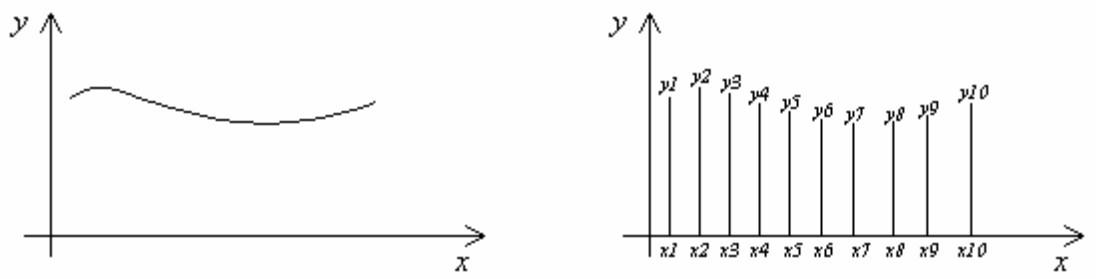


FIGURA 2.6 - Conversão analógico/digital

2.3.3 Amostragem

Amostrar um sinal é extrair amostras em intervalos fixos de tempo, ou seja, é representar um sinal variável sob uma forma contínua por uma série de sinais discretizados. Para uma melhor compreensão, suponha uma função contínua, fig. 2.7a, onde existe uma variável *x* representando o tempo ou uma freqüência, e uma variável dependente *y* referenciando uma tensão ou pressão que possuam uma ligação funcional *y=f(x)*, onde a estimação de todos os pontos, *y=f(x)*, permite um conhecimento completo do processo analisado. Mas, como *x* possuí uma variação contínua seria necessária a determinação de um número infinito de pontos para a obtenção de um conhecimento completo do processo inteiro.

A técnica de amostragem é justificada pelo fato que para o conhecimento do processo não é necessária a determinação de todos os pontos, mas somente a estimação de alguns valores de *y* que correspondam a certos valores de *x*, os quais garantem o formato da curva como um todo, fig. 2.7b, sem perda das características principais.



(a) (b)

FIGURA 2.7 - (a) Função contínua (b) Função amostrada

O teorema da amostragem especifica que quando se possui um sinal de banda limitada, com suas componentes apresentando valores abaixo de uma determinada freqüência, chamada $f_{máx}$, os mesmos devem ser descritos por um conjunto de amostras obtidas a uma taxa de amostragem de $2f_{máx}$. Garantindo que a taxa de amostragem seja rápida o suficiente para que no mínimo duas amostras sejam obtidas durante o período correspondente à sua componente espectral de freqüência mais elevada do sinal[TAU 82].

Caso a taxa de amostragem não seja alta o suficiente, pode ocorrer o efeito de *aliasing*, que introduz componentes que não faziam parte do sinal original, na sua reconstrução.

Uma boa taxa de amostragem para os sinais de voz é estabelecida em torno dos 10000 Hz, uma vez que as freqüências formantes da fala produzidas pelo idioma português brasileiro não ultrapassam a faixa de 5000 Hz. O fato de possuir 10000 Hz de amostragem em um sinal significa que 10000 amostras estarão contidas a cada segundo que compõe determinada locução.

2.3.4 Quantização

A quantização é o domínio da amplitude do sinal analógico contínuo amostrado em um determinado intervalo de tempo, ou seja, é a medida discreta da intensidade do sinal. A discretização da amplitude é usualmente definida em número de bits. Por exemplo, uma conversão de 8 bits representa 2^8 estados ou 256 níveis de quantização.

Uma conversão de 8 bits (256 níveis) é considerada satisfatória e uma conversão de 12 bits (4096 níveis) é boa o suficiente para a maioria das aplicações. O aumento no número de bits significa uma faixa menor de valores contínuos a cada valor discreto, obtendo uma maior resolução, entretanto ocasiona um aumento na quantidade de dados a serem processados[MAL 85].

Normalmente, a quantização corresponde ao intervalo da tensão elétrica captada. Por exemplo, se o intervalo da tensão elétrica é estabelecida como de 1*volt* de pico a pico do sinal e uma quantização de 16 bits é definida, isto significa que se tem 65536 níveis de quantização distribuídos neste intervalo, sendo os primeiros 32768 níveis correspondentes a faixa do sinal de –0,5V a 0V e dos 32769 aos 65536 níveis correspondentes a faixa positiva onde a tensão é superior a 0V, estendendo-se até 0,5V.

2.3.5 Normalização

A normalização procura eliminar a diferença de intensidade existente entre diversas amostras, a qual é provocada, na maioria dos casos, pela distância da boca em relação ao microfone.

Caso a intensidade do sinal seja muito elevada, ou seja, a tensão elétrica tenda a ultrapassar os 0,5V estabelecidos, por exemplo, provavelmente devido a um posicionamento muito próximo entre a boca e o microfone, a mesma deve sofrer um processo de normalização, que deverá restringir a intensidade do sinal entre a faixa de –0,5V e 0,5V pela eq. 2.1.

$$sinal\_normalizado = sinal\_amostrado*0,5/maior\_valor\_sinal \quad (2.1)$$

Onde o sinal que ultrapassar o limiar de intensidade estabelecido deverá ser multiplicado pelo valor do limiar (0,5), seguido da divisão pelo maior valor presente neste sinal, efetivando assim a sua normalização.

2.3.6 Janelas

Para se trabalhar com aplicações práticas de processamento de sinais, é necessário o uso de termos curtos ou *frames* do sinal. Isto torna-se desnecessário caso o sinal seja de curta duração. Assume-se que um *frame* $F_x(n;m)$ é definido como uma função (janela) discreta $w(n)$ de tamanho N, que termina no tempo m, multiplicada pela seqüência de voz discreta $x(n)$, onde n é a variável de tempo discreto:

$$F_x(n;m) = x(n)w(m-n) \quad (2.2)$$

As janelas são seqüências finitas utilizadas para selecionar um *frame* desejado do sinal original. Existem diversos tipos de janelas, os quais podem ser vistos na fig. 2.8.

Independente do uso de uma determinada janela, as mesmas são sobrepostas quando aplicadas sobre um sinal no sentido de reduzir as distorções causadas pela descontinuidade nos limites de cada janela. Os inícios das janelas de Hamming devem ser deslocados em no máximo $N/4$ amostras, resultando em uma sobreposição igual ou superior a 75%, enquanto que os inícios das janelas retangulares devem ser deslocados em no máximo $N/2$ amostras, o que resulta em uma sobreposição igual ou superior a 50%. O janelamento deve ser realizado com o uso de uma sobreposição mínima necessária a fim de evitar o efeito de *aliasing*[DIN 97].

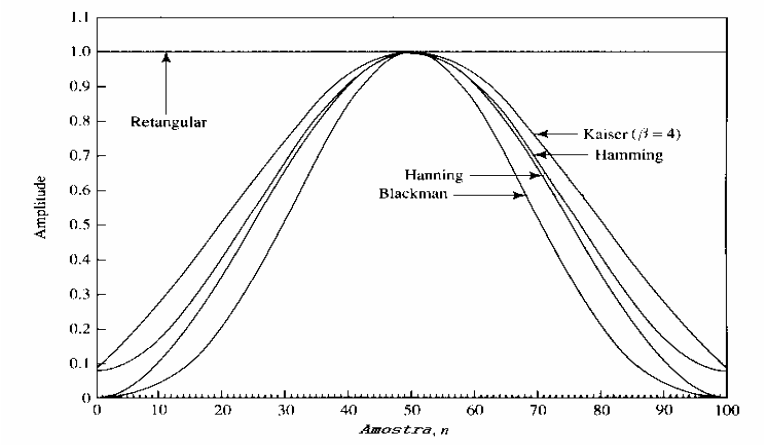


FIGURA 2.8 - Tipos de janelas

O janelamento pode ser realizado com ou sem sobreposição entre janelas. O papel da sobreposição é aumentar a correlação entre janelas sucessivas para que variações bruscas entre as características extraídas das janelas adjacentes sejam evitadas, assim como impedir mudanças abruptas nas extremidades das janelas. O único inconveniente da sobreposição é o aumento no tempo de processamento[BEZ 94].

Uma diferenciação entre dois tipos de janelas é feita comparando-se a janela retangular, a qual preserva as características temporais da forma de onda sobre o alcance dos *N* pontos, mas realiza o corte nas fronteiras da forma de onda, com as demais janelas que apresentam um corte mais suave, tendendo a distorcer a forma de onda temporal no alcance dos *N* pontos, mas com o benefício de ter um corte menos abrupto nas fronteiras.

Abaixo estão exemplificadas as expressões para a janela retangular, eq. 2.3, e para a janela de Hamming, eq. 2.4, uma das janelas mais populares utilizadas no reconhecimento da fala segundo trabalhos consultados[DIN 97][VIZ99].

$$w(n) = \begin{cases} 1, & n = 0,1,...,N-1 \\ 0, & n \text{ outros casos} \end{cases} \tag{2.3}$$

$$w(n) = \begin{cases} 0{,}54 - 0{,}46\cos(2\pi n / N - 1), & n = 0,1,...,N-1 \\ 0, & n \text{ outros casos} \end{cases} \tag{2.4}$$

Muitos sistemas de reconhecimento como aqueles baseados em redes neurais, necessitam de um número fixo de janelas, para toda e qualquer locução. Neste caso, podem ser empregados os seguintes métodos de janelamento:

- Decimação/Interpolação: este método é baseado na compressão (decimação) ou expansão (interpolação) da duração do sinal de voz, fazendo com que

todas as locuções fiquem com o mesmo número de amostras. Uma desvantagem que este método apresenta é a alteração da taxa de amostragem do sinal original, uma vez que são incluídas ou extraídas amostras, ocasionando uma mudança na duração da cada locução;

- Sobreposição Variável: visa fixar o tamanho de cada janela e determinar a sobreposição para cada locução, fazendo com que o número de janelas seja sempre o mesmo em todas as locuções, independente do tempo de duração de cada uma. Por este método as variações bruscas entre as características extraídas de janelas adjacentes são evitadas em razão do aumento da correlação entre janelas sucessivas[DIN 97];

- Janela Adaptativa: é baseado em uma sobreposição fixa e na variação do tamanho de cada janela, fazendo com que o número de janelas de cada palavra seja sempre o mesmo, independente do tempo de duração da locução. A vantagem deste método é propiciar um ajustamento temporal das locuções de maneira uniforme, uma vez que as janelas tenderão a corresponder aos mesmos fonemas pronunciados em frases idênticas com tempo de duração diferentes[BEZ 94].

2.3.7 Estimação da Energia

A medida de energia é uma forma simples de se representar um sinal. Considerando-se *x[n]* a *n*-ésima amostra do sinal *x*, a energia é obtida por:

$$E = \sum_{n=-\infty}^{\infty} x^2[n] \tag{2.5}$$

Para casos em que o sinal é não estacionário, incluindo-se os sinais da fala, a variação temporal da energia é calculada da seguinte maneira:

$$E[n] = \sum_{m=0}^{N-1} [w[m]x[n-m]]^2 \tag{2.6}$$

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

Como *E[n]* é uma potência do sinal, que apresenta características da variação temporal, ela tem a propriedade de dar grande ênfase aos sinais de maior amplitude. A fim de suavizar este efeito utilizam-se valores absolutos ao invés dos quadrados, resultando:

$$\hat{E}[n] = \sum_{m=0}^{N-1} | w[m]x[n-m] | \tag{2.7}$$

Em sistemas de reconhecimento de voz, a medida de energia pode ser usada para a determinação dos limites da palavra, bem como para separar os sons vocálicos dos não-vocálicos, uma vez que a energia dos não-vocálicos é inferior a dos sons vocálicos.

A medida de energia não é um parâmetro muito bom para o reconhecimento por ser muito sensível ao locutor e ao ambiente, mostrando-se ser dependente de características da captura do sinal como: sensibilidade do microfone e particularidades de quantização do conversor A/D.

A Figura 2.9 apresenta a evolução da energia do comando "direita" dividido em janelas de 10 ms.

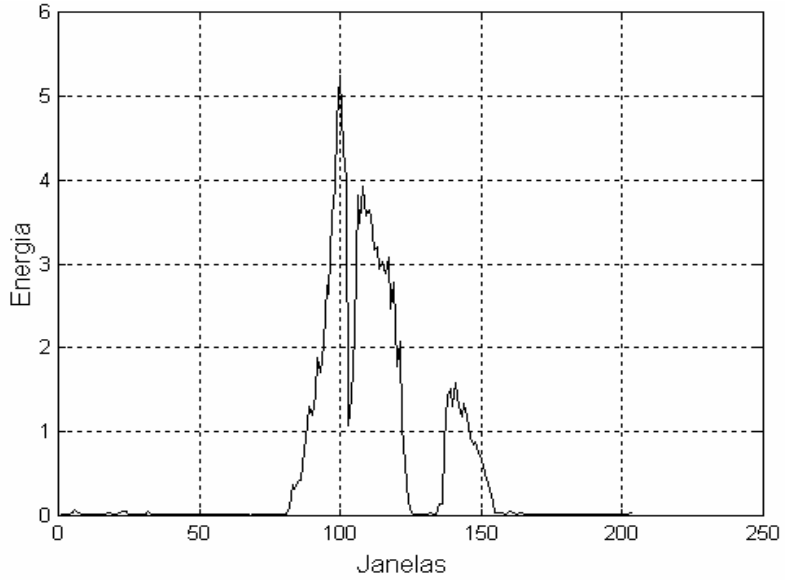


FIGURA 2.9 - Evolução da energia do comando direita

Uma medida que se mostra mais significativa no processo de reconhecimento do que a energia absoluta, é a energia diferencial, pois ela fornece informações sobre as variações relativas à amplitude do sinal[DEL 93]. A eq. 2.8 apresenta a energia diferencial:

$$DE[n] = E[n+\delta] - E[n-\delta] \tag{2.8}$$

onde δ é um valor inteiro escolhido arbitrariamente.

2.3.8 Taxa de Cruzamento por Zero

A medida da taxa de cruzamento por zero (*Zero-crossing Rate*-ZCR), assim como a energia, é uma forma simples de representação do sinal. O cruzamento por zero ocorre se sucessivas amostras têm diferentes sinais algébricos. A taxa de cruzamento por zero não é nada além do que a medida da freqüência contida em um sinal de voz.

Uma definição da ZCR é apresentada em eq. 2.9

$$ZCR = \sum_{m=-\infty}^{\infty} |\, Sinal[x(m)] - Sinal[x(m-1)]\,|\, w(n-m)\,, \tag{2.9}$$

onde *Sinal[x(n)] = 1*, quando $x(n) \geq 0$ e *Sinal[x(n)] = -1*, quando $x(n) < 0$; e $w(n)=1/(2N)$, para $0 \leq n \leq N-1$ e $w(n)=0$, para demais casos. E $w(n)$ é uma janela retangular[RAB 78].

É necessário que as amostras sejam verificadas em pares para determinar onde o cruzamento por zero ocorre e então computar a média sobre as *N* amostras consecutivas.

Resumindo, a taxa de cruzamento por zero fornece o número de vezes que ocorre o cruzamento em um determinado intervalo de tempo. Esta taxa pode ser utilizada para detecção dos limites da palavra, assim como pode avaliar se certo segmento do sinal de voz é vocálico ou não.

O modelo para produção de fala sugere que a energia dos sons vocálicos está concentrada nas freqüências abaixo de 3 KHz, enquanto que nos sons não-vocálicos esta energia encontra-se nas altas freqüências, o que implica que se ZCR for baixo podemos classificar o som como vocálico, caso contrário como não-vocálico.

Assim como a estimação da energia, este método é extremamente sensível à presença de ruído no sinal. Sendo necessário um ambiente adequado para a captura dos sinais e um microfone com boa qualidade para obter resultados satisfatórios.

Além disso, é necessário tomar cuidado no processo de amostragem, pois a taxa de cruzamento por zero é bastante afetada pela conversão A/D, pelo acoplamento de 60 Hz da rede do sinal, como também por qualquer ruído que possa estar presente na digitalização. Desta forma, a utilização de um filtro passa banda antes de um filtro passa baixa pode ser necessário antes da amostragem, para que o sinal não seja afetado por este tipo ruído[DIN 97].

Um exemplo da evolução da taxa de cruzamento por zero pode ser visualizado na fig. 2.10.

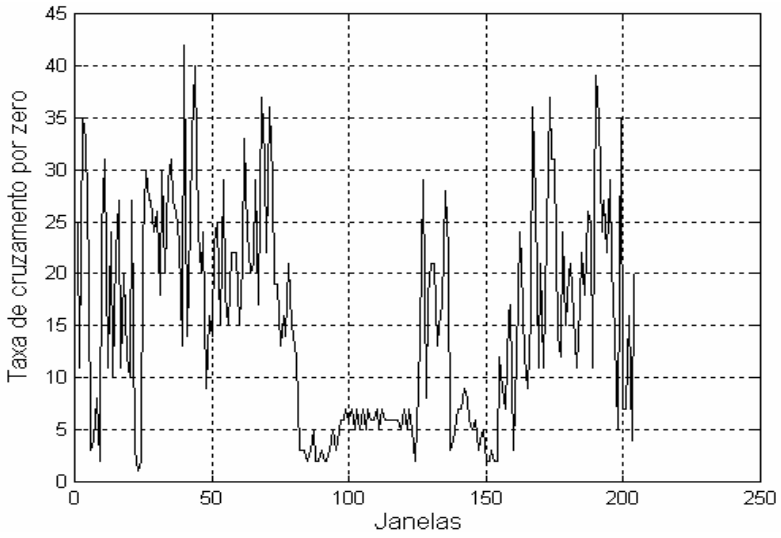


FIGURA 2.10 - Evolução da taxa de cruzamento por zero do comando direita

2.3.9 Análise Cepstral

Após as técnicas já descritas serem aplicadas ao sinal sonoro, é possível a aplicação da análise cepstral, para que as informações mais relevantes do sinal sejam extraídas de forma que representem as principais características do sinal original.

Para a extração dos coeficientes cepstrais é necessária a definição e conhecimento da Transformada de Fourier, que é um método matemático que possibilita a conversão do domínio em que o sinal sonoro está contido, a fim de tornar possível a sua manipulação. Ou seja, o sinal discreto no domínio tempo é convertido para o domínio freqüência, no sentido de possibilitar a avaliação de suas características acústicas.

O cálculo da Transformada Discreta de Fourier (*Discrete Fourier Transform*-DFT) é amplamente utilizado em aplicações computacionais para estimativa de espectros, funções de correlação e para implementação de filtros digitais[HEL 67][STO 66]. Mas sua principal função é transformar um sinal discreto no tempo em um sinal discreto no domínio freqüência.

Sejam os valores *x(k)*, que podem ser reais ou complexos, de uma série temporal de *N* pontos. Os coeficientes *x(n)* da DFT dos pontos *x(k)* são definidos por:

$$x(n) = \sum_{k=0}^{N-1} x(k) e^{-2\,jnk\pi / N} \tag{2.10}$$

onde, k=0,1,...,N-1.

Esta notação pode ser simplificada se definir:

$$w = e^{-2\,j\pi / N} \tag{2.11}$$

Com isso, os coeficientes da DFT são calculados pela expressão:

$$x(n) = \sum_{k=0}^{N-1} x(k) w^{nk} \tag{2.12}$$

como geralmente é encontrada.

O algoritmo da Transformada Rápida de Fourier (*Fast Fourier Transform*-FFT) é um método para computar a Transformada Discreta de Fourier de uma série de *N* pontos, de ordem N^2 operações - modo direto para $N \log_2 N$ operações. Portanto, a FFT é um algoritmo que possibilita o cálculo da DFT de uma série temporal mais rapidamente do que outros algoritmos existentes, tornando o emprego da DFT de uma importância notável. Este método não somente reduz o tempo de computação, como também os erros de arredondamento associados a este cálculo.

Para fornecer uma visão de tal eficiência é apresentado um exemplo da redução que este algoritmo propicia. Supondo-se que N=1024. Os métodos tradicionais para

estimação requerem um número de operações proporcionais a N^2, ou seja, 1.048.756 operações, isto é, 100 vezes maior que o exigido pela FFT (fig. 2.11).

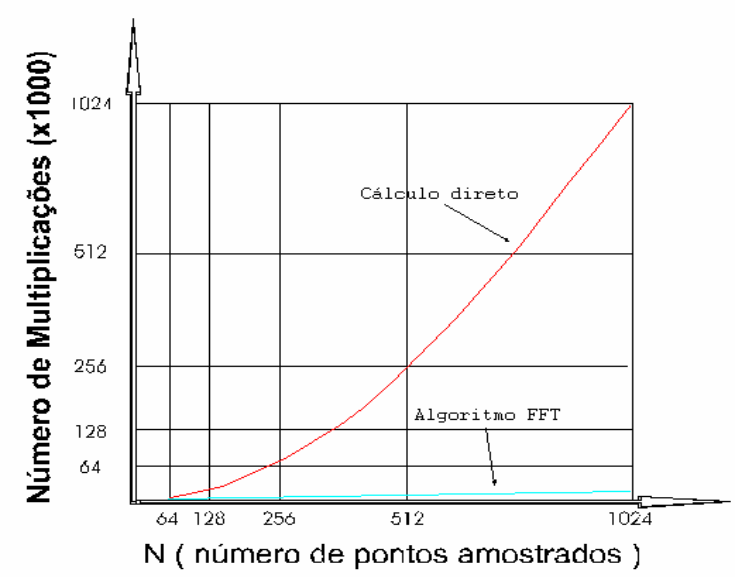


FIGURA 2.11 - Comparação de multiplicações requeridas pelo cálculo direto e o requisitado pela FFT

É possível também o cálculo da transformada inversa a partir da FFT, bastando para isto substituir os pontos pelos seus conjugados e multiplicar os resultados obtidos por *1/N*[RAB 78].

A partir desta breve descrição da Transformada de Fourier, que permite a representação do sinal discreto no domínio freqüência, já é possível se descrever a extração dos coeficientes cepstrais pela análise homomórfica, como é considerada esta metodologia para separação do sinal de excitação e o trato vocal.

Os sinais de fala são produzidos pela convolução do sinal de excitação *g(t)* com o sinal do trato vocal *h(t)*, como já visto no item 2.2 deste capítulo (fig. 2.4). O processo de deconvolução, ou análise cepstral, é responsável pelo caminho inverso, ou seja, a separação dos sinais componentes da fala.

Sabendo-se que o sinal de voz produzido pelo modelo simples de produção da voz, possui a seguinte notação:

$$s(t) = g(t) * h(t), \tag{2.13}$$

lembrando que "*" representa a convolução.

Aplicando-se a análise de Fourier sobre *s(t)*, *g(t)* e *h(t)*, respectivamente transformando-se assim em multiplicação:

$$s(w) = g(w)h(w). \tag{2.14}$$

E tomando-se o logaritmo dos módulos da eq. 2.14 é obtida uma soma de sinais no domínio freqüência o que resulta em:

$$\log|s(w)| = \log|g(w)| + \log|h(w)| \qquad (2.15)$$

Sendo que os coeficientes cepstrais (cepstros) de *s(t)*, são adquiridos por:

$$c(r) = F^{-1}\log|g(w)| + F^{-1}\log|h(w)| \qquad (2.16)$$

onde F^{-1} representa a Transformada Inversa de Fourier (*Inverse Discrete Fourier Transform*-IDFT) e *r* é o índice.

Quando se calculam os coeficientes cepstrais para uma janela de *N* amostras é utilizada a eq. 2.17[FUR 89][OPP 75].

$$c[n] = \frac{1}{N}\sum \log|x(k)|e^{j2kn\pi/N}, \qquad 0 \le n \le N \qquad (2.17)$$

Como a fonte de excitação do sinal varia mais rapidamente que a resposta impulsiva do trato vocal, os dois sinais que compõem a fala estarão distantes um do outro, sendo possível a sua distinção.

Os primeiros coeficientes cepstrais estão diretamente relacionados com a função de transferência do trato vocal, também denominada como envelope espectral, e podem ser usados para representar o sinal de voz. Geralmente, os 20 primeiros coeficientes servem para tal representação, uma vez que estes coeficientes não possuem informações relativas à fonte de excitação *g(t)* a qual é bastante variável, o que facilita o tratamento dos dados para a etapa de reconhecimento.

O processo para extração dos coeficientes cepstrais é apresentado na fig. 2.12, tornando-se clara sua aplicação.

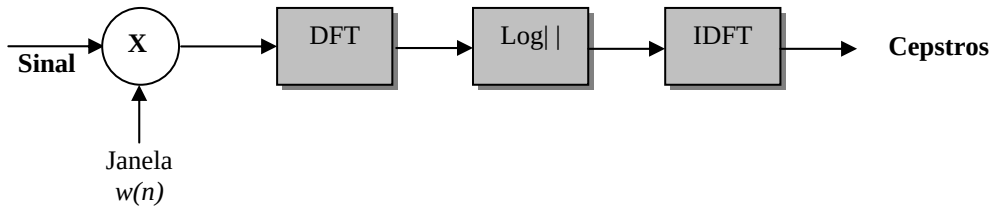


FIGURA 2.12 - Processo para obtenção dos cepstros

Resumidamente, como visualizado na figura 2.12, a análise cepstral é determinada como a Transformada Inversa de Fourier do logaritmo ao módulo da amplitude da Transformada de Fourier.

3 Redes Neurais

A computação neural surgiu inspirada no funcionamento do cérebro. O termo neural está relacionado com as redes estudadas na neurociência, as quais serviram como motivação para os modelos atualmente utilizados[HER 91]. O foco das pesquisas nesta área está voltado para o que fazem e como se comportam as redes neurais artificiais perante um determinado problema. Embora algumas redes apresentem similaridades com o modelo biológico, o mesmo segue uma área de estudos diferente desta e que está bastante relacionada ao funcionamento do cérebro.

3.1 Motivação

Um dos principais fatores que leva ao estudo das RNAs é o fato de o cérebro humano realizar um grande número de tarefas, como reconhecimento de imagens e voz, com uma capacidade muito superior a de um computador.

Porém, uma vez comparadas as características do cérebro com os computadores digitais algumas curiosidades merecem destaque.

O tempo de processamento de um neurônio é de aproximadamente 1ms, enquanto que um simples computador doméstico é capaz de trabalhar a 1GHz, isto representa a execução de uma instrução a cada 2ns[SIM 90]. Desta forma, coloca-se a dúvida de como o cérebro é tão superior na tarefa de reconhecimento, tendo em vista que a máquina possui um grande poder de processamento sobre a informação. Isto pode ser justificado pela característica do cérebro em possuir um processamento paralelo e distribuído inexistente nos computadores atuais.

Outra característica relevante é que o cérebro humano possui um número estimado de 10^{11} à 10^{14} neurônios, cada um com cerca de 10^{3} à 10^{4} conexões, representando uma unidade completa de computação[KOV 96]. Os computadores domésticos, por sua vez, possuem apenas uma unidade de processamento central.

Ainda se não bastasse, deve ser exposta a questão do armazenamento do conhecimento em ambos os sistemas, uma vez que em um o conhecimento é armazenado em um local endereçável, caso dos computadores, e no outro está armazenado sob uma forma dispersa e adaptativa, onde não é permitido que novas informações aprendidas pelo cérebro, eliminem outras já existentes. Outro tópico relativo ao armazenamento é que a forma distribuída como o cérebro guarda as informações torna possível regenerar um conhecimento global a partir de apenas uma de suas partes. Isto faz do cérebro um sistema tolerante a falhas, uma vez que removida uma de suas partes o conhecimento global não é afetado de forma signifiaiva.

Desta forma, as redes neurais são utilizadas e estudadas com a intenção de explorar as características de paralelismo e processamento altamente distribuído do cérebro.

3.2 Inspiração Biológica

O neurônio tem um corpo celular chamado de soma, e diversas ramificações conhecidas como dendritos, que são responsáveis por conduzir sinais das extremidades para o corpo celular. Outra ramificação existente, geralmente única, é o axônio, que transmite um sinal do corpo celular para suas extremidades. A grosso modo, os dendritos recebem as informações, ou impulsos nervosos, provenientes de outros neurônios, conduzindo-os ao soma, no corpo celular a informação é processada, produzindo novos impulsos que são passados através do axônio até os dendritos dos nós seguintes. O ponto de contato entre a extremidade axônica do neurônio e o dendrito de outro é chamado de sinapse. É pelas sinapses que os neurônios se unem constituindo as redes neurais. A fig. 3.1 apresenta um modelo do neurônio biológico.

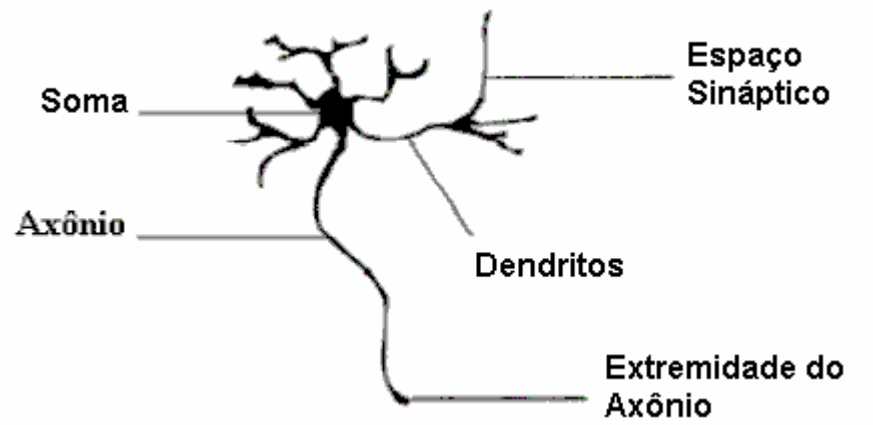


FIGURA 3.1 - Neurônio Biológico

3.2.1 Potencial de Ação

A ação da membrana dos neurônios é que cria a habilidade de produzir e transmitir os sinais, envolvida no exterior do corpo do neurônio possuí a capacidade de gerar impulsos nervosos. O corpo é responsável por combinar os sinais recebidos, avaliando se o valor resultante está acima de um limiar de excitação do neurônio, caso esteja um impulso elétrico é produzido e propagado através do axônio para os nós seguintes.

A diferença de potencial (em *volts*) entre o interior e o exterior do neurônio, resultante da diferença de concentração de potássio (interna à célula) e sódio (externa à célula) é o que ocasiona o disparo do neurônio. A concentração de íons de potássio dentro da célula produz um potencial elétrico de –70 mV (potencial de repouso) em relação ao exterior, quando os impulsos das sinapses reduzem este nível para cerca de - 50 mV é que o fluxo de sódio e potássio é invertido, tornando o interior da célula positivo em relação ao exterior. Isto faz com que o impulso nervoso seja transmitido pelo axônio até suas conexões sinápticas[BRA 98]. Ao chegar ao terminal de um axônio, os canais controlados se abrem, permitindo a liberação de moléculas de vários tipos com o nome genérico de neurotransmissores que se difundem no espaço entre o terminal do axônio e o dendrito de outro neurônio. Dependendo do tipo de neurotransmissores liberados a sinapse poderá ser excitatória ou inibitória.

3.3 O que são as RNAs ?

As Redes Neurais Artificiais são sistemas paralelos e distribuídos compostos por unidades de processamento simples, neurônios (artificiais) ou nós, responsáveis por realizar determinadas funções matemáticas (geralmente não lineares). Os nós estão dispostos em uma ou mais camadas e interligados por um grande número de conexões, quase sempre unidirecionais. A maioria dos modelos apresenta conexões que estão associadas a pesos, que armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida da rede.

A maneira como tais redes solucionam problemas é bastante interessante, superando o desempenho de modelos convencionais, tanto por apresentar um paralelismo natural como pela forma interna como é representada. Inicialmente, a rede é submetida a uma fase de aprendizado, ou seja, exemplos são apresentados à rede, que extrai características necessárias com a finalidade de representar a informação fornecida e posteriormente produzir respostas relativas ao problema apresentado.

As RNAs não atuam somente como mapeadores de funções de entrada e saída; a sua capacidade de aprendizado e generalização da informação, que está associada à capacidade de a rede aprender através de um conjunto reduzido de exemplos, fornecendo respostas coerentes para dados não conhecidos, torna as RNAs uma ferramenta computacional poderosa e interessante na solução de problemas complexos[BRA 98].

3.4 O Surgimento das RNAs

O primeiro modelo de neurônio foi desenvolvido por McCulloch e Pitts em 1943 [MCC 43]. Este modelo se concentrou mais em descrever um modelo de um neurônio artificial e em apresentar suas capacidades computacionais, ao invés de tratar técnicas de aprendizado.

Em 1949 Hebb, observando as mudanças nas sinapses dos neurônios, desenvolveu a “Teoria do Aprendizado Neural”, onde determinou que a conexão entre duas unidades ativadas ao mesmo tempo é reforçada[HEB 49]. A Regra de Hebb, como é conhecida a sua teoria na comunidade de RNAs, foi interpretada do ponto de vista matemático e é utilizada atualmente em vários algoritmos de aprendizado. Outra regra de aprendizado, baseada no método do gradiente para a minimização do erro na saída de um neurônio com resposta linear, surgiu na década de 60, foi elaborada por Widrow e Hoff e ficou conhecida como Regra Delta.

Frank Rosemblatt apresentou em 1958 o seu modelo *Perceptron*, onde as RNAs com nós MCP (originários de McCulloch e Pitts) podiam ser treinadas para classificar certos tipos de padrões. O *Perceptron* mais simples descrito por Rosemblatt possui três camadas: a primeira recebe as entradas do exterior e possui conexões fixas (*fotoperceptron*), a segunda recebe os impulsos da primeira por meio das conexões, ocorrendo um ajuste nos pesos e por fim os valores são repassados à terceira camada produzindo uma resposta. Marvin Minsky e Seymour Papert mostraram algumas limitações dos *Perceptrons* que somente solucionavam uma classe de problemas, e que

problemas não linearmente separáveis como a função XOR (ou exclusivo) não eram capazes de serem computados pelo *Perceptron* elementar[MIN 69]. Estes resultados e observações feitas por Minsky e Papert deixaram as pesquisas relativas às RNAs em segundo plano durante a década de 70 até o inicio dos anos 80. Entretanto, surgiu um novo paradigma que buscava a simulação do cérebro humano do ponto de vista físico para simular a atividade mental.

Em 1982, John Hopfield publicou um artigo que chamou a atenção para as propriedades associativas das RNAs, o que incentivou as pesquisas na área. Hopfield mostrou a relação entre redes recorrentes auto-associativas e sistemas físicos. Mais tarde, em 1986, a impotência das redes *Perceptron* na solução do problema de associação de padrões para um conjunto de padrões não linearmente separável foi eliminada por Rumelhart, Hilton e Williams, com a utilização da Regra Delta Generalizada, surgindo a rede *Backpropagation*[RUM 86].

Um fator importante que forneceu interesse à área foi sem dúvida o avanço da tecnologia, sobretudo a microeletrônica, a qual vem permitindo a realização física de modelos de nós e sua interconexão de modo nunca proposto[BRA 98].

3.5 Neurônio Artificial

O modelo geral de neurônio mostrado na fig. 3.2, criado por McCulloch e Pitts, é representado para se observar o funcionamento básico do neurônio artificial.

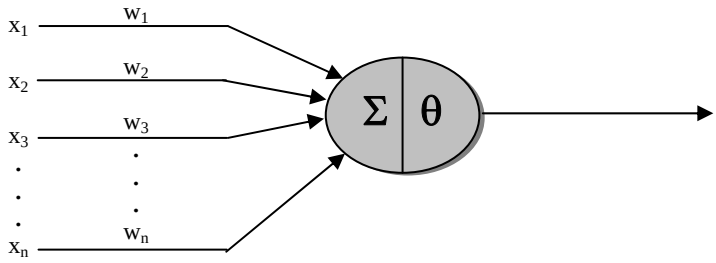


FIGURA 3.2 - Neurônio Artificial de McCulloch e Pitts

Este modelo é uma simplificação do neurônio biológico, sua descrição matemática resultou em um modelo com n terminais de entrada $x_1,x_2,x_3,...,x_n$ (representando os dendritos) e um terminal de saída (correspondente ao axônio). Com a finalidade de emular o comportamento das sinapses, os terminais de entrada possuem pesos acoplados $w_1,w_2,w_3,...,w_n$, cujos valores podem ser positivos ou negativos, conforme as sinapses forem inibitórias ou excitatórias. Os pesos determinam quando o neurônio deve considerar sinais de disparo que ocorrem naquela conexão.

A ativação do neurônio é obtida através da aplicação de uma "função de ativação", que habilita ou não a saída, dependendo do valor da soma ponderada das suas entradas. No modelo MCP, a função de ativação é dada pela função de limiar da eq. 3.1, e o neurônio terá sua saída ativada quando:

$$\sum_{i=1}^{n} x_i w_i \geq \theta \,, \qquad (3.1)$$

onde n é o número de entradas do neurônio, w_i é o peso associado à entrada x_i e θ é o limiar de ativação do neurônio.

3.6 Funções de Ativação

Diversas são as funções de ativação que podem ser aplicadas aos nós para produzir uma saída qualquer e não necessariamente zero ou um. A fig. 3.3 apresenta algumas funções utilizadas.

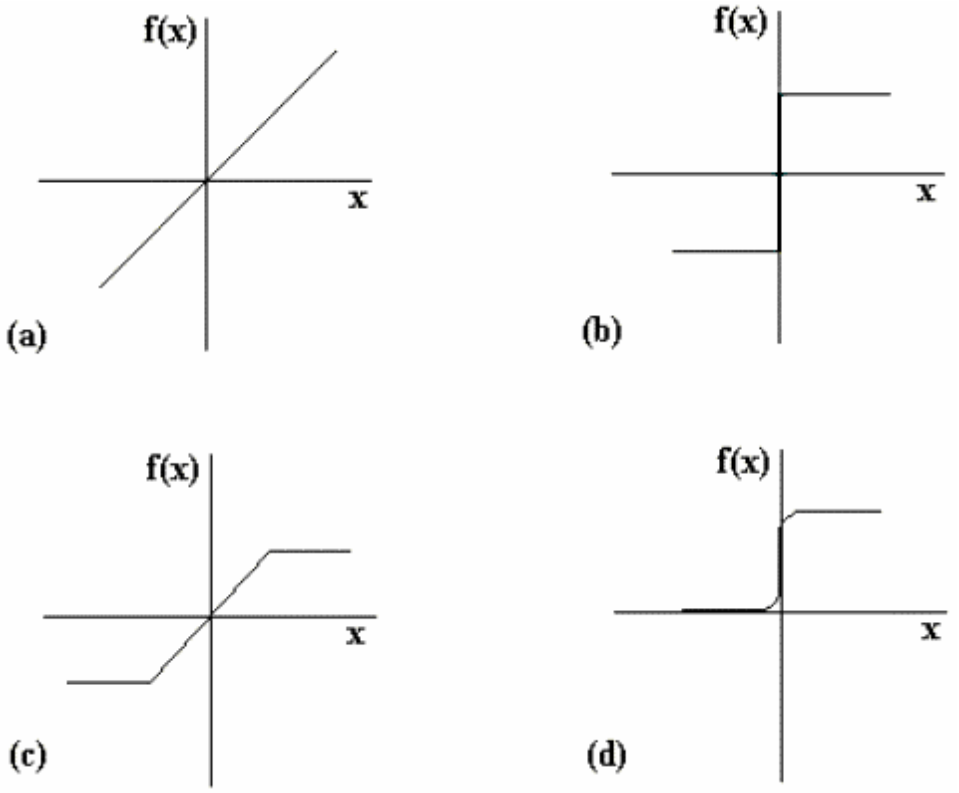


FIGURA 3.3 - Funções de Ativação

A função de ativação linear está representada na Fig. 3.3(a) e é definida por:

$$y = \alpha x \,, \qquad (3.2)$$

sendo α um número real que define a saída linear y, para os valores de entrada x.

A função degrau, fig. 3.3(b), produz uma saída + φ para os valores de x maiores que zero e – φ caso contrário. Está função está definida por:

$$y = \begin{cases} +\varphi, & \text{se } x > 0 \\ -\varphi, & \text{se } x \leq 0 \end{cases} \tag{3.3}$$

A função rampa é uma derivação da função linear, podendo produzir valores em uma faixa de [-φ, +φ], como representada na eq. 3.4 e visualizada na fig. 3.3(c).

$$y = \begin{cases} +\varphi, & \text{se } x \geq +\varphi \\ x, & \text{se } |x| < +\varphi \\ -\varphi, & \text{se } x \leq -\varphi \end{cases} \tag{3.4}$$

E por último é apresentada uma função sigmoide em forma de “S” na fig. 3.3(d), cobrindo o intervalo [0,+1]. Comparativamente à função rampa, ela aumenta o poder discriminante da rede em problemas de separação de classes de padrões. Dentre as principais características atribuídas destaca-se o fato de ser uma função contínua, monotônica, não linear e diferenciável em qualquer ponto, tornando este tipo de função um dos mais utilizados pelas redes neurais, juntamente com a função tangente hiperbólica que se diferencia desta pela extensão de seu intervalo em [-1,+1].

3.7 Aprendizado

A aprendizagem das redes neurais é definida como o processo pelo qual os parâmetros da rede são ajustados através de uma forma continuada de estímulo, pelo ambiente no qual a mesma está operando; o tipo de aprendizagem realizada é definido pela maneira particular como ocorrem os ajustes realizados nos parâmetros[MEN 70].

O aprendizado pode ser definido como uma mudança de comportamento nos procedimentos de treinamento. O desempenho da rede é medido antes e depois do treinamento, sendo que a diferença na medida de desempenho será o fator responsável por indicar o quanto a rede aprendeu[LOE 96].

A maneira conforme ocorre o ajuste nos parâmetros da rede pode ser feita sob dois modos diferente:

Por Lote (Batch): onde os parâmetros são atualizados ao final de cada época, ou seja, após todo o processamento de um conjunto de observações é que ocorrerão as devidas atualizações nos parâmetros. Este modo não possibilita que a ordem de apresentação dos padrões exerça grandes influências no treinamento, mas ocasiona uma redução na velocidade de treinamento da rede;

Incremental: o ajuste nos parâmetros ocorre a cada observação processada pela rede, importando a ordem em que os padrões são apresentados, influenciando na velocidade em que a rede convergirá, ou melhor, terá aprendido.

Os métodos para treinamento estão divididos em duas abordagens que são vistas nos itens subsequentes.

3.7.1 Aprendizado Supervisionado

A razão deste método possuir tal nome está ligada ao seu funcionamento, onde a entrada e a saída desejadas para a rede são fornecidas por um "supervisor", responsável por indicar explicitamente um comportamento bom ou ruim, ou seja, seu papel é ajustar os parâmetros da rede, no sentido de encontrar uma ligação entre os pares de entrada e a saída fornecida. A fig. 3.4 mostra o funcionamento do aprendizado supervisionado. Quando um padrão de entrada é apresentado para a rede, a saída desejada é comparada com a resposta calculada, e os pesos sofrem um ajuste no sentido de minimizar uma medida de erro.

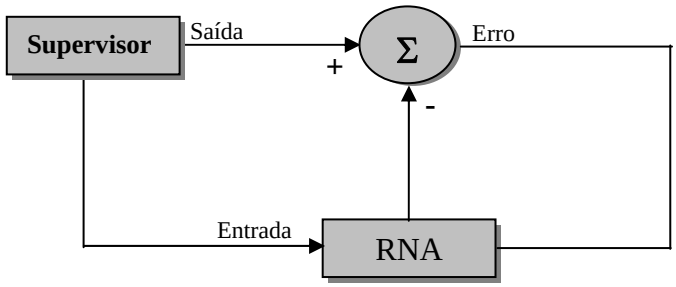


FIGURA 3.4 - Aprendizado Supervisionado

Na implementação deste tipo de aprendizado o treinamento pode ser tratado sob duas formas: o *off-line*, que uma vez treinados os dados para resolução de determinado problema os mesmos tornam-se fixos, desde que não sejam adicionados novos dados ao conjunto de treinamento, o que implicaria em um novo treinamento; e o *on-line*, onde o conjunto de dados muda continuamente, o que faz a rede permanecer sempre em processo de adaptação.

3.7.2 Aprendizado Não-Supervisionado

Este método é facilmente diferenciado do anterior devido ao fato de não apresentar um "supervisor", como mostrado na fig. 3.5. No aprendizado não-supervisionado não são usadas informações relativas sobre se a resposta da rede foi correta ou não, no sentido de ajustar conexões. Mas, sim, se a rede deve responder de modo semelhante a exemplos semelhantes, ou seja, no momento que a rede estabelece uma harmonia com regularidades estatísticas de entrada, ela torna-se capaz de formar representações internas para codificar características da entrada e produzir novas classes automaticamente. Com isto conclui-se que este método somente é possível com a existência de dados redundantes na entrada da rede.

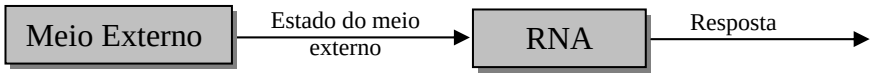


FIGURA 3.5 - Aprendizado Não-Supervisionado

3.8 Caracterização Geral das RNAs

Existem diversas arquiteturas que comportam as RNAs, a definição da mais adequada está diretamente associada ao tipo de problema a ser solucionado, visando um melhor arranjo na representação das interconexões entre os neurônios.

Na forma geral, grupos de neurônios estão distribuídos em camadas dentro da estrutura da rede, freqüentemente dispostas em uma camada de entrada, seguida de uma ou mais camadas intermediárias, terminando em uma camada de saída. Quanto maior o número de camadas maior será a complexidade e o tempo de processamento da rede. A definição do número de neurônios na camada de entrada está relacionada ao tamanho dos vetores a serem apresentados para a rede, assim como a quantidade de neurônios na saída está diretamente ligada ao número desejado de classes que a rede deverá classificar. Quanto a quantidade de neurônios que deve existir na camada escondida (intermediária) não há uma definição especifica, estando esta mais relacionada com o tipo de problema que está sendo tratado e ao comportamento que a rede apresentar, ressaltando que um maior número de neurônios por camada propicia um aumento no grau de liberdade da função de transferência, implicando no acréscimo de variáveis livres e consequentemente na redução da capacidade de generalização da rede.

A conectividade dos neurônios pode ser estruturada de duas formas:

Para Frente (*Feedforward*): esta estrutura de conexão entre os neurônios se caracteriza por não permitir que a saída de um neurônio na *i-ésima* camada da rede seja utilizada como entrada dos neurônios das camadas com índice menor ou igual a *i*; a mesma segue o sentido entrada-saída e completamente conectada, possuindo as saídas de todos os neurônios de uma referida camada ligados a todos os neurônios da camada posterior, conforme a fig. 3.6.

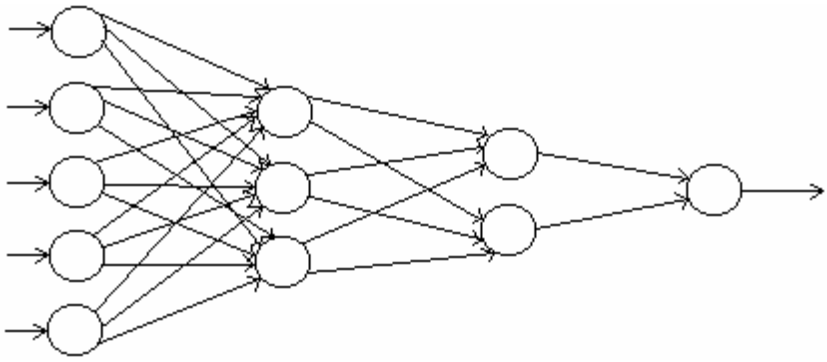


FIGURA 3.6 - Rede *Feedforward* com 4 camadas

Recorrente: a diferença na conexão dos neurônios das redes recorrentes, ao contrário das redes *feedforward*, é que a saída de algum neurônio na *i-ésima* camada pode ser utilizada como entrada dos neurônios com índice menor ou igual a *i*; não existindo um sentido único para as conexões que podem existir entre neurônios de mesma camada como visualizado na fig. 3.7.

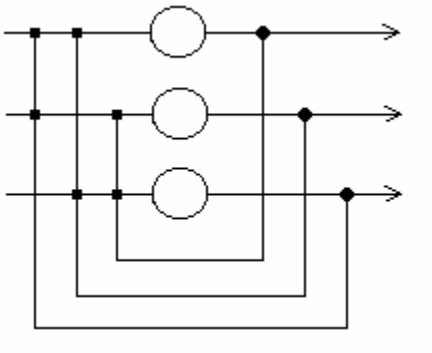


FIGURA 3.7 - Rede Recorrente de uma camada

Independente das diferentes arquiteturas que as RNAs apresentam sua forma básica é composta por uma ou mais camadas que geralmente recebem uma matriz de pesos (*W*), um vetor de polarização ("*bias*"), uma função de ativação, uma função de propagação que pode ser a mesma ou não em todas as camadas da rede e um vetor de saída. Com relação a camada de entrada é importante especificar que sua utilização refere-se mais a uma formalização da rede, quando utilizada, servindo somente para a passagem dos dados que entram na rede.

3.9 A Rede *Backpropagation*

Inicialmente o termo *Backpropagation* dizia respeito ao algoritmo utilizado para o aprendizado da rede. Em virtude da sua expansão e uso seu nome foi incorporado a denominação da própria rede.

3.9.1 Arquitetura

A arquitetura é baseada em um projeto hierárquico, formado por camadas que contêm vários elementos de processamento (neurônios). O número de camadas e neurônios que fazem parte da rede é determinado de acordo com a aplicação a que se destina. A quantidade de elementos de processamento da primeira camada (entrada) está diretamente relacionada à dimensão do espaço de entrada, da mesma forma que a camada de saída está com o espaço de saída. As demais camadas da rede não possuem uma regra fixa que estabeleça um número de neurônios, variando de forma empírica até a obtenção de resultados satisfatórios.

Segundo o Teorema de Kolmogorov, o objetivo da rede é mapear uma função $y=f(x)$ de R^m em R^n, através de um conjunto de treinamento de k pares de vetores $(x_1,y_1),...,(x_i,y_i),...,(x_k,y_k)$, onde para cada vetor de entrada x_i de m elementos, existe um vetor de saída correspondente y_i de n elementos[HEC 90].

3.9.2 Funcionamento

Quando apresentado à rede um padrão de entrada (x_i), este é propagado pela(s) camada(s) intermediária(s) até chegar a camada de saída, onde a rede apresentará a sua

resposta (o_i), correspondente ao estímulo. O padrão de saída calculado pela rede (o_i) é comparado ao padrão de saída desejado (y_i), sendo então calculado o erro para cada unidade de saída. A função de erro que deve ser minimizada é apresentada na eq. 3.5, e é conhecida como Erro Médio Quadrado (EMQ).

$$E = \frac{1}{2}\sum_{p=1}^{M}\sum_{i=1}^{N}\left(y_{p,i} - o_{p,i}\right)^2, \tag{3.5}$$

onde:

M = número de padrões ou vetores de entrada;
N = número de neurônios na saída ou dimensão do vetor de saída;
$y_{p,i}$ = saída desejada do *i-ésimo* neurônio, para o *p-ésimo* padrão apresentado;
$o_{p,i}$ = saída obtida pela rede do *i-ésimo* neurônio, para o *p-ésimo* padrão apresentado.

O erro calculado é retropropagado para os neurônios da camada intermediária imediatamente anterior à camada de saída. O valor recebido por cada unidade é proporcional à sua contribuição para o sinal de saída e consequentemente para o erro total. Este valor é diretamente influenciado pelo peso (W_{ij}) associado a conexão entre o elemento da camada *i* e o da camada de saída *j*; a regra de aprendizado atualiza os pesos das conexões por meio do erro produzido.

3.9.3 Regra de Aprendizado

A regra de aprendizado utilizada na rede *Backpropagation* busca obter pesos que permitam a realização de mapeamentos não-lineares entre os padrões de entrada e saída, e é conhecida como Regra Delta Generalizada, em virtude de ser uma extensão da Regra Delta.

A Regra Delta foi criada por Widrow e Hoff, que a aplicaram a um algoritmo adaptativo, *Least Mean Square*-LMS, destinado a redes neurais de apenas duas camadas com unidades de saída lineares. O objetivo principal desta regra é ajustar os pesos no sentido de minimizar o erro médio quadrático total.

A limitação em aplicar-se a Regra Delta sobre redes com mais de duas camadas surgiu devido ao fato desta ser baseada na diferença entre os valores reais e os desejados para o ajuste dos pesos. Como em redes com mais de duas camadas os valores produzidos pelos neurônios da camada intermediária são desconhecidos, torna-se impraticável o cálculo da diferença entre a saída real e a desejada.

Em busca da superação dessa limitação é que foram intensificados os estudos sobre redes com uma ou mais camadas, surgindo assim a Regra Delta Generalizada, a qual utiliza o método de descida do gradiente em busca da minimização do erro, ao contrário da Regra Delta, onde o ajuste dos pesos é proporcional a diferença entre a saída desejada e a real. A sua generalização propõe um passo na direção do gradiente descendente em busca do mínimo da função de erro, conforme a eq. 3.6 os pesos são alterados da seguinte maneira:

$$\Delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}}, \tag{3.6}$$

onde, η é a taxa de aprendizado e $\frac{\partial E}{\partial W_{ij}}$ é a derivada parcial do erro em relação ao peso da respectiva conexão.

A fórmula final da Regra Delta Generalizada é dada pela eq. 3.7 baseada na eq. 3.6:

$$\Delta W_{ij} = \eta \delta_i X_i, \tag{3.7}$$

em que δ_i é o gradiente local associado a unidade *i* e X_i é a entrada da unidade *i*. Estes valores são referentes a um padrão *p* específico, podendo a equação 3.7 ser representada por:

$$\Delta_p W_{ij} = \eta \delta_{pi} X_{pi} \tag{3.8}$$

O gradiente local é calculado conforme a unidade *i* ser referente ou não a camada de saída, o cálculo do erro está definido na eq. 3.9 segundo a unidade a que ele pertence.

$$\delta_{pi} = \begin{cases} \left(o_{pi} - y_{pi}\right) f_i'\left(net_{pi}\right), & \text{se } i \in \text{à camada de saída} \\ \left(\sum_k \delta_k W_{ki}\right) f_i'\left(net_{pi}\right), & \text{se } i \notin \text{à camada de saída} \end{cases} \tag{3.9}$$

Onde y_{pi} corresponde a saída desejada da unidade *i* correspondente ao padrão *p*. O erro é calculado recursivamente para as unidades que não estão na camada de saída, em que o somatório $\sum_k$ representa a contribuição de todos os neurônios da camada subsequente. O valor de net_{pi} refere-se à entrada líquida do neurônio (soma das entradas ponderadas pelos pesos) e *f(.)* é a sua função de transferência. *f '(.)* é a derivada de *f* em relação ao seu argumento.

3.9.4 O Fator de Convergência

A convergência é alcançada quando a rede atinge o seu objetivo, ou seja, o mínimo global. Embora a Regra Delta Generalizada tenha superado certas limitações apresentadas por outros métodos aplicados para o mesmo fim, é possível ainda a sua permanência em mínimos locais, provocando oscilações nos pesos como também o abandono do processo de treinamento da rede[SIM 90].

No sentido de evitar que durante o treinamento a rede permaneça presa em mínimos locais, alguns fatores como taxa de aprendizado, número de neurônios na camada intermediária, número de camadas intermediárias e o número de padrões utilizados no treinamento merecem uma atenção especial. A aplicação das RNAs está muito relacionada à heurística, onde a experimentação ainda é o melhor fator para a determinação dos parâmetros para o treinamento da rede.

Algumas técnicas que já foram aplicadas e obtiveram resultados satisfatórios relatam a alteração do modelo original do *Backpropagation*, como: modificação na arquitetura (*Backpropagation* recorrente), inclusão do termo de momento ou ainda da função de custo.

O treinamento do *Backpropagation* permanece sendo baseado ou no decréscimo do erro a níveis fixados, ou a execução de um determinado ciclo de treinamento iterativo ou mesmo quando a rede atingir um estado em que todos os padrões de treinamento estejam classificados corretamente.

3.9.5 Os Parâmetros Aprendizado e Momento

No algoritmo de aprendizado do *Backpropagation* a definição de dois parâmetros merece destaque pela influência que exercem no treinamento desta rede:

- taxa de aprendizado: sua definição está relacionada com a atualização dos pesos, uma vez que quanto menor for a taxa de aprendizado, menores serão as variações nos pesos, tornando o treinamento lento e propício a sua estabilidade em mínimos locais. Por outro lado, elevadas taxas de aprendizado conduzem à saturação ou mesmo à oscilação. A definição desta taxa é um fator importante, pois permite agilidade na convergência do erro desejado, como também para evitar que a rede oscile. É importante observar que se durante o treinamento o Erro Médio Quadrado aumentar, a taxa de aprendizado deve ser reduzida, ocasionando um decaimento rápido; caso contrário, se o erro venha a diminuir, a taxa de aprendizado poderá ser elevada gradualmente para que o mínimo global não seja ultrapassado;

- termo de momento: sua principal função é acelerar o aprendizado sem produzir oscilações. A probabilidade da convergência esbarrar em mínimos locais é reduzida, uma vez que o termo ignora as variações de alta freqüência na superfície do erro. A inclusão do momento se baseia em fazer com que as mudanças nos pesos das conexões sejam somadas a uma fração da última alteração nestes pesos determinada pela regra de aprendizagem. Desta forma, se a alteração anterior foi realizada num determinado sentido da superfície do erro, parte da atualização atual nos pesos das conexões será realizada no mesmo sentido. Conforme expresso na eq. 3.10

$$\Delta W_{ij}(k+1) = -m \cdot \Delta W_{ij}(k) + (1-m) \cdot \eta \delta_{pi} \cdot a_{pj}, \qquad (3.10)$$

onde:

m = termo de momento;
η = taxa de aprendizado;
δ_{pi} = função de erro;
a_{pj} = ativação da *j-ésima* entrada (ou a saída do *j-ésimo* neurônio da camada anterior) para o *p-ésimo* padrão apresentado;

conforme relatado por Thomé apud [DIN 97], quando o momento se reduz a zero, a mudança nos pesos é baseada no gradiente e quando ele apresenta valor unitário a mudança atual é igual à última mudança realizada nos pesos e o gradiente é ignorado.

3.9.6 O Algoritmo *Backpropagation*

Uma definição básica do algoritmo *Backpropagation* para treinamento incremental da rede é descrita nos seguintes passos:

1) Inicialização dos pesos, polarizações, taxa de aprendizado, momento e definições dos critérios de parada;

2) Apresentação para a rede de um padrão de entrada do conjunto de treinamento e computação de sua saída;

3) Cálculo do erro para os neurônios da camada de saída, ou seja, diferença entre a saída desejada e a obtida;

4) Cálculo para atualização dos pesos da camada de saída;

5) Retropropagação do erro para as camadas escondidas. Uma vez que não há uma saída desejada para os neurônios que compõem a(s) camada(s) intermediária(s), o cálculo do erro destes neurônios deve ser obtido a partir dos neurônios pertencentes à camada de saída e das conexões que os interligam;

6) Cálculo do erro acumulado da rede, verificando se o erro médio total de todos os padrões pode ser considerado desprezível, ou seja, se o mesmo atingiu um valor abaixo do limiar estipulado. Caso isto ocorra é deduzido que a rede aprendeu, caso contrário retorna-se ao passo 2.

3.10 Teoria da Ressonância Adaptativa

A rede neural ART (*Adaptive Resonance Theory*), proposta por Carpenter e Grossberg, representa um sistema que auto-organiza padrões de entrada em categorias de reconhecimento. O principal objetivo desta rede é solucionar o problema de plasticidade e estabilidade, ou seja, como um sistema adaptativo pode permanecer flexível quando padrões estranhos estimulam a rede criando novas categorias de reconhecimento e ainda assim continuar estável quando agrupa padrões similares na mesma categoria de reconhecimento. A solução para este dilema foi a implantação de um mecanismo de realimentação entre a camada competitiva e a camada de entrada, permitindo o aprendizado de uma nova informação relevante sem perda do conhecimento já adquirido.

Basicamente a família ART é composta pelos modelos:

ART 1: processa apenas padrões binários;

ART 2: processa padrões contínuos;

ART 3: processa padrões contínuos e aborda a ação de neurotransmissores nos mecanismos de sinapse;

Fuzzy ART: trata os conceitos nebulosos na arquitetura *ART 1*, permitindo o tratamento de padrões analógicos;

ARTMAP: apresenta uma arquitetura preditiva, composta por dois módulos ART;

Fuzzy ARTMAP: apresenta a arquitetura preditiva do *ARTMAP* utilizando conceitos nebulosos.

Neste trabalho é abordado o modelo conexionista *Fuzzy ARTMAP* e desta forma para sua melhor compreensão é descrita também a rede *Fuzzy ART* que constitui a base de funcionamento dos elementos que compõem a *Fuzzy ARTMAP*.

3.10.1 *Fuzzy ART*

O modelo *Fuzzy ART* é muito semelhante ao modelo *ART 1*. A principal diferença entre ambos está na escolha de uma categoria e no processo de aprendizado, onde em *ART 1* é utilizada a tradicional teoria dos conjuntos, enquanto que em *Fuzzy ART* é aplicada a abordagem da lógica difusa. Assim, em *ART 1* é utilizado o operador de intersecção($\cap$), que em *Fuzzy ART* é substituído pelo operador de mínimo($\wedge$). Além disso, uma operação adicional de pré-processamento dos padrões de entrada, conhecida como codificação complementar, é abordada no modelo *Fuzzy*.

Na escolha de categorias são calculados escores de casamento entre os padrões que entram na rede e avaliada uma medida de similaridade entre tais padrões, determinando se o padrão corrente é pertencente ou não a uma determinada classe já classificada, ou se ele deve dar origem a uma nova categoria. No cálculo do casamento entre padrões ocorrem dois processos: um conhecido como *bottom-up* que diz respeito às conexões alimentadas adiante e que fazem o papel de filtro adaptativo, ampliando os contrastes de um padrão que entra na rede; e outro processo chamado de *top-down* que é responsável pela realimentação da rede e visa encontrar similaridades entre os padrões já classificados e os novos padrões. Como será visto na figura abaixo o processo *bottom-up* trata dos pesos de baixo para cima, enquanto que o processo *top-down* realiza a tarefa inversa, administrando os pesos de cima para baixo.

Conforme apresentado na fig. 3.8, a rede *Fuzzy ART* é composta por campos de atividade dos vetores, incluindo: um campo F_0 de nós que representam o vetor de entrada corrente, um campo F_1 que recebe ambas entradas *bottom-up* de F_0 e a entrada *top-down* do campo F_2, que representa o código de atividade ou categoria. O vetor ativo em F_0 é representado por $I=(I_1,...,I_M)$, e cada componente I_i possui um valor no intervalo [0,1] com o índice $i=1,...,M$. O vetor de atividade em F_1 é representado por $x=(x_1,...,x_M)$ e o vetor de atividade em F_2 é representado por $y=(y_1,...,y_N)$. O número de nós de cada campo é arbitrário.

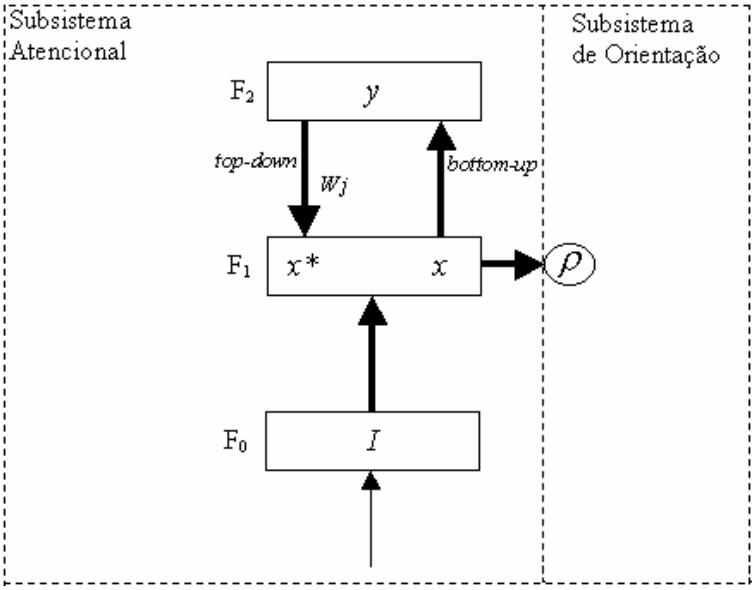


FIGURA 3.8 - Arquitetura de *Fuzzy ART*

Existem dois subsistemas que atuam em um módulo ART para processar os padrões de entrada:

- Subsistema Atencional: realiza o processamento dos padrões de entrada familiares, determinando respostas e representações internas mais precisas de tais padrões;

- Subsistema de Orientação: atua sobre os padrões estranhos, inibindo o subsistema atencional quando estes padrões são apresentados à rede[GUA 93].

Na categoria F_2 existem nós $j(j=1,...,N)$ que estão associados a um vetor de pesos adaptativos $W_j \equiv (w_{j1},...,w_{jM})$, que inicialmente estão configurados para o valor 1, não possuindo nenhuma categoria comprometida. Quando uma categoria é selecionada ela se torna comprometida e o peso é atualizado, não crescendo monotonicamente através do tempo, convergindo para um limite.

O funcionamento do modelo *Fuzzy ART* ocorre da seguinte maneira:

- quando um padrão de entrada estimula a rede ele se encontra na camada F_0 e recebe a denominação de *I*;

- este padrão *I* é então repassado para a camada F_1 recebendo o rótulo de *x*. *x* é idêntico ao padrão *I*, o que ocorreu foi somente a mudança de camada;

- através das conexões *bottom-up* o padrão *x* tem seus contrastes ampliados, quando propagado pelas conexões entre as camadas F_1 e F_2 (*Long-Term Memory - LTM*), produzindo na camada F_2 o padrão *y*, resultante do

processo *WTA* (*Winner Take All*), onde o neurônio com maior ativação recebe o valor 1 e aos demais neurônios é atribuído o valor zero;

- obtido o padrão *y* em F_2 as conexões *top-down* atuam no sentido de produzir a realimentação em F_1, gerando um padrão protótipo nesta camada, *x**, que representará o grau de similaridade entre o padrão corrente e o protótipo produzido. Desta forma é definido, por meio de uma regra que avalie *x** e *I*, se o padrão de entrada será estabilizado, ou seja, pertencerá ao neurônio vencedor, ou se outro neurônio, que representará as características deste novo padrão, deverá ser utilizado.

3.10.2 Codificação Complementar

A codificação complementar é utilizada para resolver um problema proposto por Moore que prevê a proliferação de categoria na camada F_2[CAR 91]. Sua dinâmica representa a porção ligada e desligada do padrão de entrada, ou seja, as características presentes e ausentes deste padrão, respectivamente; preservando a informação de amplitude. A definição da operação de codificação complementar é bastante simples: supondo que $\boldsymbol{a}$ representa um padrão de entrada, $\boldsymbol{a^c}$ representa seu complemento obtido na eq. 3.11.

$$a^c = 1 - a \tag{3.11}$$

Sendo a codificação complementar de uma entrada *I* na camada F_1 descrita por um vetor de tamanho 2*M*, com *M* representando a dimensão do vetor de entrada, a forma geral da codificação complementar é apresentada na eq. 3.12.

$$I = (a, a^c) \equiv (a_1, ..., a_M, a_1^c, ..., a_M^c) \tag{3.12}$$

3.10.3 Aprendizado em *Fuzzy ART*

A rede *Fuzzy ART* possui um processo de aprendizagem rápida que possibilita que padrões de entrada esporádicos e importantes sofram uma adaptação instantânea ao sistema. A aprendizagem rápida está associada a uma taxa de esquecimento (enfraquecimento da conexão representativa de alguma característica), a qual é conhecida como *fast commit slow recode*[CAR 92].

Existem três parâmetros que guiam o aprendizado da rede *Fuzzy ART*: o parâmetro de escolha $\alpha > 0$; o parâmetro de vigilância $\rho \in [0,1]$ e a taxa de aprendizado $\beta \in [0,1]$.

Quando o valor de β é igual a 1 significa que a aprendizagem rápida está habilitada. No entanto, se a propriedade *fast commit slow recode* for desejada para neurônios que não estejam comprometidos por nenhuma categoria, o valor de β deve ser reduzido ($\beta < 1$) depois da primeira adaptação daquele neurônio[GUA 94].

Os pesos das conexões entre as camadas F_1 e F_2 são inicializados com o valor 1 e representados pelo vetor w_j, conforme:

$$W_{j1} = W_{j2} = \ldots = W_{jM}\,, \tag{3.13}$$

onde M indica o número de neurônios em F_1 e N o número de neurônios em F_2 com o valor de j variando de 1 até N.

Uma categoria j é escolhida em F_2, por um determinado padrão de entrada, segundo a eq. 3.14 que representa a função de escolha T_j

$$T_j(I) = \frac{|I \wedge w_j|}{\alpha + |w_j|}, \tag{3.14}$$

onde o operador E-nebuloso($\wedge$) é definido por:

$$(x \wedge y)_i \equiv min(x_i, y_i) \tag{3.15}$$

e a norma | . | é definida por:

$$|x| = \sum_{i=1}^{M} x_i \quad . \tag{3.16}$$

O neurônio com maior ativação de entrada, pelo processo *WTA* (eq. 3.17) será o escolhido

$$T_J = max\{T_j : j = 1,\ldots,N\}\,, \tag{3.17}$$

onde J indica o índice do neurônio vencedor. Caso mais de um neurônio obtenha a ativação máxima, a regra determina que o nó escolhido deve ser aquele que apresentar o menor índice j.

Carpenter descreve que o sistema pode estar dentro de um *limite conservativo*, em que o valor de α, utilizado na escolha da função da categoria j em F_2 (eq. 3.14), deve receber um valor muito próximo a zero. Isto faz com que o padrão de entrada tente ativar uma determinada categoria, já existente, que represente nos seus respectivos pesos um subconjunto nebuloso do padrão de entrada. Evitando a modificação dos pesos da *LTM*, o que dá origem a expressão *limite conservativo*.

A ressonância que dá origem à aprendizagem, ocorre quando a proposição da eq. 3.18 for verdadeira, indicando que uma categoria estável foi encontrada.

$$\frac{|I \wedge w_j|}{|I|} > \rho \tag{3.18}$$

Assim, a rede pode realizar a adaptação do padrão de entrada, alterando o valor das conexões da categoria ativada por:

$$w_J(t+1) = \beta\left(I \wedge w_J(t)\right) + (1-\beta) w_J(t). \tag{3.19}$$

Entretanto, se a proposição da equação 3.18 for falsa o subsistema de orientação gera um sinal inibitório de *reset*, inibindo o neurônio *J* durante a apresentação do padrão de entrada corrente, para que este não seja novamente selecionado, fazendo com que outro neurônio em F_2 seja escolhido.

3.10.4 *Fuzzy ARTMAP*

A rede *Fuzzy ARTMAP* é composta por dois módulos *Fuzzy ART*: o módulo ART_a e o módulo ART_b; mais o módulo inter-ART, também conhecido por *map-field*, responsável por realizar um mapeamento associativo entre os módulos *Fuzzy ART*. Esta rede realiza um aprendizado supervisionado por associar os padrões de entrada com suas saídas correspondentes.

Desta forma, durante o treinamento, o módulo ART_a recebe uma seqüência de padrões de entrada $[a^{(p)}]$, enquanto que o módulo ART_b recebe uma seqüência de padrões $[b^{(p)}]$, onde $b^{(p)}$ prediz a categoria correta dado $a^{(p)}$.

Como pode ser visualizado na fig. 3.9, os modelos *Fuzzy ART* são formados por 3 camadas: a camada F_0, onde o padrão de entrada sofre a codificação complementar; a camada F_1, em que a entrada codificada é atribuída a x; e a camada F_2, onde são geradas as categorias de cada módulo. Em ART_a os neurônios de saída da camada F_1^a, representados por x^a, dão entrada na camada F_2^a, indexada por $j=1,...,N_a$, produzindo

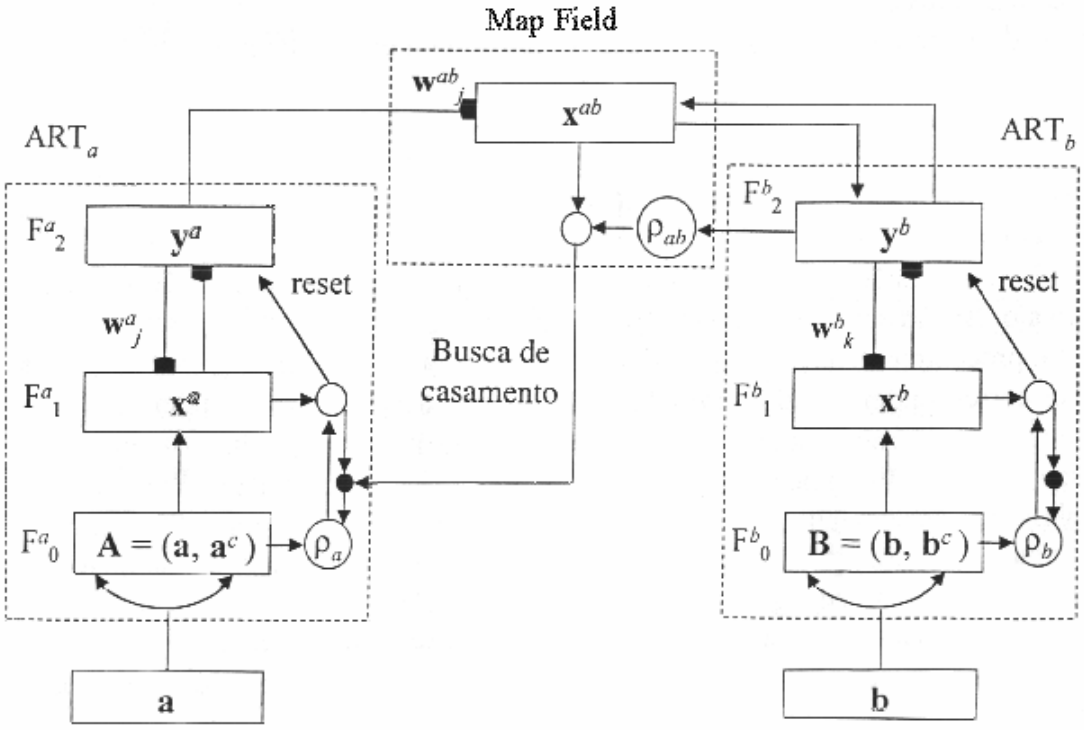


FIGURA 3.9 - Arquitetura da *Fuzzy ARTMAP*

sinais de saída binários y_j^a. Já em ART_b os neurônios de saída de F_1^b, x^b, entram na camada F_2^b, indexada por $k=1,...,N_b$, produzindo sinais de saída binários y_k^b. Os pesos entre as camada F_1 e F_2 são representados por w_j^a em ART_a e w_k^b em ART_b. J representa o índice do neurônio ativo em F_2^a, enquanto que K representa tal índice em F_2^b. F^{ab} representa o módulo inter-ART, associando os sinais de F_2^a e F_2^b, e é indexado segundo os nós da camada F_2^b por existir uma correspondência de um para um entre os neurônios de ambas as camadas. Os sinais de saída produzidos pelos neurônios de F^{ab} são representados por x^{ab} e os pesos da conexão entre o *j-ésimo* nó de F_2^a e F^{ab} são representados pelo vetor w^{ab}.

3.10.5 O Aprendizado de *Fuzzy ARTMAP*

A ligação dos módulos *Fuzzy ART* é realizada por uma rede associativa e um controlador interno que assegura a operação autônoma do sistema em tempo real[ENG 99]. Este controlador interno ao mesmo tempo que maximiza a generalização minimiza um erro preditivo, que é fruto da ativação de uma categoria em F_2^a por um determinado padrão ***a*** que prediz outra categoria em F_2^b que não a mesma ativada pelo padrão ***b*** corrente.

Aos valores de x^a, y^a, x^b, y^b, e x^{ab} é atribuído o valor 0 (zero) entre as apresentações das entradas e o peso w_J^{ab} é inicializado com o valor 1.

A ativação do módulo inter-ART depende da ativação de ART_a, ou de ART_b, podendo ocorrer 3 situações distintas:

- se o neurônio J de F_2^a é ativado, as conexões w_J^{ab} transmitirão seu sinal de ativação até F^{ab};
- se o neurônio K de F_2^b é ativado, o neurônio K em F^{ab} também é ativado, pois existe a correspondência de um para um entre F_2^b e F^{ab};
- se ambos ART_a e ART_b são ativados, F^{ab} é ativado somente se ART_a predizer a mesma categoria ativada em ART_b através das conexões.

Desta forma, o valor resultante de x^{ab} obedece a eq. 3.20

$$x^{ab} = \begin{cases} y^b \wedge w_J^{ab} & \text{se o } J\text{-ésimo nó de } F_2^a \text{ está ativo e } F_2^b \text{ está ativo} \\ w_j^{ab} & \text{se o } J\text{-ésimo nó de } F_2^a \text{ está ativo e } F_2^b \text{ está inativo} \\ y^b & \text{se o } J\text{-ésimo nó de } F_2^a \text{ está inativo e } F_2^b \text{ está ativo} \\ 0 & \text{se o } J\text{-ésimo nó de } F_2^a \text{ está inativo e } F_2^b \text{ está inativo} \end{cases} \tag{3.20}$$

Pela equação 3.20, constata-se que $x^{ab} = 0$ se a previsão w_J^{ab} for desconfirmada por y^b, na aplicação do *E*-nebuloso, causando um erro preditivo (eq. 3.21). Quando isso ocorre um processo chamado *Match Tracking* busca uma categoria melhor.

$$\left|x^{ab}\right| = \left|y^b \wedge w_J^{ab}\right| < \rho_{ab}\left|y^b\right| \tag{3.21}$$

O *Match Tracking* é responsável por gerar um sinal inibitório vindo de inter-ART, que é enviado para o subsistema de orientação de ART_a. Inicialmente, os parâmetros de vigilância do *map field*, ρ_{ab}, e de ART_b, ρ_b, podem ser fixados em qualquer valor no intervalo [0,1], enquanto que o valor de ρ_a é um valor básico (baixo), $\overline{\rho}_a$. Quando ocorrer o descasamento entre o rótulo previsto e o real, o valor de ρ_a é incrementado até apresentar um valor suficientemente alto para que o critério de vigilância em ART_a rejeite a categoria *J*, ou seja, iniba o neurônio vencedor de F_2^a. O valor de ρ_a deve ser incrementado até que satisfaça a eq. 3.22.

$$\rho_a > \frac{\left|I \wedge w_J^a\right|}{\left|I\right|}, \tag{3.22}$$

onde *I* é o padrão de entrada na forma complementada.

Sabendo que a regra de ativação de F_1^a é dada pela eq. 3.23

$$\left|x^a\right| = \left|I \wedge w_J^a\right| \quad , \tag{3.23}$$

após o incremento de ρ_a, ART_a busca um outro neurônio *J* que satisfaça os critérios de vigilância exigidos pela eqs. 3.24 e 3.21.

$$\left|x^a\right| = \left|I \wedge w_J^a\right| \geq \rho_a\left|I\right| \tag{3.24}$$

Quando a equação 3.21 é satisfeita, ou seja, a categoria *J* de ART_a está ativa e obedece ao critério de vigilância do módulo inter-ART, significa que ocorre a ressonância. Desta forma, w_J^{ab} se aproxima do vetor de saída do mapa x^{ab}, indicando que as conexões w_J^{ab}, entre as camadas F_2^a e F^{ab} devem ser atualizadas conforme a eq. 3.25.

$$w_J^{ab}(t+1) = \beta(x^{ab} \wedge w_J^{ab}(t)) + (1-\beta)w_J^{ab}(t) \tag{3.25}$$

Na aprendizagem rápida ($\beta = 1$), quando *J* em F_2^a aprende a predizer a categoria *K* em F_2^b, essa associação torna-se permanente, ou seja, w_J^{ab} vai ser sempre igualado a 1 para todo o instante de tempo.

4 Ferramentas e Métodos

Grande parte do procedimento para o desenvolvimento deste trabalho foi realizado no ambiente do software Matlab 5.3, da empresa MathWorks, salvo a aquisição das amostras para formar as bases de treinamento e teste, a qual foi realizada com a utilização do software Creative WaveStudio. Os procedimentos desenvolvidos são atualmente executados somente no ambiente do próprio software em que foram projetados, uma vez que o principal enfoque desta ferramenta é a simulação de eventos.

Os itens, logo abaixo citados, seguem a seqüência em que foram desenvolvidos e aplicados. Existe uma subdivisão no processo de reconhecimento que é descrita em duas etapas: a aquisição dos dados para treinamento, item 4.1, e a aquisição das amostras para teste. Numa primeira etapa os dados, em grande quantidade, foram coletados e armazenados para posteriormente serem processados, assim como numa etapa posterior ocorreu a coleta de amostras para teste, sendo possível também a aquisição de um único sinal adquirido pelo próprio ambiente de simulação do Matlab, o Simulink, com um breve armazenamento seguido do reconhecimento pela rede neural.

Todo o procedimento do sistema de reconhecimento de comandos pode ser visualizado na fig. 4.1, que apresenta a arquitetura do sistema proposto, e melhor compreendido na seqüência do texto.

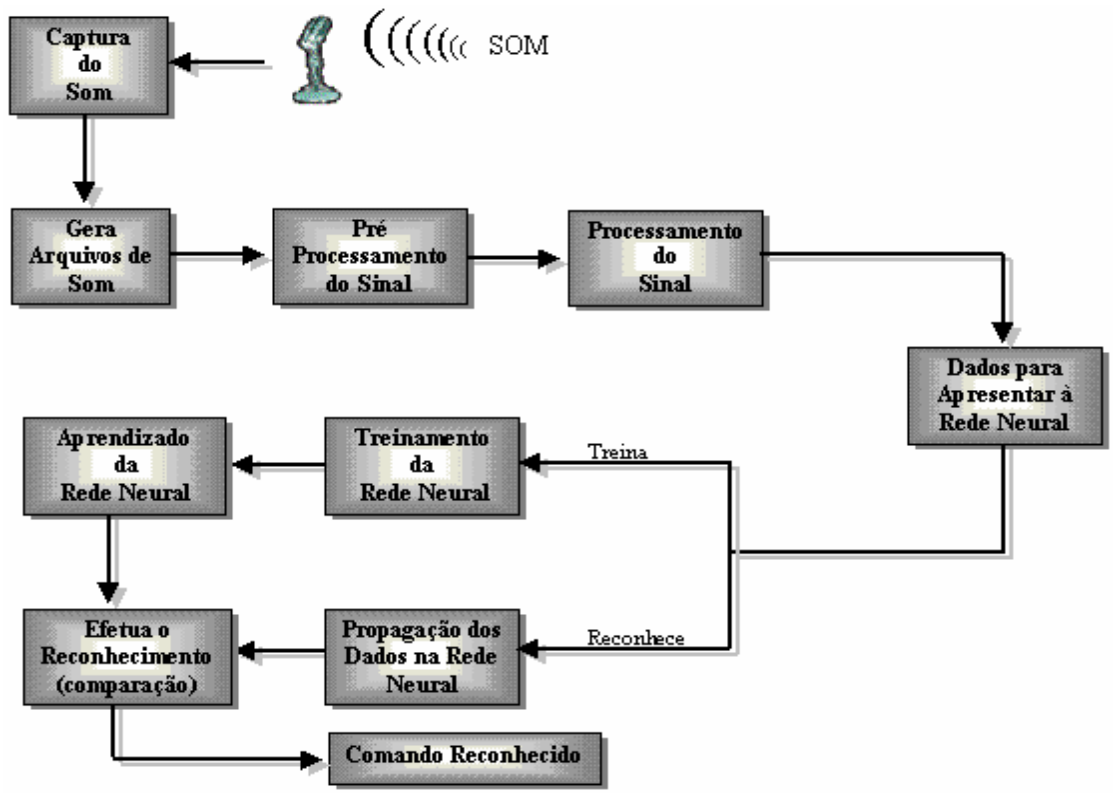


FIGURA 4.1 - Modelo do sistema proposto e implementado para o reconhecimento de comandos falados

4.1 Captura da Fala para Geração dos Padrões de Treinamento

O sinal de fala foi obtido através de um microfone ligado a um microcomputador PC, provido de uma placa de som Sound Blaster.

A aquisição do sinal foi realizada pela coleta da fala de 20 usuários de ambos os sexos, 10 masculinos e 10 femininos, que pronunciaram palavras relacionadas a comandos de direcionamento. Cada usuário que concedeu amostras sonoras repetiu cada um dos 5 comandos 10 vezes, compondo um total de 50 locuções por usuário, totalizando uma base sonora de 1000 locuções. Os comandos de direcionamento foram formados pelas palavras: **direita**, **esquerda**, **siga**, **pare** e **recue**. As sessões para gravação destes comandos foram realizadas em dias e horários diferentes, com a finalidade de abranger a variabilidade em que pronúncias repetidas do mesmo comando poderiam apresentar. O ambiente onde foram coletados os dados não era provido de qualquer isolamento acústico, estando exposto ao som de carros e pessoas passando na rua.

Para coleta e gravação dos arquivos de som ".WAV", foi utilizado o software Creative WaveStudio, que acompanha a placa de som utilizada. A manipulação e geração de arquivos sonoros por este programa é simples e prática. A seguir são citados alguns de seus comandos básicos de configuração e operação. A fig. 4.2 apresenta a tela do WaveStudio, onde são determinadas as configurações do arquivo a ser gerado, nela são possíveis as definições de: canais, mono para um canal e estéreo para dois canais; taxa de amostragem do sinal, 11025 Hz para voz, 22050 Hz para cassete, e 44100 Hz para CD; e determinação do tamanho da amostragem para qualidade do som, 8 e 16 bits (2^8 = 256 níveis de quantização e 2^{16} = 65536 níveis de quantização, respectivamente). Também pode ser configurada a habilitação ou não de componentes (meios) que serão utilizados, assim como a definição de seus respectivos volumes para gravação, fig. 4.3. O ambiente do programa que possibilita a gerência deste aplicativo é mostrado na fig. 4.4, juntamente com as definições dos principais botões que permitem a edição dos arquivos de som, onde:

1- permite a abertura de um novo arquivo para gravação;

2- permite a abertura de um arquivo sonoro já existente;

3- permite salvar alterações realizadas sobre determinada amostra, assim como os dados adquiridos para um novo arquivo;

4- permite a reprodução de um arquivo sonoro;

5- permite a gravação de uma amostra sonora e propicia o acesso a tela da fig. 4.5;

6- possibilita a determinação das configurações do novo arquivo;

7- possibilita a determinação dos componentes que serão utilizados na geração e edição dos arquivos sonoros e seus respectivos volumes.

Quando pressionado o botão 5 para gravação de um novo arquivo, uma tela é apresentada, fig. 4.5, em que podem ser vistas as configurações definidas para gravação, assim como a definição do nome do arquivo e sua localização.

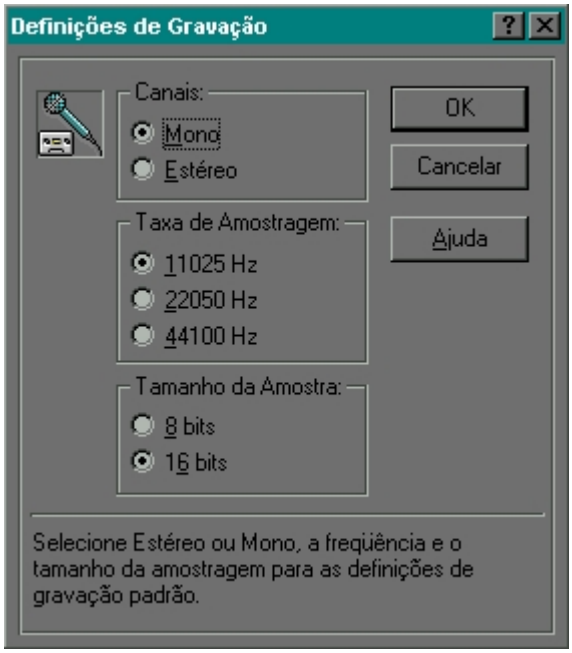


FIGURA 4.2 - Tela de configurações do Creative WaveStudio

Na conversão dos dados, analógico para digital, foi utilizada uma taxa de amostragem de 11025Hz e quantização de amostragem em 16 bits, em formato mono para todas as amostras geradas, o que garantiu uma boa qualidade para o sinal de voz adquirido. Como a maior parte da energia de um sinal de voz encontra-se abaixo dos 4 KHz, a taxa de amostragem escolhida atende ao teorema de Nyquist, onde um sinal limitado em banda pode ser representado por amostras adquiridas a uma taxa duas vezes maior que a freqüência máxima contida no sinal a ser amostrado[OPP 75]. Na aquisição destes dados não foi utilizada uma placa *anti-alias* para filtragem do sinal que entra no microfone.

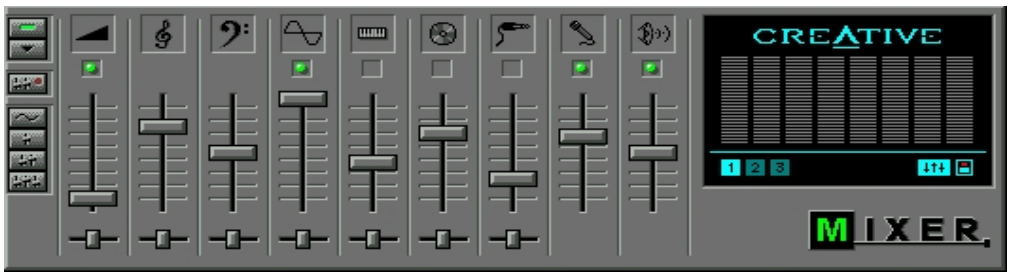


FIGURA 4.3 - Tela para determinação dos meios e volumes utilizados

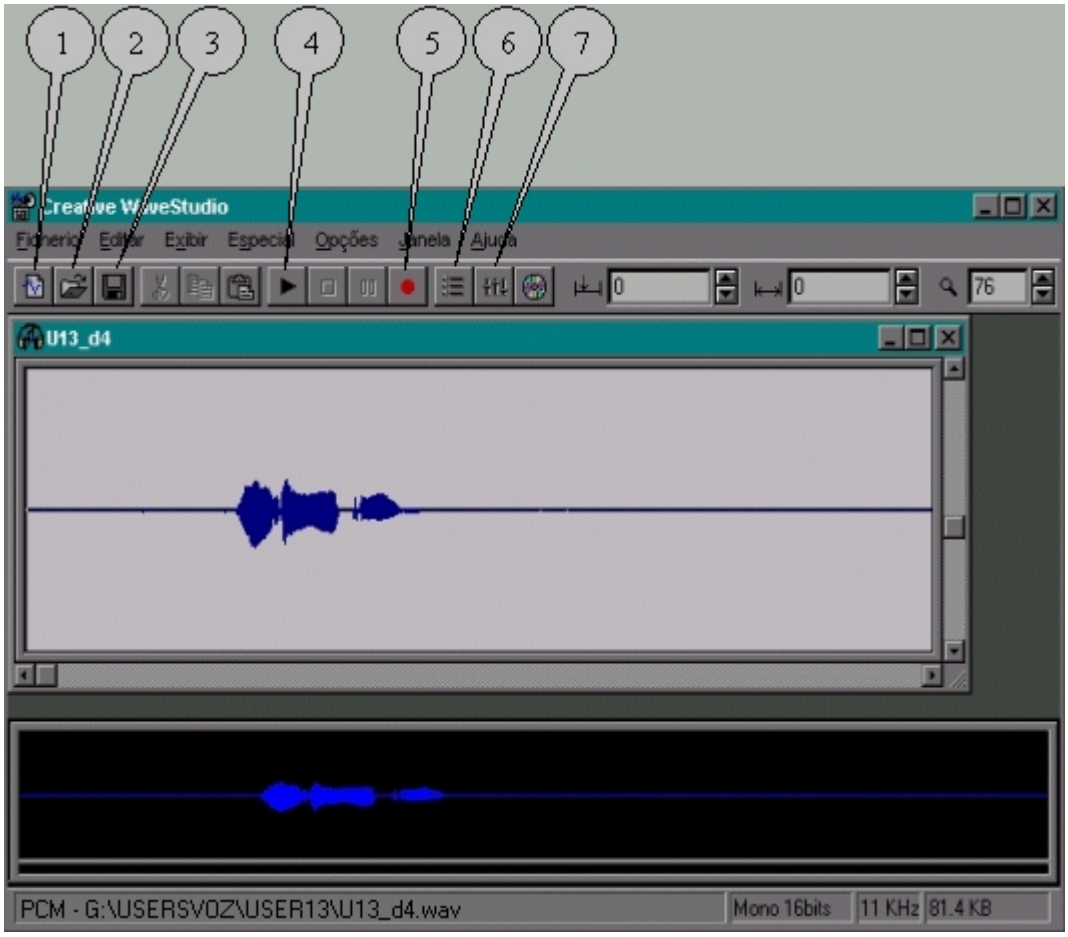


FIGURA 4.4 - Tela principal do Creative WaveStudio

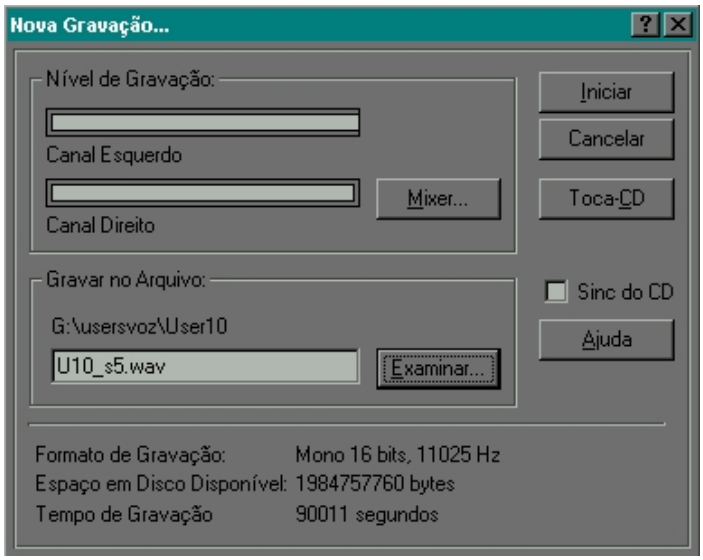


FIGURA 4.5 - Tela para determinação e gravação de novas amostras

Cada amostra capturada era precedida e também sucedida, por um período de silêncio ou ruído de fundo, o qual foi obtido juntamente com o sinal sonoro, a fim de validar os métodos de determinação dos pontos extremos da locução.

Os arquivos adquiridos seguiram a seguinte nomenclatura: Ux_Cy.wav, onde a letra 'U' significa usuário, 'x' o número do usuário, de 1 a 20, 'C' representa a letra inicial do respectivo comando, podendo ser: 'd', 'e', 's' ,'p' ou 'r'; e 'y' é a indicação do número da repetição do comando em questão. Por exemplo, o arquivo U10_s5.wav, indica a quinta repetição do comando siga pelo usuário 10. Uma vez que um dos objetivos do sistema é ser independente do locutor, não foi necessário uma distinção entre o sexo dos locutores na nomenclatura dos arquivos amostrados. O conjunto de amostras de determinado locutor foi gravado em um diretório específico com a indicação do usuário referente ao conjunto, sendo posteriormente obtidas deste, para leitura e pré-processamento.

4.2 Pré-processamento do Sinal Capturado

Uma vez obtidas todas as amostras gravadas em sua forma bruta, torna-se necessário dar início a fase de pré-processamento, onde cada amostra será submetida a um processo de normalização do sinal e posteriormente a detecção de seu início e fim, ou melhor, a eliminação dos dados ruidosos que foram registrados juntamente com a informação sonora. Como dados ruidosos pode-se considerar o ruído de fundo do próprio ambiente de obtenção das amostras.

4.2.1 A Normalização

O processo de normalização tem por objetivo padronizar diferenças entre amostras, fazendo com que as amostras possuam 1V (*volt*) de pico a pico, ou seja, limitar a amplitude entre +500mV e –500mV, no sentido de equalizar as amplitudes dos sinais de voz gravados com níveis de volumes diferentes, ocasionados pela distância entre o microfone e a boca ou pela diferença de entonação das pronúncias.

4.2.2 Determinação de Limiares

Os sinais na forma que são capturados, sob condições do próprio ambiente de gravação estão sujeitos aos ruídos de fundo, que certamente estarão presentes na geração da amostra. Este tipo de som não é de interesse do sistema uma vez que somente a informação referente a pronúncia é que deve ser processada. O principal motivo que leva a determinação dos pontos extremos de uma locução é a redução considerável do tamanho da amostra que foi produzida, reduzindo substancialmente o tempo de processamento requerido.

Tal determinação não é tão simples, salvo em ambiente que o nível de ruído é tratado como, por exemplo, uma sala de gravação a prova de som. Em condições reais, sistemas de processamento de fala não conseguem simplesmente pela medição da energia do sinal distinguir certos sons da fala, como fricativas fracas, em relação ao ruído de fundo.

O algoritmo utilizado para encontrar os pontos extremos de uma locução foi o proposto por Rabiner e Sambur[RAB 75], onde são estimadas duas medidas extraídas do sinal: a energia de curto prazo e a taxa de cruzamento por zero, aplicadas em janelas de 10 ms de duração do sinal. Este algoritmo apresenta como vantagens: um

processamento simples, boa precisão e adaptabilidade em relação ao ruído de fundo. O diagrama em blocos do algoritmo para estimação dos *endpoints* pode ser observado na fig. 4.6, sendo considerado que nos 100 ms iniciais e finais de gravação não existe voz. A extração da média deste intervalo fornece as estimativas do ruído de fundo e de silêncio. Tais estimativas são compostas pela medida e desvio padrão da taxa de cruzamento por zero, e pela medida da energia. O cálculo da energia é realizado pelo somatório simples do módulo das amostras contidas nas janelas fixadas. O cálculo com a utilização do módulo ao invés do quadrado de cada uma das amostras evita a possibilidade de *overflow* que poderia ocorrer com o uso da exponenciação, assim como reduzir o tempo de processamento do algoritmo na ordem de *n*k* vezes (*n* é o número de amostras por janela e *k* é o número de janelas), correspondente ao tempo necessário para elevar cada amostra ao quadrado.

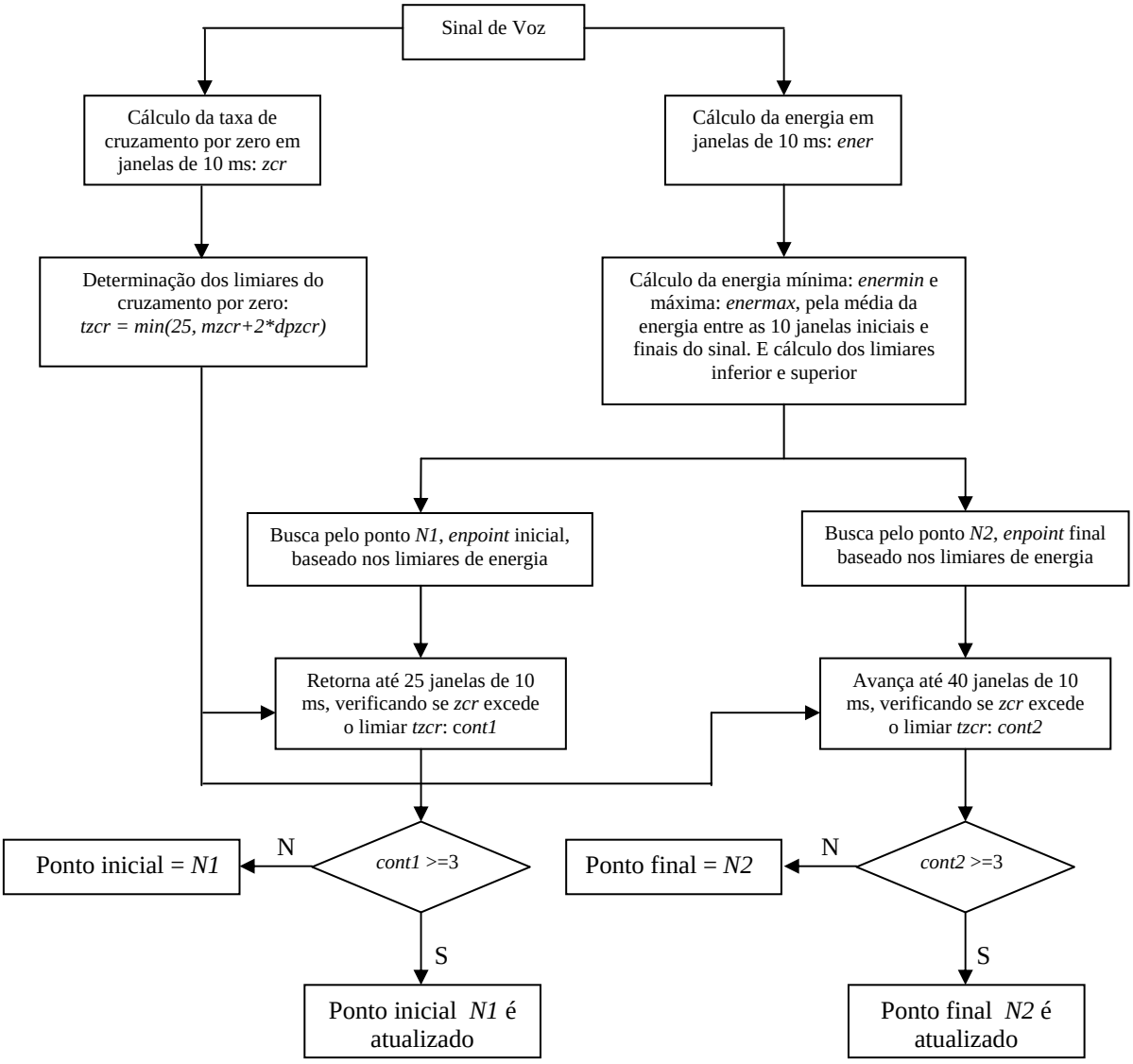


FIGURA 4.6 - Diagrama em bloco do algoritmo de determinação dos *endpoints*

A partir do cálculo da energia de tempo curto sobre cada janela, é determinado um valor máximo (*enermax*) e mínimo (*enermin*), os quais são utilizados para a estimação dos limiares, inferior (*liminf*) e superior (*limsup*) da amostra. A determinação dos limiares segue os seguintes passos:

$$I1 = enermin+0.03*(enermax-enermin) \quad (4.1)$$

$$I2 = 4*enermin \quad (4.2)$$

$$liminf = min(I1,I2) \quad (4.3)$$

$$limsup = 5*liminf \quad (4.4)$$

Com a estimação dos dois limiares é possível percorrer a amostra em busca da primeira janela com um valor de energia superior ao limiar inferior. Mantendo-se o nível de energia por mais duas janelas consecutivas maiores ou iguais ao limiar superior, sem sofrer nenhuma queda abaixo do limiar inferior, esta janela é considerada como o *endpoint* inicial da amostra. Caso contrário um novo ponto inicial que exceda o limiar inferior e depois exceda o limiar superior deve ser encontrado. A determinação do *endpoint* final segue um algoritmo similar a este, com a ressalva de que ele é processado no sentido inverso, ou seja, a partir da última amostra em direção ao início da mesma.

Os algoritmos para determinação dos pontos extremos, baseados na energia, são apresentados nas figuras 4.7 e 4.8, onde os pontos de início e fim da locução são expressos por *N1* e *N2*, respectivamente.

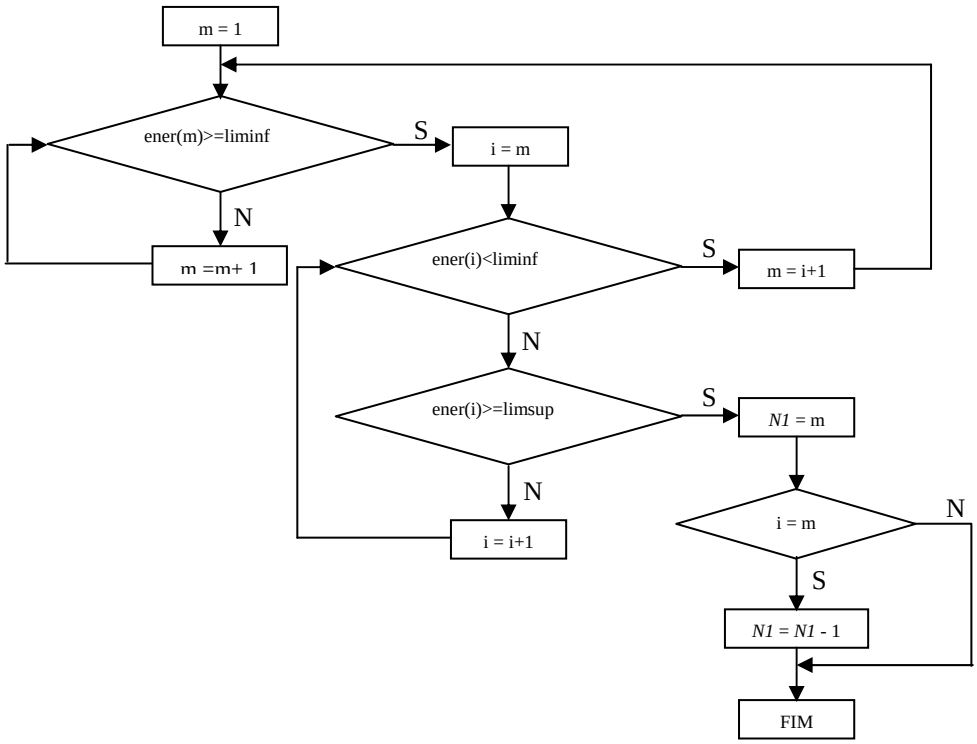


FIGURA 4.7 - Diagrama para estimação do *endpoint* inicial baseado na energia

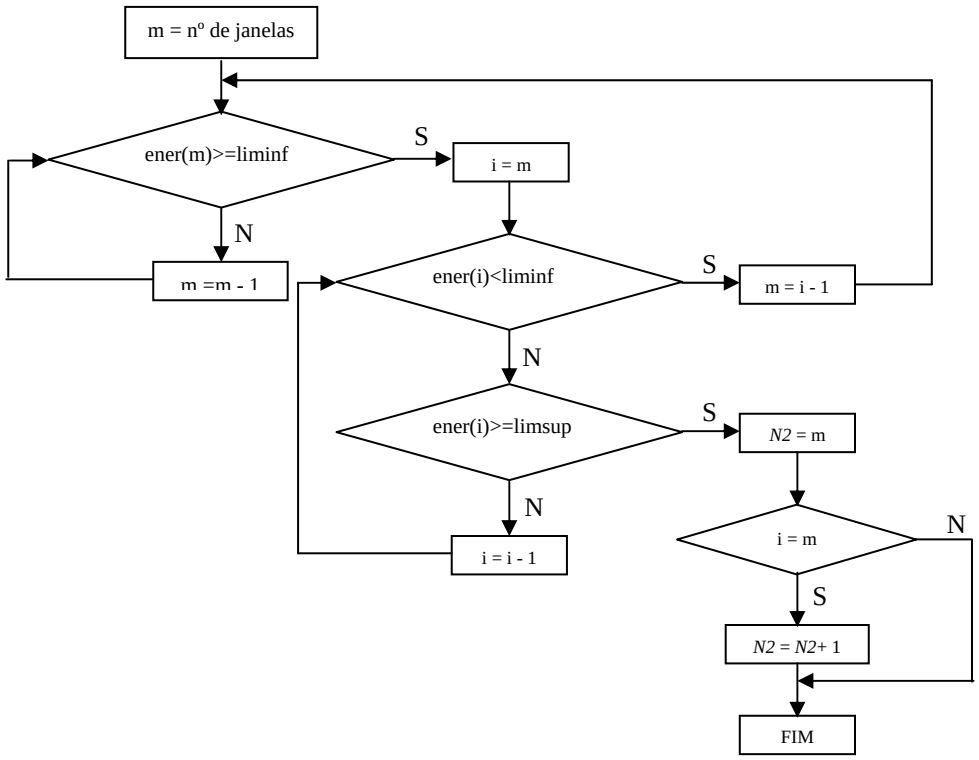


FIGURA 4.8 - Diagrama para estimação do *endpoint* final baseado na energia

Para a obtenção da taxa de cruzamento por zero, *tzcr*, deve ser escolhido o valor mínimo de um limiar fixado em 25 cruzamentos por zero em 10 ms, e o somatório da média da taxa de cruzamento por zero durante a zona considerada como de silêncio ou ruído de fundo, *mzcr*, mais duas vezes o desvio padrão da taxa de cruzamento por zero, *dpzcr*, neste período, como expresso na eq. 4.5:

$$tzcr = min(25, mzcr + 2*dpzcr) \qquad (4.5)$$

Obtida a taxa de cruzamento por zero, *tzcr*, são analisadas as 25 janelas anteriores a *N1*, ponto inicial determinado pela energia, assim como as 40 janelas posteriores a *N2*, ponto final determinado pela energia, verificando se o número de janelas com o número de cruzamentos por zero ultrapassar *tzcr* por mais de três vezes os *endpoints*, *N1* e *N2*, serão atualizados. Caso contrário serão mantidos.

A redução considerável no tamanho das amostras pode ser observado na fig. 4.9, onde o sinal sonoro adquirido possuía um tamanho original de 3,41 segundos; e após a aplicação do algoritmo de determinação dos limiares inicial e final, o mesmo foi reduzido a 0,838 segundos, demostrando-se de fundamental importância a sua aplicação.

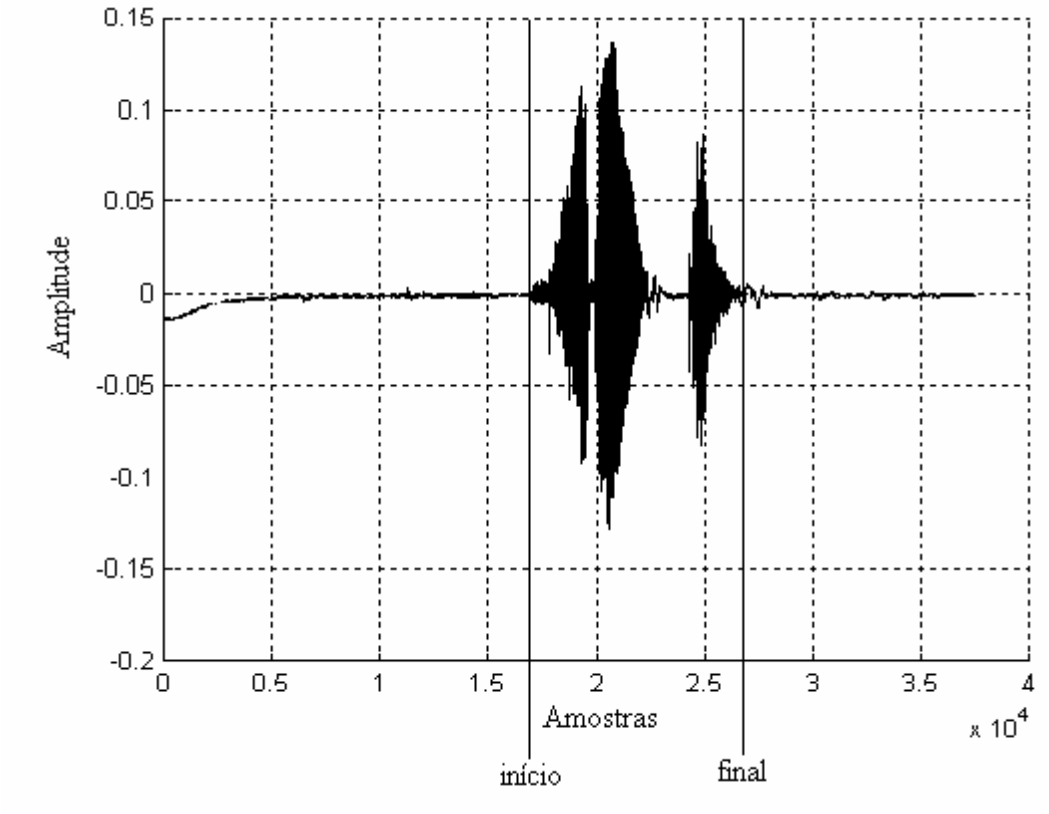


FIGURA 4.9 - Determinação dos *endpoints* de uma locução do comando direita

O algoritmo para detecção dos pontos extremos da locução seguiu a lógica utilizada por Rabiner e Sambur, seguida de uma pequena modificação do algoritmo original, conforme proposto por Diniz[DIN 97], onde a estimativa da taxa de cruzamento por zero e das medidas de energia é realizada sobre os 100 ms iniciais e finais da locução, e não como descrito no algoritmo original que se baseia somente nos 100 ms iniciais, obtendo-se assim bons índices na detecção dos limiares da pronúncia.

4.3 Processamento para Obtenção de Padrões

Após as amostras serem pré-processadas, esta etapa busca extrair do sinal as informações relevantes, que representem a amostra sonora original de tal forma que as características principais contidas nela não sejam afetadas ou modificadas, propiciando uma redução na quantidade de dados sem perda do conteúdo que caracteriza a informação primitiva. Na seqüência do texto são relatados os passos para extração dos padrões de referência.

4.3.1 Determinação das Janelas

Neste trabalho foi fixado que o número de janelas para qualquer locução, independente de sua duração é de 100 janelas, uma vez que será utilizada uma rede neural do tipo “para frente” (*feedforward*) a qual sempre requer um número fixo de entradas para um determinado conjunto de dados, tanto no seu treinamento como no reconhecimento. A determinação de 100 janelas está associada ao tamanho de cada

janela, que deve ser de até 40ms, onde se considera que o sinal seja invariante, podendo ser aplicada a extração dos coeficientes cepstrais de cada janela.

Como todas as locuções, inclusive as pronunciadas por um único locutor do mesmo comando, dificilmente apresentarão tamanhos idênticos, foi utilizado o método de janelas adaptativas[BEZ 94], em que é determinada uma sobreposição entre as janelas, enquanto que o tamanho de cada janela é que sofre uma variação. A sobreposição foi estipulada como sendo de 76%, obedecendo o critério de espaçamento estipulado pela janela de Hamming, como visto no item 2.3.5. Desta forma, a sobreposição entre janelas consecutivas ocasiona um ajuste temporal das locuções, onde as janelas tenderão a corresponder aos mesmos fonemas pronunciados em locuções idênticas com tempo de duração diferente. As locuções maiores, relativas a pronúncias mais lentas, serão avaliadas por janelas maiores, fig. 4.10. E as locuções menores por sua vez serão particionadas por janelas proporcionalmente menores, fig. 4.11. Sendo que as 6 janelas apresentadas nas figuras citadas foram somente adaptadas para ilustração, uma vez que número de janelas na prática foi estipulado em 100 janelas.

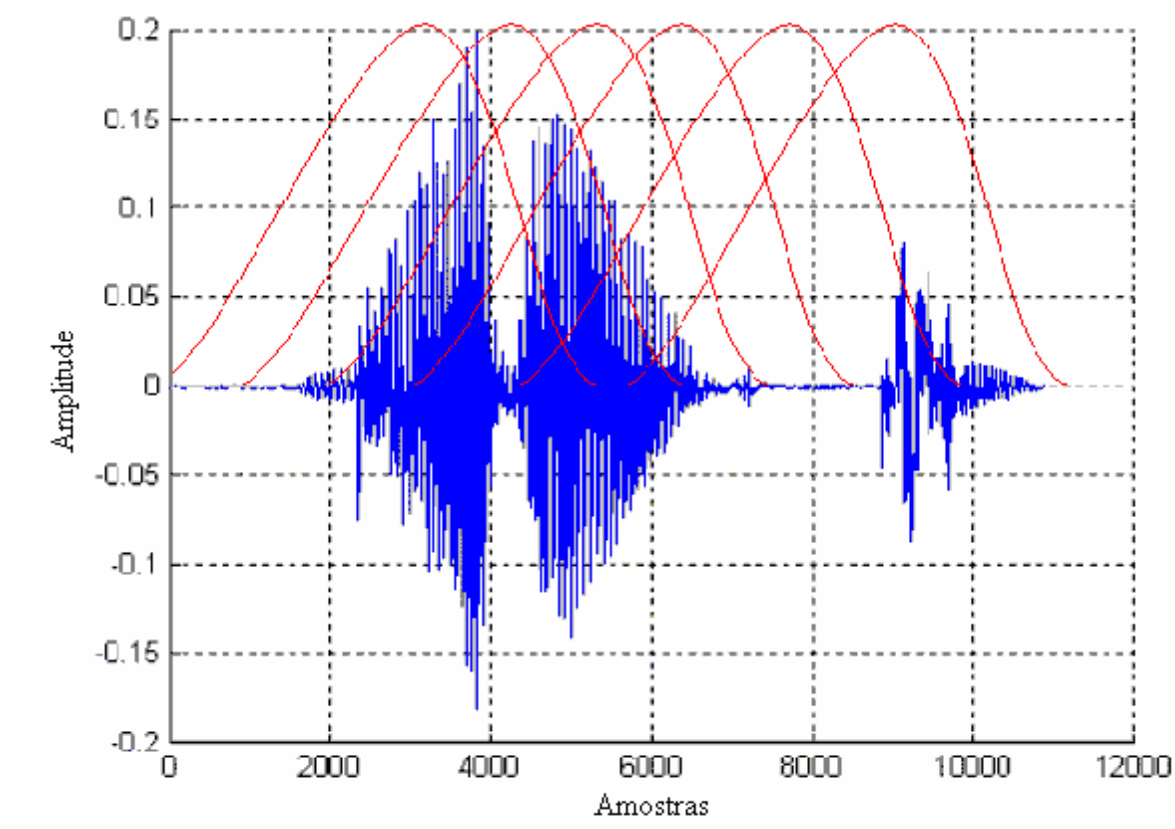


FIGURA 4.10 - Comando direita com 0,988 segundos de duração, aplicadas 6 janelas adaptativas com 76% de superposição e 4950 amostras por janela

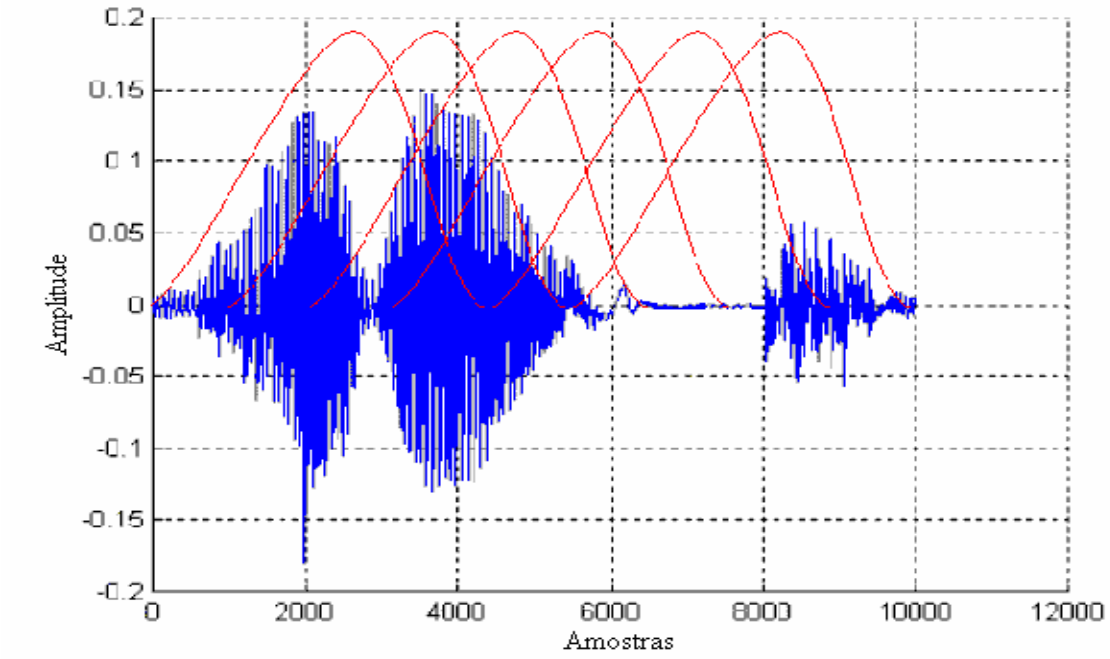


FIGURA 4.11 - Comando direita com 0,908 segundos de duração, aplicadas 6 janelas adaptativas com 76% de superposição e 4550 amostras por janela

A determinação do comprimento de cada janela em sua respectiva locução é obtida por meio da eq. 4.6.

$$Na = N/((n\_jan-1)*(1-sobrep/100)+1), \quad (4.6)$$

onde:

Na : é o número de amostras por janela;
N : é tamanho total de determinada locução;
n_jan : é o número de janelas desejado, no caso 100; e
sobrep : é o valor da sobreposição expressa em porcentagem.

De posse de *Na*, já é possível determinar a janela de Hamming que será aplicada sobre cada uma das janelas para a próxima fase do processamento que é a obtenção dos coeficientes cepstrais.

4.3.2 Extração de Coeficientes Cepstrais

Foi determinado como 14 o número de coeficientes cepstrais que seriam obtidos de cada janela. Considerando que a informação de interesse situa-se na faixa dos 10 aos 20 primeiros coeficientes cepstrais de cada janela, desconsiderando-se sempre o primeiro coeficiente, o qual refere-se a energia do período analisado.

Sobre cada uma das 100 janelas sobrepostas foi aplicada a janela de Hamming que foi responsável por produzir uma atenuação nas extremidades das janelas, não

ocorrendo uma distorção tão abrupta, da onda temporal, nas extremidades como ocorre com o uso da janela retangular.

Desta forma os coeficientes cepstrais foram adquiridos pela aplicação da Transformada Inversa de Fourier do logaritmo da amplitude da Transformada de Fourier sobre o sinal de voz particionado em janelas. Com a extração do segundo elemento até o décimo quinto de cada janela foi composto um vetor com 1400 elementos, o qual contém as principais informações que caracterizam o sinal original de determinada locução.

Foram extraídos os cepstros de todas as 1000 locuções adquiridas os quais produziram 1000 arquivos '.TXT', cada um contendo os dados cepstrais relativos a cada locução. Estes arquivos foram salvos com os mesmos nomes dos originais '.WAV' sendo somente diferenciada a extensão de seus respectivos nomes.

4.4 Apresentação dos Padrões para as Redes Neurais

A escolha da rede neural *Backpropagation* aplicada neste trabalho, é devido a extensão de seu uso em diversos problemas que envolvem o reconhecimento de padrões, enquanto que a escolha pela rede *Fuzzy ARTMAP* está relacionada a sua capacidade de aprendizado rápido e geração autônoma de categorias.

Os dados utilizados para treinamento das redes neurais utilizadas neste trabalho são compostos pelos coeficientes cepstrais das 1000 locuções coletadas. Também foram capturadas amostras de usuários desconhecidos, ou seja, que não contribuíram com amostras para formação da base de treinamento; estas amostras seguiram o mesmo sistema de captura que as amostras coletadas para composição da base, conforme descrito no item 4.1 deste capítulo, com a utilização do software Creative WaveStudio, assim como a mesma forma de nomenclatura. Para criação da base de teste foram utilizados 10 usuários de ambos os sexos, desta vez, 5 masculinos e 5 femininos, que pronunciaram os respectivos comandos com uma série de 5 repetições por comando, totalizando 25 amostras por usuário, produzindo uma nova base, somente para teste de 250 amostras. É importante destacar que dos 10 usuários que forneceram amostras para teste, 40% deles estiveram presentes na produção da base para treinamento. Isto foi realizado no sentido de revalidar a sua participação com novas amostras a serem propagadas pela rede.

Primeiramente, as redes neurais devem ser treinadas para que os padrões semelhantes possam ser agrupados em classes afins. Após esta etapa de aprendizado, os dados sofrem o processo de validação do modelo neural treinado, onde os mesmos dados utilizados para treinamento das redes são apresentados novamente, e a rede terá que classificá-los.

Sobre os dados sonoros, provenientes de usuários desconhecidos, é aplicado o mesmo processo de extração dos coeficientes cepstrais e estes são apresentados para as duas redes realizarem a classificação. Abaixo são descritas as particularidades referentes a cada rede neural utilizada.

4.4.1 A Rede *Backpropagation*

Uma das redes definida para treinamento e propagação dos dados foi a rede *Backpropagation Feedforward* com três camadas. Sua escolha ocorreu em virtude da variedade de aplicações que utilizam este modelo e que têm obtido bons resultados na classificação de padrões. Foram utilizadas diversas configurações, mas a base da rede é composta por: 1400 neurônios na camada de entrada, referentes aos coeficientes cepstrais armazenados e 5 nós na camada de saída correspondente aos 5 comandos a reconhecer, a arquitetura básica é ilustrada na fig. 4.12.

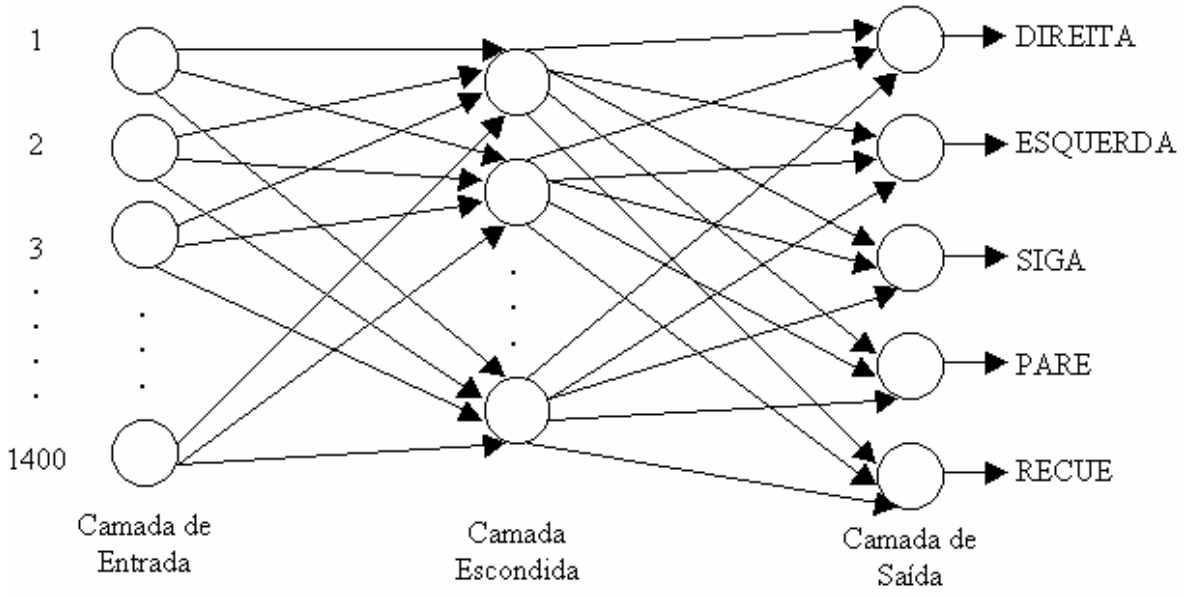


FIGURA 4.12 - Arquitetura da rede neural utilizada

Quanto a definição do número de neurônios da camada intermediária foram estipulados cinco diferentes quantidades de elementos de processamento: 10, 20, 30, 40 e 60 neurônios e observado o comportamento da rede. Não existindo nenhuma regra que fixe um valor exato.

Inicialmente, os arquivos que contêm os vetores resultantes da extração dos coeficientes cepstrais foram todos carregados em uma matriz tridimensional, onde cada comando representava uma dimensão e suas colunas representavam os 200 padrões referentes a cada comando. Já que a rede é supervisionada foi necessário a definição da saída desejada, representada por uma matriz bidimensional de 5x5, onde cada uma das colunas refere-se a um comando, sendo o neurônio que deseja-se ativar induzido ao valor 1 (um) e os demais ao valor 0 (zero).

Os parâmetros definidos para treinamento desta rede foram configurados de forma que a mesma obtivesse um melhor desempenho. A função de ativação utilizada tanto na camada intermediária como na camada de saída foi a logística sigmóide, pelo fato de os valores de entrada variarem entre 0 e 1. No treinamento foi utilizado o método do gradiente descendente com termo de momento, no sentido de acelerar o treinamento e minimizar o EMQ. Desta forma, o processo de treinar a rede ocorreu

diversas vezes, até que visualmente, pelo decaimento e estabilização do erro, como também pela observação dos valores resultantes do aprendizado, determinou-se a taxa de aprendizado fixada em $1x10^{-2}$, a constante de momento em 0,5 e o EMQ em $1x10^{-3}$.

Os pesos foram definidos pelas próprias funções do Matlab, com relação aos valores que deram entrada na rede, assim como a polarização, o valor de tendência dos neurônios. Os critérios de parada determinados foram: ou a rede atingir o seu objetivo, ou seja, o Erro Médio Quadrado, ou executar o número máximo estipulado de 3000 iterações, ou ainda, o decaimento do valor do gradiente a um índice menor que $1x10^{-6}$, garantindo então o aprendizado da rede.

Após todos os parâmetros definidos e a rede neural ter conseguido com sucesso atingir uma das metas definidas significa que o treinamento está concluído. A estrutura da rede treinada pode então ser salva em um arquivo em formato determinado pelo Matlab, para posteriormente ser carregado e utilizado para que os dados de teste possam ser propagados pela rede neural.

O próximo passo é a aplicação de amostras sonoras para utilização e validação do modelo, realizando testes com as amostras da sua própria base de dados, como também com novas amostras sonoras, dos referidos comandos, por usuários desconhecidos pelo sistema.

O resultado da propagação dos dados pela *Backpropagation* treinada, independente se os mesmos foram obtidos da base de dados ou diretamente da locução de um novo usuário, será responsável por ativar um dos cinco neurônios correspondentes aos comandos em questão.

Os valores que são produzidos pela ativação do neurônio vencedor tendem ao valor um, enquanto que os neurônios inibidos ficam próximos ou iguais ao valor zero. A decisão referente ao reconhecimento de um determinado comando apresentado à rede ocorre quando um dos valores dos neurônios da camada de saída ultrapassar um limiar estipulado em 50% de reconhecimento, ou seja, superior a 0,5, seguido pela comparação entre a saída obtida com a desejada para cada comando, classificando a saída da rede conforme ela obtiver uma aproximação com os valores de saída referentes a cada comando.

4.4.2 A Rede *Fuzzy ARTMAP*

Diferente da leitura de dados realizada para o aprendizado da rede *Backpropagation*, em que os dados estão organizados em uma matriz tridimensional, aqui os dados dos coeficientes cepstrais são carregados em uma matriz bidimensional, onde cada coluna desta matriz representa as características de uma determinada amostra sonora. Os dados estão organizados de maneira alternada, não seguindo uma forma ordenada, ou seja, uma seqüência de padrões que representem a mesma classe, e sim cada coluna indica um dos comandos a ser reconhecido proveniente de um usuário diferente. Esta forma de apresentação dos dados faz com que a rede tenha um aprendizado melhor, não sofrendo influência da ordem em que as amostras são apresentadas para a criação de categorias.

Cada amostra possui uma saída correspondente associada a ela, já que o aprendizado desta rede também é supervisionado. Tanto as amostras que entram em ART_a, como as que dão entrada em ART_b sofrem o processo de codificação complementar, que faz com que a entrada de ART_a, na camada F_1^a, possua 2800 neurônios, enquanto que a entrada de ART_b, na camada F_1^b , possua 10 nós.

Conforme as amostras são apresentadas, a rede automaticamente gera categorias de reconhecimento que determine conveniente, segundo a configuração inicial de seus parâmetros: ρ, β e α.

Inicialmente, todos os pesos recebem o valor 1, indicando que nenhuma categoria está comprometida. Cada amostra deve ser passada para a rede uma única vez e a rede deve ser capaz de avaliar a criação de novas categorias ou a identificação de classes semelhantes, agrupando os dados nas devidas classes. A rede é considerada treinada após todos os vetores de entrada terem sido apresentados à rede. No momento em que a rede agrupa os dados em categorias que julga similares, os pesos das conexões são atualizados, ocorrendo a ressonância da *Fuzzy ARTMAP*.

Ao fim do treinamento o vetor correspondente aos pesos é salvo para que possa ser utilizado na etapa de validação do modelo em que os dados de treinamento e teste são apresentados.

Quando uma amostra de teste é passada à rede sua propagação ocorre somente pelo módulo ART_a, pois não existe um rótulo correspondente que de entrada em ART_b para os dados de teste, uma vez que a rede já aprendeu a partir da etapa de treinamento. A classificação ocorre no momento em que o padrão de teste propagado pela rede corresponder a uma das categorias geradas pela rede ativando o nó correspondente a mesma.

4.4.3 Testes com o Simulink

Como descrito no início deste capítulo, foi desenvolvido um módulo no ambiente de simulação do Matlab, versão 5.3, o Simulink, com a seguinte estrutura de funcionamento (ver a figura 4.13): o usuário pronuncia, em um tempo estipulado de 2 segundos, qualquer um dos 5 (cinco) comandos para teste; sua locução é salva em um arquivo WAVE, denominado “samp.wav”, o qual é processado tão logo o botão que ativa o reconhecimento do respectivo comando seja acionado. Então este arquivo sofrerá todo o processo já referenciado anteriormente, e seus coeficientes cepstrais serão apresentados à rede neural que se encarregará de classificar esta pronúncia. Vale ressaltar que o usuário pronunciará somente um dos comandos a cada execução deste módulo.

É possível a adequação de qualquer uma das redes neurais treinadas neste módulo do Simulink, bastando para isto possuir programas separados que referenciem o reconhecimento por determinada rede. Nas propriedades do botão que efetua o reconhecimento é necessário referenciar o programa respectivo, ou seja, fazer a sua chamada, habilitando este módulo para realizar o reconhecimento de amostras de maneira fácil e prática.

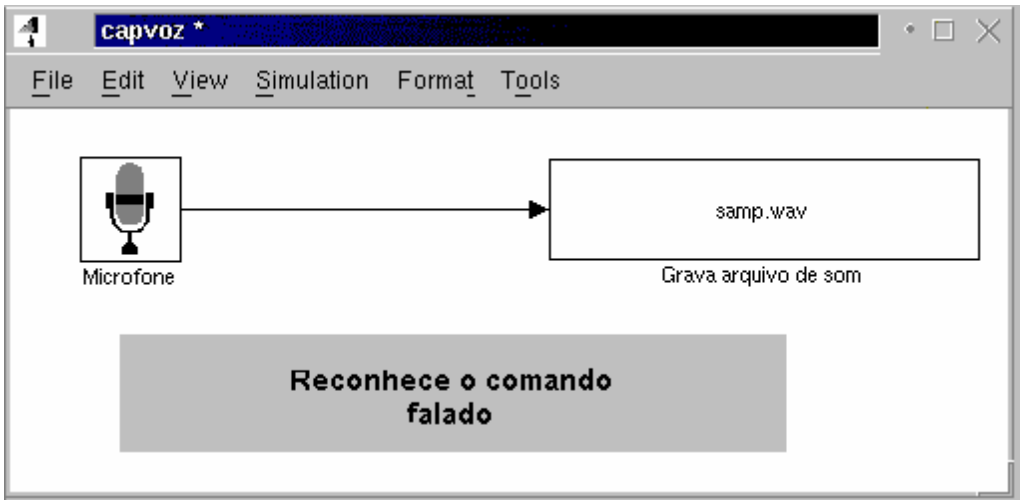


FIGURA 4.13 - Módulo para captura e reconhecimento de comandos

5 Resultados Obtidos

Para validar o sistema proposto foram realizados vários testes. Um deles realizado sobre a base de comandos de direcionamento, descrita no capítulo anterior, e mais dois testes avaliando o desempenho do sistema proposto: um utilizando parcialmente a base de dados do Projeto REVOX, desenvolvido no Instituto de Informática da UFRGS; e outro fazendo uso da base de dados do Laboratório de Processamento de Sinais e Imagens (LaPSI), do Departamento de Engenharia Elétrica da UFRGS, utilizada no projeto: Implementação de um Sistema de Controle Vocal de Equipamentos de Automação Industrial. As bases descritas para avaliação sofreram o processo de extração dos coeficientes cepstrais que foram submetidos às redes *Backpropagation* e *Fuzzy ARTMAP* para treinamento e teste. É importante destacar que o treinamento e teste realizado com a rede *Backpropagation*, sobre as bases que não foram produzidas por este trabalho, foi realizado somente com a configuração da rede que obteve os melhores resultados no treinamento e teste da base de comandos de direcionamento aplicada.

5.1 Resultados da Base de Comandos de Direcionamento

5.1.1 Comandos de Direcionamento Aplicados à *Backpropagation*

Inicialmente foram criadas 5 redes neurais, as quais foram treinadas e seus parâmetros ajustados em função de seu comportamento. Todas as redes seguem o modelo escolhido, o que diferencia uma da outra é a quantidade de neurônios que compõem a camada intermediária.

As redes estão especificadas na Tab. 5.1, onde são apresentados os valores obtidos em decorrência do treinamento respectivo a cada uma em separado.

TABELA 5.1 – Resultados alcançados no treinamento das redes neurais BP com diferentes parâmetros

Especificações / Redes	número de neurônios na camada intermediária	EMQ alcançado	Número de iterações
R1	30	0,00301818	2994
R2	20	0,00462638	2977
R3	60	0,00563109	2509
R4	40	0,00464447	3000
R5	10	0,00403261	3000

O critério pelo qual as redes R1, R2 e R3 concluíram seu ciclo de treinamento foi o valor do gradiente ter alcançado o limiar estipulado em $1x10^{-6}$. Enquanto que as redes R4 e R5 atingiram o número máximo fixado de iterações, mas sem alcançar o limiar do gradiente determinado como meta para o treinamento.

Primeiro foram aplicadas cada uma das 1000 locuções que compõem a base de treinamento para serem propagadas por cada uma das redes definidas. Uma a uma as amostras eram apresentadas e a rede se encarregava de classificá-las através das características cepstrais extraídas das amostras. O mesmo processo também ocorreu com as outras 250 amostras provenientes de usuários desconhecidos ao sistema, que foram igualmente apresentadas às redes.

Os resultados estão apresentados nos gráficos a seguir; cada um apresenta a taxa de reconhecimento obtido para determinado comando em termos de percentagem para as cinco redes treinadas. Os gráficos demonstram não somente os resultados, como também um comparativo entre o desempenho das redes.

Na fig. 5.1, que representa o comando **direita**, percebe-se que todas as redes apresentaram resultados superiores a 90%, indicando que os dados da base foram bem classificados para este comando e que os mesmos possuíam uma boa uniformidade. O mesmo índice não foi obtido para o comando **esquerda**, fig. 5.2, em que a melhor taxa percentual não ultrapassou os 72,5% de acerto.

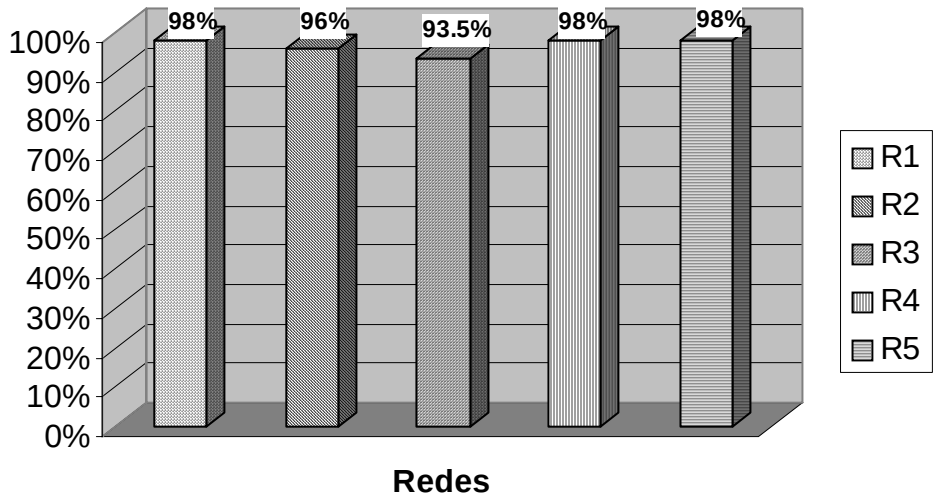


FIGURA - 5.1 Gráfico dos resultados obtidos para o comando **direita** relativos à base de treinamento sobre as redes BP

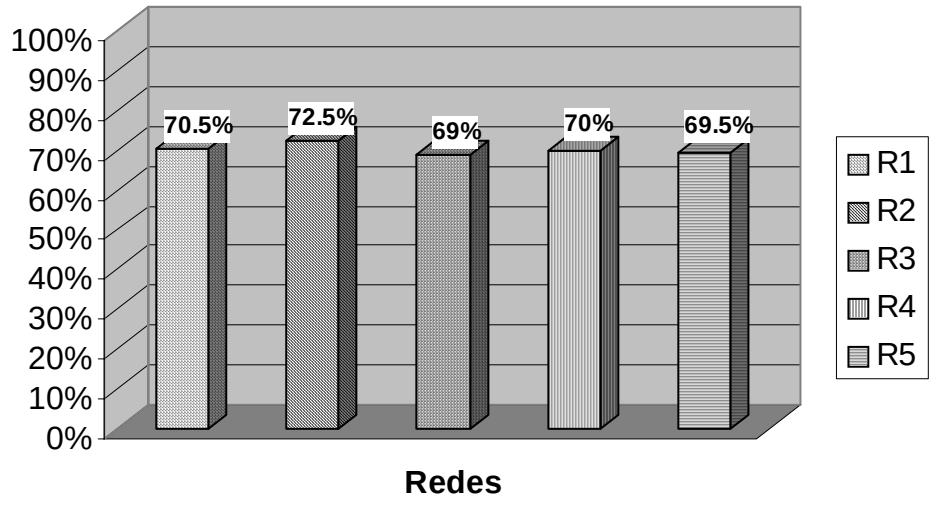


FIGURA - 5.2 Gráfico dos resultados obtidos para o comando **esquerda** relativos à base de treinamento sobre as redes BP

O comando **siga** da base de treinamento, fig. 5.3, obteve uma diferença de 5,5 pontos percentuais entre a rede que melhor o classificou e a rede que conseguiu o menor índice resultante.

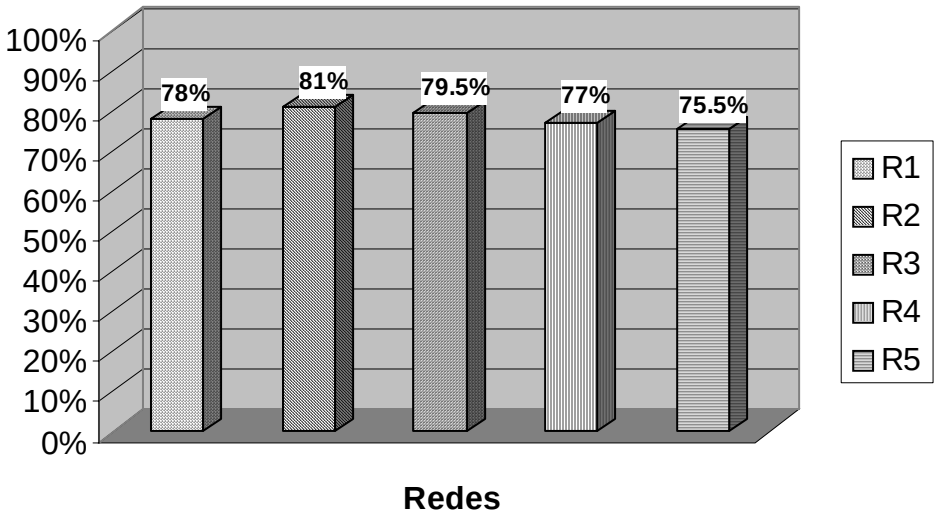


FIGURA - 5.3 Gráfico dos resultados obtidos para o comando **siga** relativos à base de treinamento sobre as redes BP

Já o comando **pare**, fig. 5.4, assim como o comando **direita**, fig. 5.1, apresentou bons resultados acima de 90% de reconhecimento em todas as redes. Bem como o comando **recue**, fig. 5.5, com taxas de acerto na faixa de 85%.

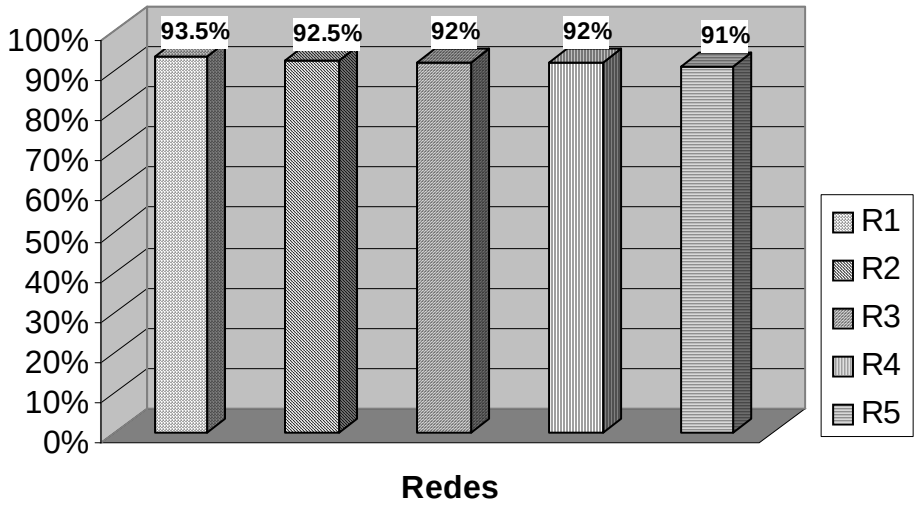


FIGURA - 5.4 Gráfico dos resultados obtidos para o comando **pare** relativos à base de treinamento sobre as redes BP

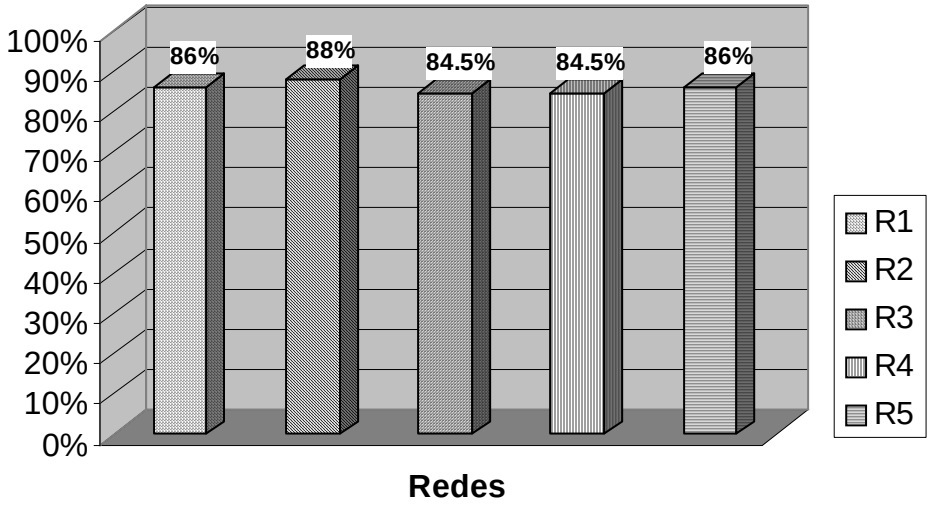


FIGURA - 5.5 Gráfico dos resultados obtidos para o comando **recue** relativos à base de treinamento sobre as redes BP

O desempenho geral obtido pelas redes sobre os dados utilizados também no treinamento pode ser visualizado na fig. 5.6, onde são apresentados os percentuais de acerto individuais de cada rede. A partir deste, constata-se que a rede R2 obteve um melhor desempenho frente as demais, entretanto não deve ser desconsiderado o índice obtido pelas redes R1, R3, R4 e R5, uma vez que a diferença entre a rede que obteve o melhor resultado e a que apresentou uma taxa mais baixa de reconhecimento ficou em 2,3 pontos percentuais.

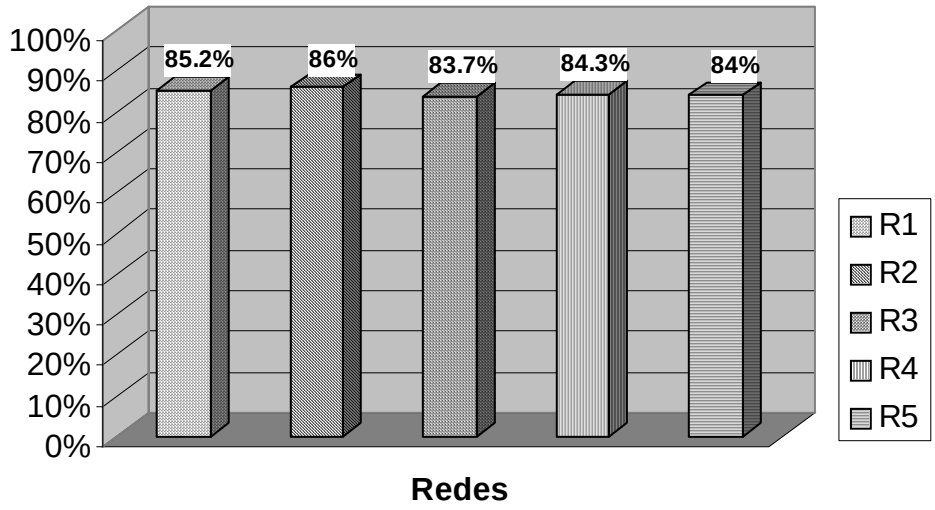


FIGURA - 5.6 Gráfico de desempenho geral das rede neurais BP a todos os comandos de voz que fizeram parte da base de treinamento

Isto demonstra que a independência dos dados, ou seja, a variação existente entre diferentes comandos pronunciados por diversos locutores, provenientes da base utilizada no treinamento da rede foi bem aceita diante dos resultados constatados, indiferente da quantidade de neurônios que compõem a camada intermediária de cada rede.

Como exemplo da saída obtida na classificação dos dados de treinamento propagados pela rede neural R2, são apresentados no Anexo 1 os valores produzidos pelos neurônios. Estes resultados são referentes somente a alguns comandos devido a grande quantidade de informações e visa propiciar uma visão dos valores de saída que foram gerados pelo processo de classificação da rede.

No reconhecimento das amostras de usuários desconhecidos ao sistema o comportamento foi um pouco diferente conforme visto na seqüência do texto. Como pode ser observado na fig. 5.7, o reconhecimento do comando **direita** manteve-se na faixa de 70% de acerto.

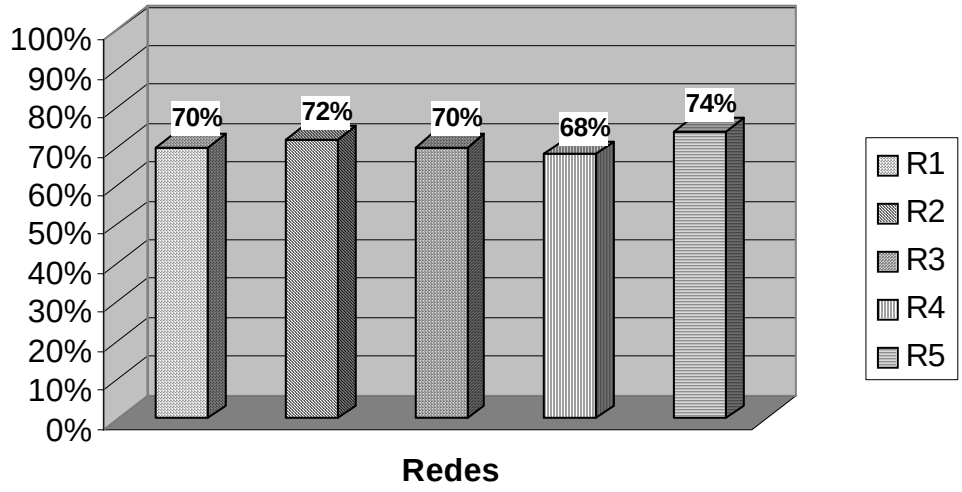


FIGURA - 5.7 Gráfico dos resultados obtidos para o comando **direita** aos novos dados desconhecidos apresentados as redes BP

O comando **esquerda**, fig. 5.8, por sua vez apresentou uma excelente taxa de reconhecimento de 92% na maioria das redes, embora a rede R1 tenha obtido um índice de 74%, demostrando que a quantidade de neurônios existentes na camada intermediária não é a adequada para o referido comando.

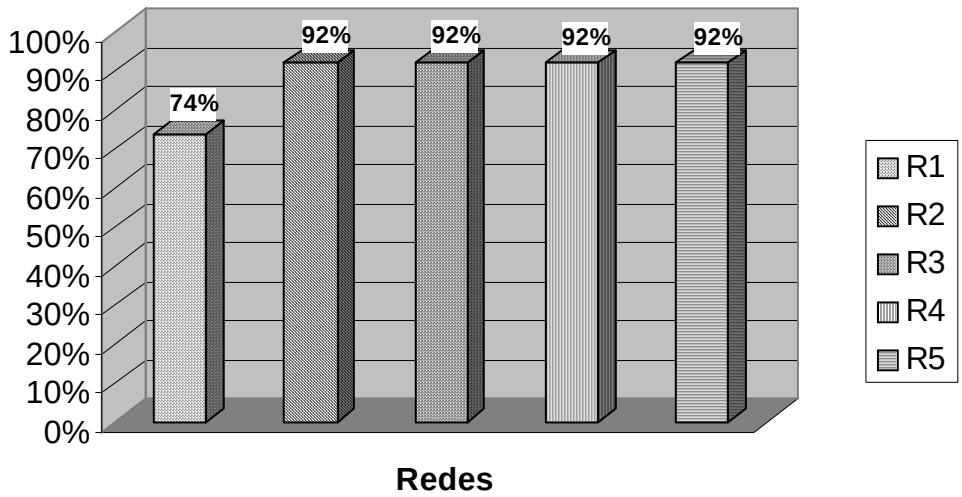


FIGURA - 5.8 Gráfico dos resultados obtidos para o comando **esquerda** aos novos dados desconhecidos apresentados as redes BP

Já os resultados para o comando **siga** apresentados na fig. 5.9, revelam uma grande sensibilidade deste comando em relação a independência de locutores, pois o índice máximo de reconhecimento não atingiu 75% para a rede R2, chegando ao nível de 42% para redes compostas por 40 e 60 neurônios na camada intermediária, respectivamente.

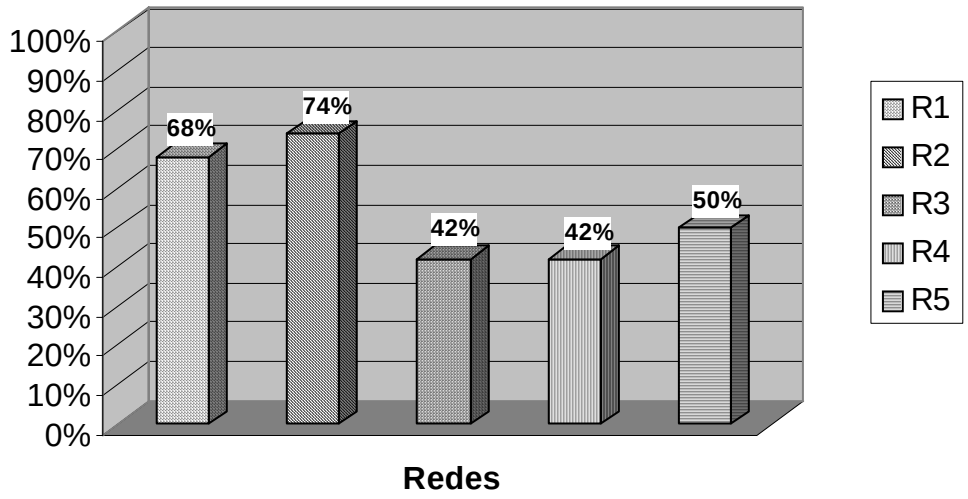


FIGURA - 5.9 Gráfico dos resultados obtidos para o comando **siga** aos novos dados desconhecidos apresentados as redes BP

Por outro lado para o comando **pare**, fig. 5.10, os resultados demonstraram-se excelentes quase atingindo um reconhecimento de 100%.

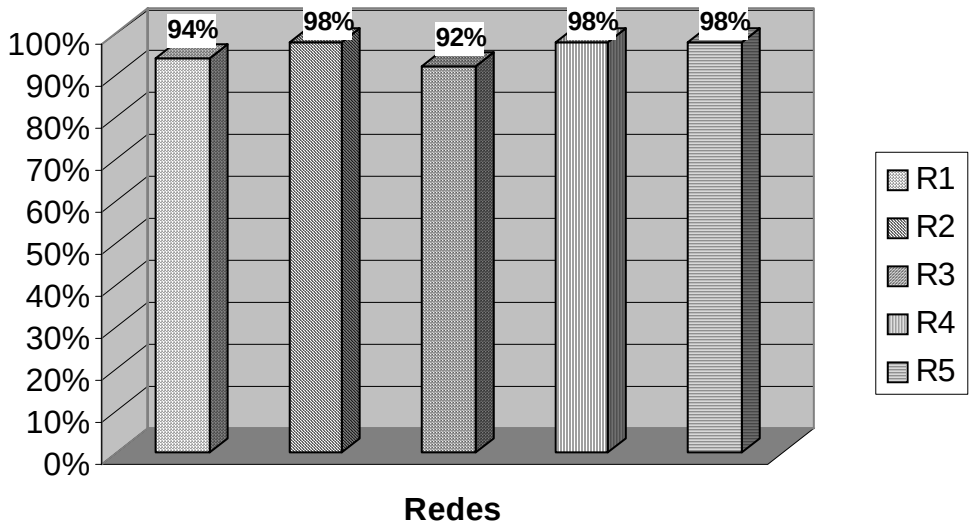


FIGURA - 5.10 Gráfico dos resultados obtidos para o comando **pare** aos novos dados desconhecidos apresentados as redes BP

O comando **recue**, fig. 5.11, assim como o comando **direita**, fig. 5.7, obteve resultados satisfatórios na ordem de 60% a 70% para todas as redes.

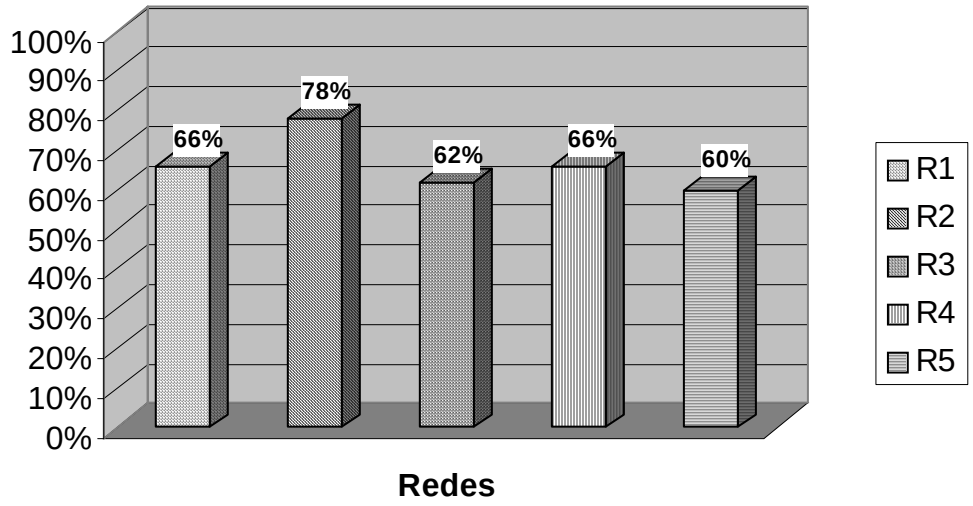


FIGURA - 5.11 Gráfico dos resultados obtidos para o comando **recue** aos novos dados desconhecidos apresentados as redes BP

Para os dados que não faziam parte da base original que foi utilizada para treinar a rede, o desempenho geral das redes sobre estes dados se demonstrou satisfatório conforme visto na fig. 5.12. Ocorreu uma queda já esperada entre os resultados de ambos os dados propagados, os utilizados no treinamento e os novos, em virtude das variações da pronúncia dos diferentes usuários.

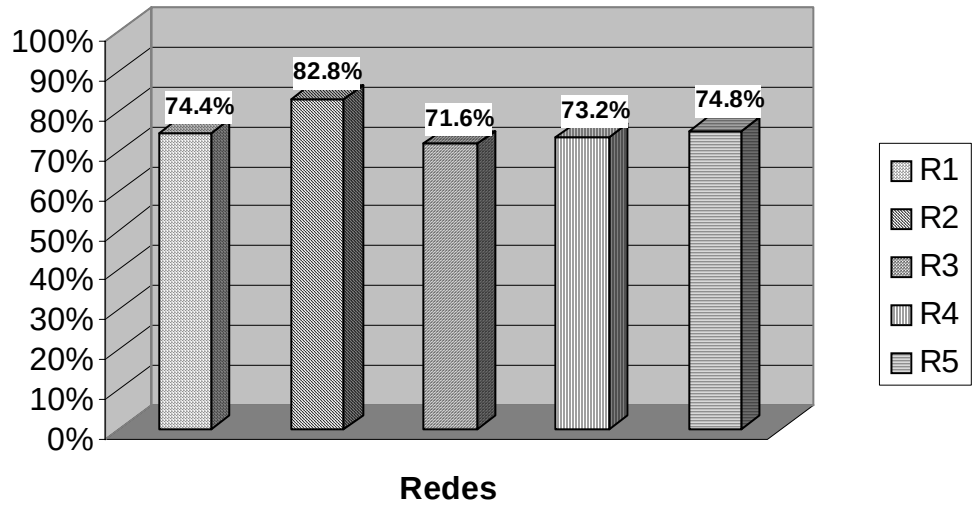


FIGURA - 5.12 Gráfico de desempenho geral das rede neurais BP a todos os comandos de voz provenientes do grupo de amostras desconhecidas

Embora tenha ocorrido uma variação maior entre os resultados relativos aos dados que não foram utilizados para o treinamento das redes em relação aos utilizados,

os mesmos se mantiveram uniformes. Com exceção do comando **siga**, que obteve o comportamento de seus resultados bastante variável, com diferença de até 32% entre as taxas de maior e pior classificação. Isto pode ser justificado pelo fato desta palavra ter seu início com um som fricativo, o que dificulta bastante a caracterização de suas amostras.

5.1.2 Comandos de Direcionamento Aplicados à *Fuzzy ARTMAP*

Ao contrário do que ocorreu no treinamento da *Backpropagation*, aqui não é determinada a quantidade de neurônios que compõem a camada intermediária da rede, muito menos é estipulado um número de épocas destinadas para a rede atingir um determinado objetivo. Na *Fuzzy ARTMAP* cada amostra é apresentada somente uma única vez à rede e esta se encarrega de produzir categorias que julgue conveniente.

A rede *Fuzzy ARTMAP* treinada com a mesma base aplicada para o treinamento da *Backpropagation* foi responsável por produzir 30 categorias para os 5 comandos a reconhecer. Destas 30 categorias produzidas: 7 representam o comando **direita**, 6 o comando **esquerda**, 7 o comando **siga**, 6 o comando **pare** e 4 o comando **recue**. A propagação dos dados que formam a base de treinamento obtiveram um resultado de 100% de reconhecimento.

Para os dados provenientes de usuários desconhecidos ao sistema, compostos por 250 amostras, a rede apresentou uma queda na taxa de reconhecimento, atingindo um desempenho geral de 65,2% de acerto. As taxas individuais de reconhecimento para cada comando são apresentadas na fig. 5.13, onde visivelmente são percebidas as baixas taxas apresentadas pelos comandos **direita** e **siga**, as quais, provavelmente, tenham a influência da composição fonética particular destes comandos.

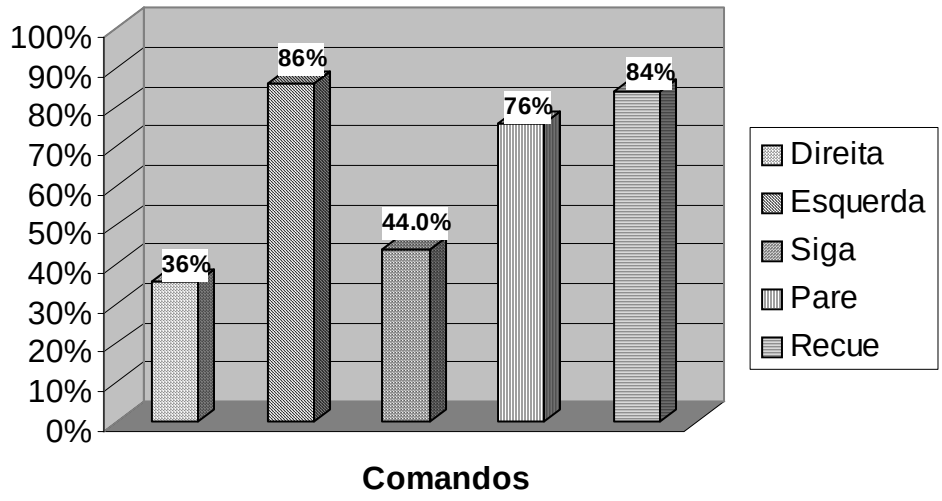


FIGURA - 5.13 Gráfico de resultados obtidos com a rede *Fuzzy ARTMAP* sobre comandos de voz provenientes das amostras de usuários desconhecidos da base de comandos de direcionamento

5.2 Resultados de Dados Parciais da Base do Projeto Revox

Esta base foi desenvolvida para um trabalho de identificação automática de pessoas pela voz, que combina o reconhecimento de comandos discretos com a verificação da identidade do locutor para efetivar o controle de elevadores.

A mesma é composta por comandos que representam os números ordinais de 1 a 9, ou seja, comandos de primeiro a décimo, mais a palavra "andar", provenientes de um conjunto de 100 usuários de ambos os sexos, os quais forneceram 6000 amostras relativas aos comandos e mais 6000 amostras referentes a palavra "andar".

Para o treinamento das redes propostas neste trabalho foram escolhidos os comandos: **primeiro**, **segundo**, **terceiro**, **quarto** e **quinto**, provenientes de um grupo de 20 usuários, 10 masculinos e 10 femininos, totalizando um conjunto de treinamento composto por 300 amostras, todas adquiridas por meio de um microfone conectado a um filtro passa baixas, *anti-alias*, cuja a saída é conectada à placa de som instalada no PC. A taxa de amostragem destas amostras é de 11025Hz, com resolução de 16 bits.

Também foram selecionadas amostras dos mesmos comandos, de usuários desconhecidos, para serem propagadas pelas redes. Os dados de usuários desconhecidos são representados por 10 usuários, 5 masculinos e 5 femininos, com 3 repetições de cada comando, formando um conjunto de 150 amostras de teste. Nenhum dos usuários selecionados para compor este conjunto de teste participou do conjunto utilizado para o treinamento das redes.

5.2.1 Comandos da Base do REVOX Aplicados à BP

Conforme especificado no início deste capítulo a rede BP utilizada para treinamento possui a mesma configuração que obteve os resultados mais expressivos na classificação dos dados da base de comandos de direcionamento, ou seja, a rede composta por 20 nós na camada intermediária.

No treinamento deste conjunto de dados foi obtido um erro de 0,001358 em 791 épocas, o treinamento encerrou quando o gradiente alcançou o limiar fixado de $1x10^{-6}$.

A propagação dos próprios dados utilizados para treinamento da rede obteve um desempenho geral de 98,9% de reconhecimento. O reconhecimento individual de cada comando é apresentado na fig. 5.14.

Quanto a propagação dos dados de usuários desconhecidos, ao treinamento efetuado nesta rede, também foi obtido um resultado muito satisfatório, uma vez que a taxa de reconhecimento atingiu 88% de acerto. Os percentuais de acerto relativos a cada comando são visualizados na fig. 5.15.

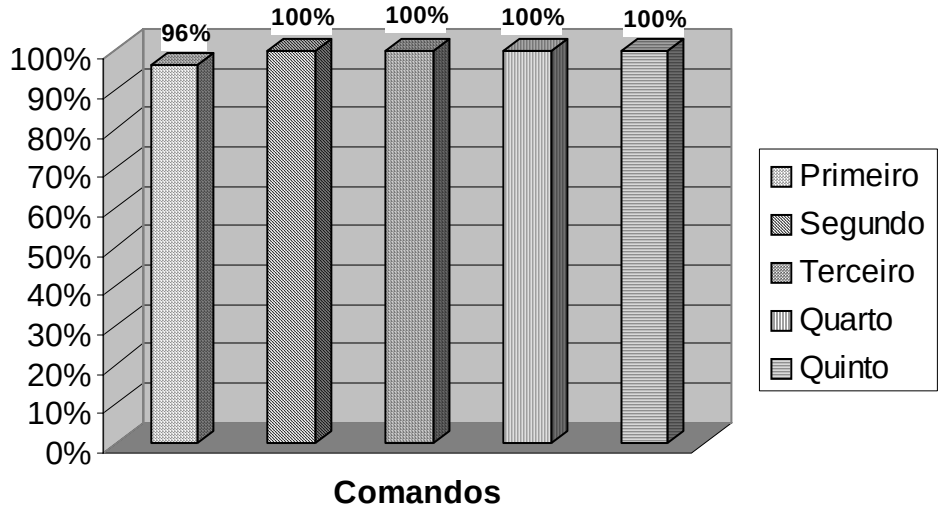


FIGURA - 5.14 Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes amostras de treinamento da base de comandos do REVOX

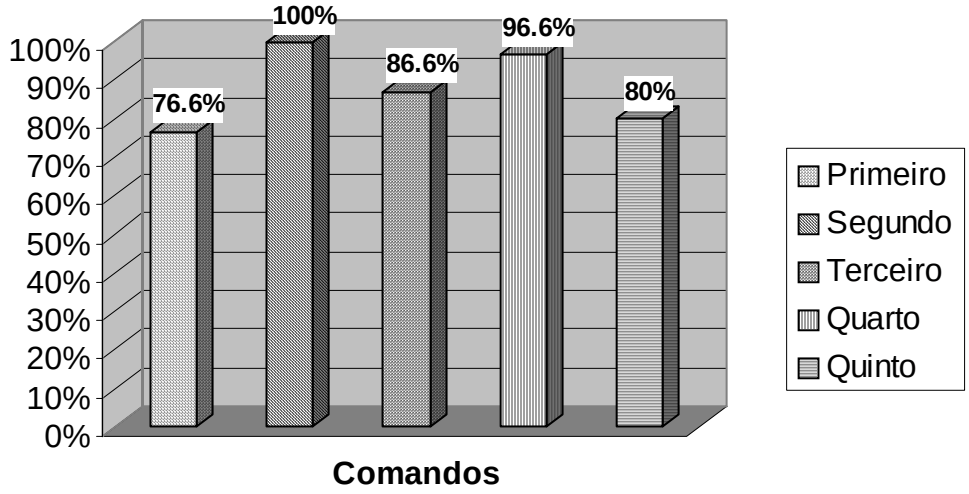


FIGURA - 5.15 Gráfico de resultados obtidos com a rede BP sobre comandos de voz de usuários desconhecidos provenientes da base de comandos do REVOX

5.2.2 Comandos da Base do REVOX Aplicados à *Fuzzy ARTMAP*

A rede *Fuzzy ARTMAP* treinada com o conjunto de 300 amostras produziu 10 categorias, as quais apresentaram: 3 categorias referenciando o comando **primeiro**, 1 o comando **segundo**, 3 o comando **terceiro**, 1 o comando **quarto** e 2 o comando **quinto**. Quando apresentados os próprios dados do treinamento para validação do modelo a rede obteve 100% de reconhecimento.

A propagação dos dados referentes aos usuários desconhecidos obteve um reconhecimento melhor que o obtido com a *Backpropagation* na mesma circunstância. O desempenho geral obtido foi de 90,7%, os percentuais de acerto individuais de cada comando são apresentados na fig. 5.16. A queda na taxa de acerto somente para o comando **primeiro** pode ter sido causada pela incapacidade das técnicas de processamento de sinais utilizadas para extrair algumas características próprias do sinal que compõe este comando, agravada pela apresentação de amostras de locutores estranhos aos usados no treinamento. Esta suposição é baseada também pelo fato de que a classificação dos mesmos dados pela rede BP (Fig. 5.15) apresentar o índice mais baixo de reconhecimento para este comando.

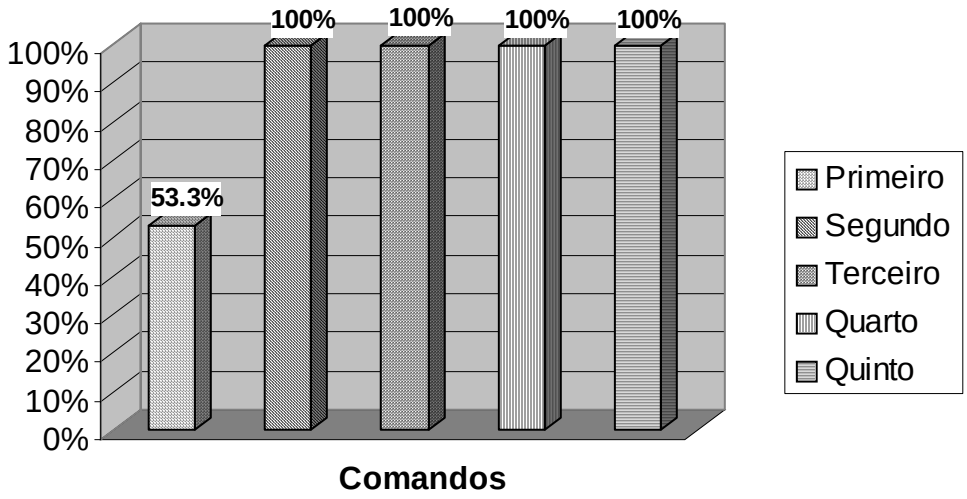


FIGURA - 5.16 Gráfico de resultados obtidos com a rede *Fuzzy ARTMAP* sobre comandos de voz de usuários desconhecidos provenientes da base de comandos do REVOX

5.3 Resultados dos Dados da Base do LaPSI

A base de dados do LaPSI, desenvolvida pelo Departamento de Engenharia Elétrica, foi gerada visando o controle vocal de equipamentos de automação industrial. Os comandos que compõem esta base estão, em sua maioria, relacionados a comandos para manipulação de um elevador. São 29 comandos, provenientes de um conjunto de 30 usuários, os quais forneceram 3 repetições de cada comando.

Foi testado um conjunto de 10 comandos que compõem a base; esta mesma avaliação também foi realizada pelo projeto que deu origem a esta base e a comparação entre resultados é realizada ao final deste capítulo. Para o treinamento das redes propostas neste trabalho foram utilizados os comandos: **primeiro**, **segundo**, **terceiro**, **quarto, elevador, ok, ventilador, porta, beijo** e **alarme**, provenientes de um grupo de 30 usuários, 24 masculinos e 6 femininos, totalizando um conjunto de treinamento composto por 900 amostras, todas as amostras foram adquiridas por meio de um microfone conectado a um filtro passa baixas, *anti-alias*, cuja a saída é conectada à placa de som instalada no PC[LAB 95]. A taxa de amostragem destas amostras é de

44100Hz, com resolução de 16 bits. Estas amostras foram coletadas utilizando uma cabine de isolamento acústico.

Como não havia amostras de usuários desconhecidos foram realizados testes utilizando metade dos dados do conjunto selecionado para treinamento e reservada a outra metade para a servir como amostras de usuários desconhecidos. Também foi realizado um treinamento utilizando 70% dos dados do conjunto selecionado para aprendizado das redes e reservado os outros 30% para a servirem como amostras de usuários desconhecidos. É importante especificar que ambas as divisões realizadas sobre o conjunto de amostras selecionadas tratou de dividir os dados proporcionalmente em relação ao sexo dos usuários, por exemplo, no treinamento que foi realizado com a metade dos dados foram apresentados para as redes 12 usuários masculinos e 3 femininos. E para o treinamento com 70% do conjunto haviam 17 usuários masculinos e 4 femininos.

5.3.1 Comandos da Base do LaPSI Aplicados à BP

A aplicação de todo o conjunto de dados para treinamento da rede *Backpropagation*, também composta por 20 neurônios na camada intermediária, como a rede aplicada aos dados do REVOX, produziu um erro de 0,003054, atingindo o número máximo estipulado de 3000 épocas, sendo o treinamento encerrado pelo gradiente que alcançou o limiar de $1x10^{-6}$.

A propagação dos dados de treinamento pela rede BP obteve uma taxa de acerto de 96,6%, as taxa de acerto individuais para cada comando podem ser visualizadas na fig. 5.17.

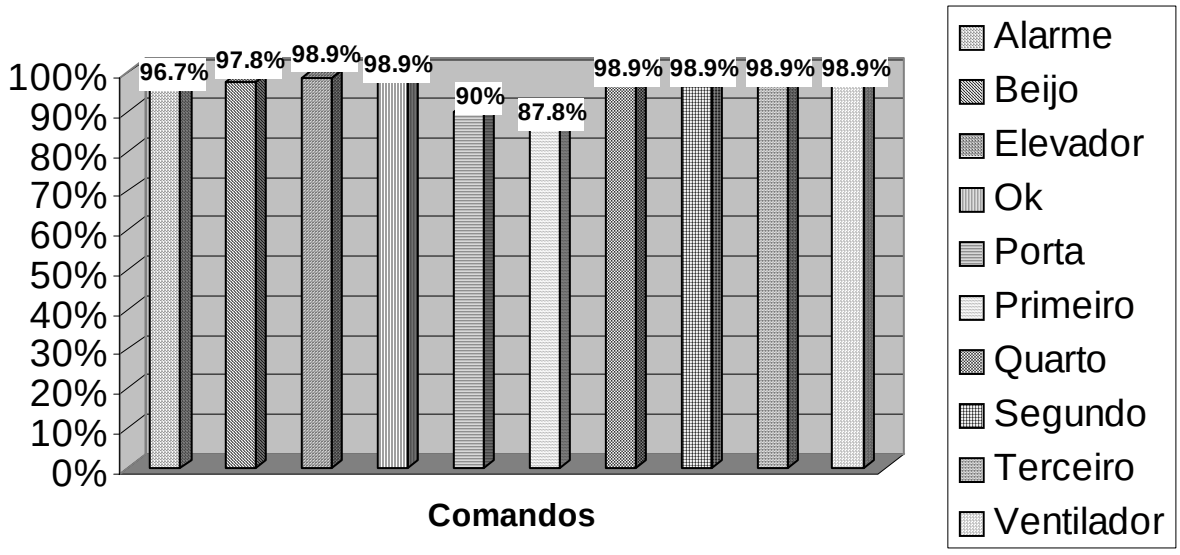


FIGURA - 5.17 Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes do conjunto de amostras utilizadas no treinamento do conjunto da base de comandos do LaPSI

A fig 5.18 apresenta os resultados que os comandos obtiveram quando a metade dos dados utilizados para o treinamento da mesma rede foi propagado. Esta rede atingiu

um erro de 0,0025038 em 2662 épocas e teve seu treinamento encerrado quando o gradiente que alcançou o limiar de $1x10^{-6}$. O desempenho geral obtido no reconhecimento ficou em 97,1%.

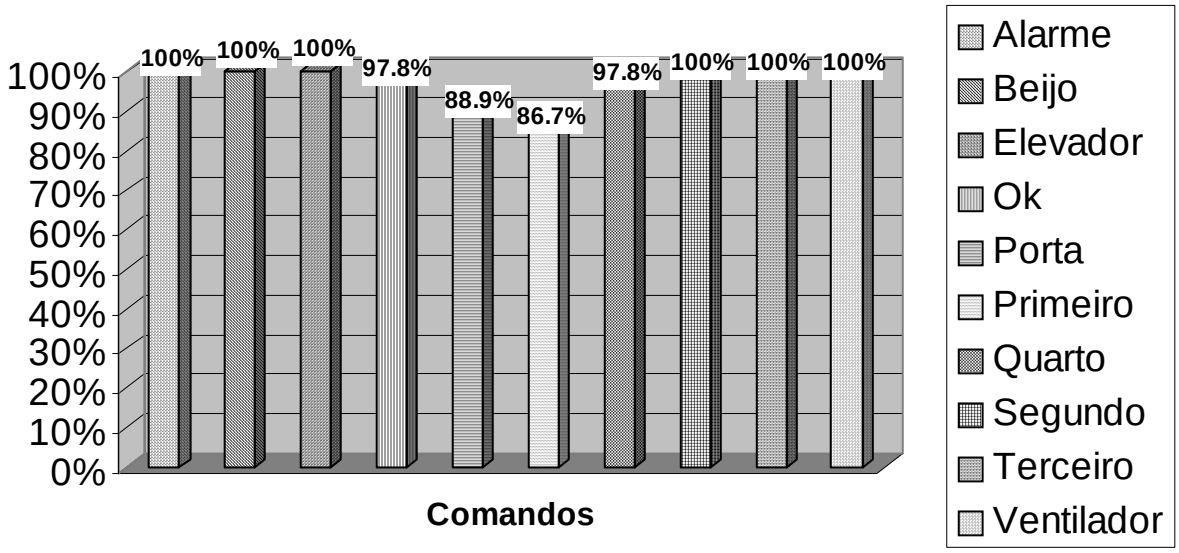


FIGURA - 5.18 Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes da metade do conjunto de amostras utilizadas no treinamento da base de comandos do LaPSI

Quando aplicada a outra metade dos dados, ou seja, a metade considerada como dados referentes a usuários desconhecidos que não foram utilizados no treinamento, ocorreu uma queda significativa no percentual de desempenho geral, atingindo 64,9% de acerto. As taxas de acerto individuais de cada comando são apresentada na fig. 5.19.

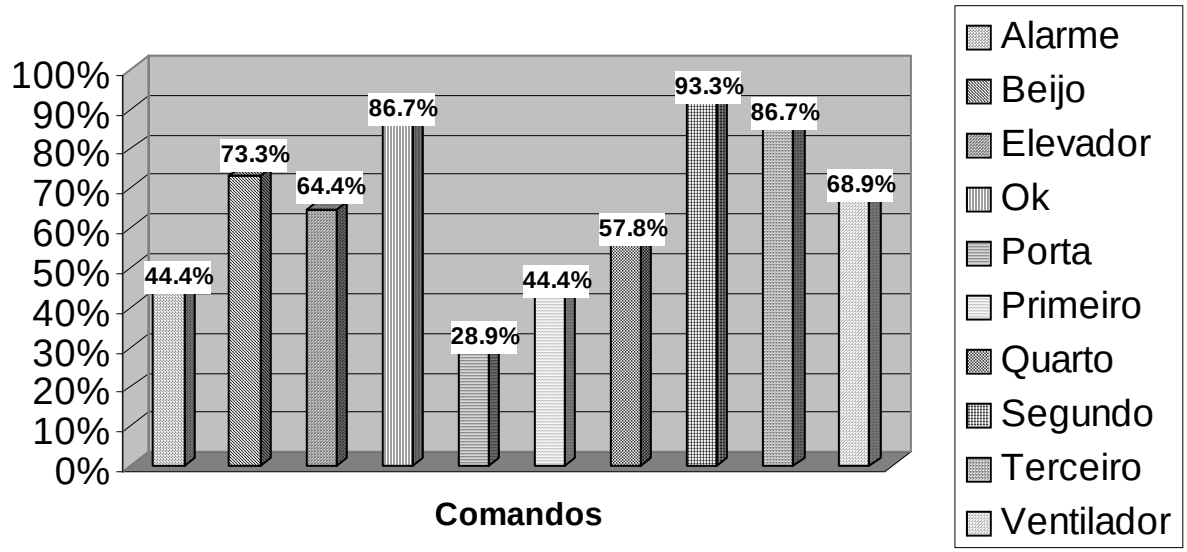


FIGURA - 5.19 Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes da metade de amostras desconhecidas do conjunto da base de comandos do LaPSI

O treinamento com 70% dos dados do conjunto original produziu um erro de 0,001617 em 2922 épocas, que teve seu término quando o limiar do gradiente ficou abaixo de $1x10^{-6}$. A propagação dos dados utilizados neste treinamento obteve um desempenho geral de 97,6% e os resultados individuais relativos a cada comando são visualizados na fig. 5.20.

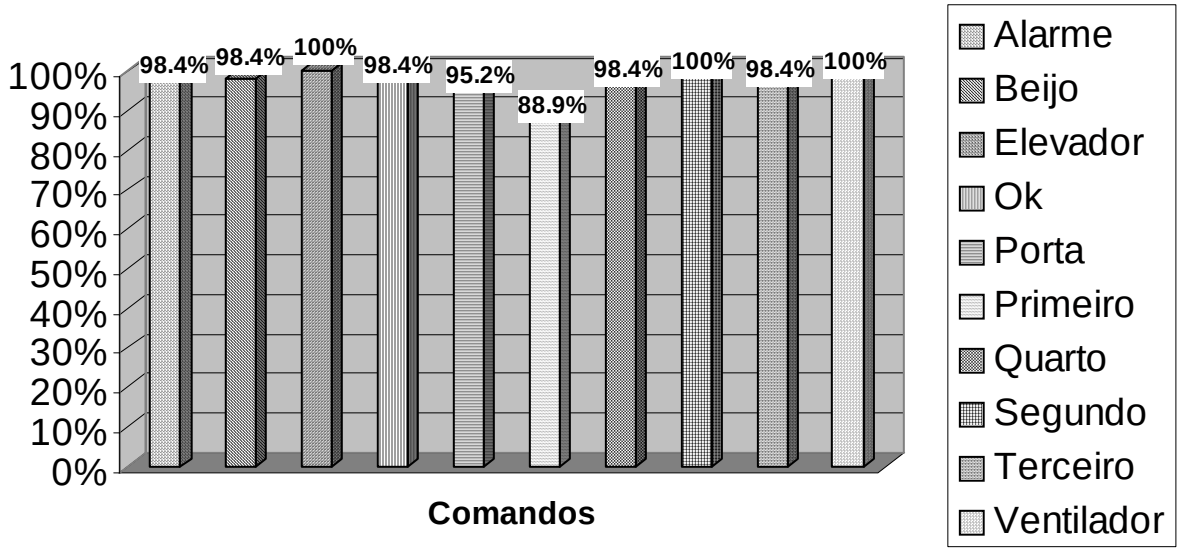


FIGURA - 5.20 Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes do treinamento de 70% do conjunto da base de comandos do LaPSI

Assim como a propagação da metade dos dados desconhecidos pela rede anterior, também foram propagados os 30% dos dados destinados a teste, que foram reservados do conjunto original neste treinamento, obtendo resultados satisfatórios. Os resultado individuais de cada comando são apresentados na fig. 5.21 e o desempenho geral de acerto obtido na propagação destes dados atingiu 64,4% de reconhecimento.

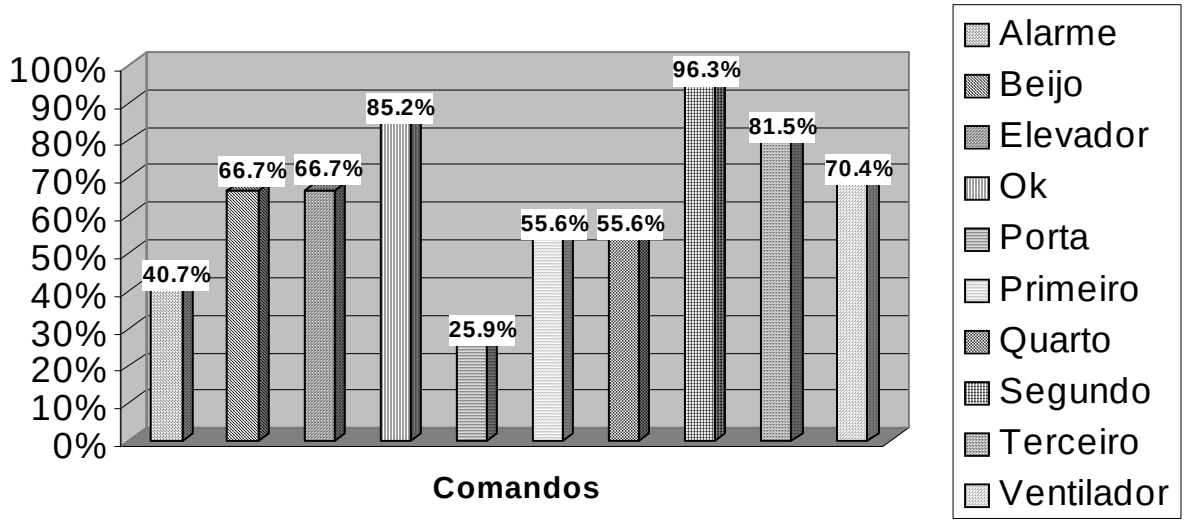


FIGURA - 5.21 Gráfico de resultados obtidos com a rede BP sobre comandos de voz provenientes de 30% de amostras desconhecidas do conjunto da base de comandos do LaPSI

5.3.2 Comandos da Base do LaPSI Aplicados à *Fuzzy ARTMAP*

Tanto o treinamento com o conjunto original das amostras da base de dados do LaPSI, quanto o treinamento utilizando metade do conjunto original proporcionou a criação de 47 categorias, que para o conjunto original produziu: 6 categorias para o comando **alarme**, 6 para **beijo**, 6 para **elevador**, 3 para **ok**, 6 para **porta**, 6 para **primeiro**, 4 para **quarto**, 2 para **segundo**, 3 para **terceiro** e 5 para **ventilador**; e para a metade dos dados: 7 categorias para o comando **alarme**, 5 para **beijo**, 2 para **elevador**, 2 para **ok**, 8 para **porta**, 7 para **primeiro**, 5 para **quarto**, 3 para **segundo**, 1 para **terceiro** e 7 para **ventilador**. Enquanto que o treinamento com 70% do conjunto original produziu 42 categorias, distribuídas em: 5 categorias para o comando **alarme**, 3 para **beijo**, 5 para **elevador**, 4 para **ok**, 6 para **porta**, 5 para **primeiro**, 4 para **quarto**, 2 para **segundo**, 2 para **terceiro** e 6 para **ventilador**. Todos os conjuntos treinados obtiveram um reconhecimento de 100% sobre a propagação dos próprios dados utilizados no treinamento.

Na apresentação da metade dos dados do conjunto original, composto por amostras de usuários que não participaram do treinamento, sobre a rede treinada com a outra metade, o desempenho geral obtido pela rede foi de 60,2% com a taxa de reconhecimento individual de cada comando apresentada na fig. 5.22.

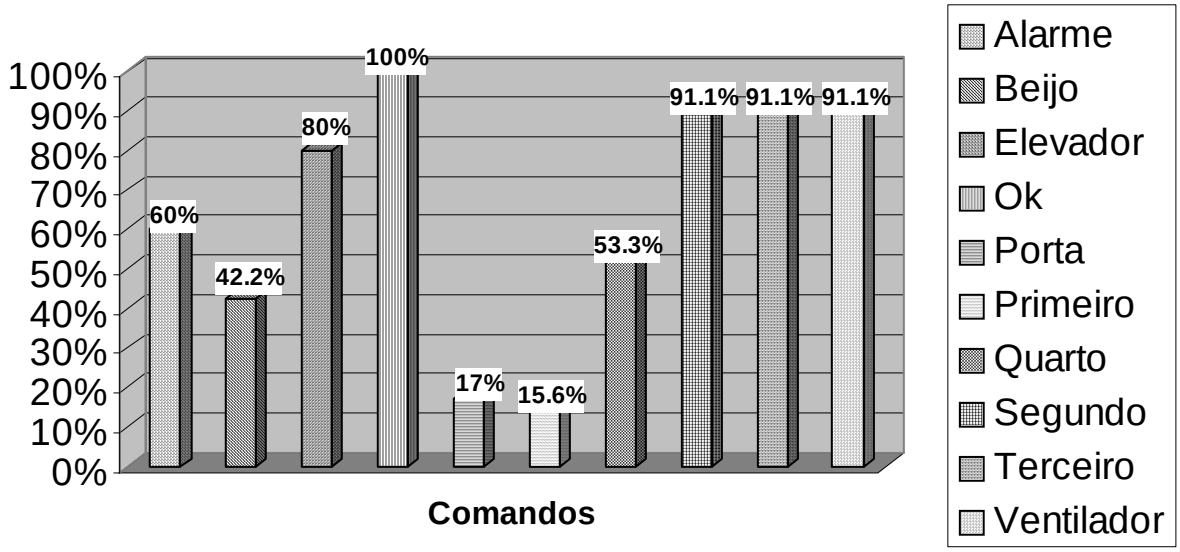


FIGURA - 5.22 Gráfico de resultados obtidos com a rede *Fuzzy ARTMAP* sobre comandos de voz provenientes da metade de amostras desconhecidas do conjunto da base de comandos do LaPSI

Já na apresentação dos 30% dos dados do conjunto original, que não teve seus usuários inseridos no treinamento, a rede obteve uma taxa de acerto de 68,5% com percentuais individuais dos comandos apresentados na fig. 5.23

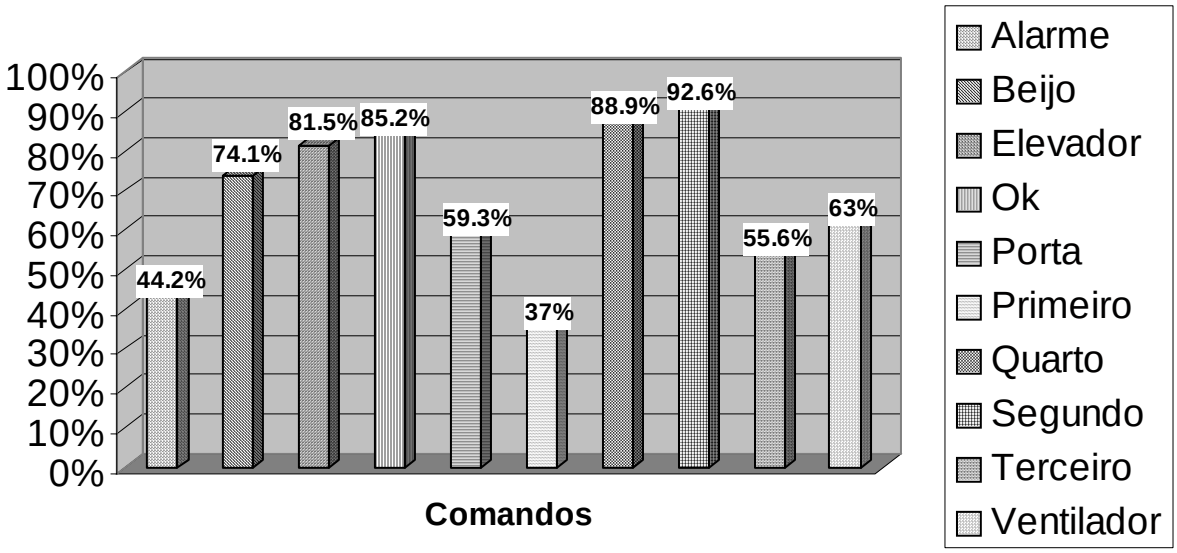


FIGURA - 5.23 Gráfico de resultados obtidos com a rede *Fuzzy ARTMAP* sobre comandos de voz provenientes de 30% de amostras desconhecidas do conjunto da base de comandos do LaPSI

5.4 Comparações e Comentários sobre os Resultados Obtidos

Os resultados obtidos, Tab. A2 no Anexo 2, sobre as bases aplicadas à metodologia, de extração dos coeficientes cepstrais e a apresentação destes às redes neurais, proposta por este trabalho, apresentaram resultados semelhantes aos obtidos com os experimentos relativos a criação de cada base utilizada para avaliação dos métodos empregados neste experimento.

Petry relata que obteve uma taxa de reconhecimento de comandos superior a 97%[PET 99]. Os métodos utilizados para obtenção de tais resultados foram a utilização de coeficientes *mel*-cepstrais e a técnica de Quantização Vetorial Multisecção (MSVQ). Embora o foco do trabalho em que foi aplicada a base de dados do Projeto REVOX seja relativo a identificação de locutores, os resultados obtidos neste trabalho sobre a mesma base são similares comprovando a eficácia dos métodos utilizados.

Uma comparação interessante é que as amostras da base do REVOX e as da base de comandos de direcionamento possuem a mesma configuração, ou seja, foram amostradas a 11025 Hz, 16 bits em monocanal. A única diferença entre ambas é o uso do filtro passa baixas. Isto leva a suposição que a utilização de tal filtro, na captura das amostras que compuseram a base de comandos de direcionamento, poderia elevar a taxa de reconhecimento para esta base em 10%.

Quanto ao resultado apresentado com os experimentos desenvolvidos e aplicados pelo LaPSI sobre o conjunto de 10 comandos de sua própria base de treinamento foi obtido 97% de reconhecimento, utilizando coeficientes cepstrais derivados da análise dos Coeficientes de Predição Linear (LPC) submetidos aos HMM

com algoritmo reconhecedor de Viterbi[LAB 95]. Alguns testes utilizando dados parciais desta base foram realizados no sentido de tentar uma classificação de dados que poderiam, de certa forma, serem considerados de usuários desconhecidos, uma vez que determinada parte destes dados não estava presente no conjunto reservado para treinamento das redes neurais. A taxa de reconhecimento obtida sobre a propagação dos dados utilizados para o treinamento obteve índices de acerto similares. No entanto, realizando outros testes que consideravam informações parciais desta base como dados desconhecidos, as taxas de acerto baixaram significativamente. É importante destacar que, um dos motivos que levou à obtenção de índices pouco expressivos foi a alta freqüência de amostragem das amostras que compõem esta base, associada ao método de extração dos 1400 coeficientes cepstrais de cada amostra, proposto por este trabalho. Provavelmente o número de coeficientes cepstrais determinado não foi suficiente para obter algumas características essenciais destes dados.

Esta análise comparativa nos leva a crer que a queda de desempenho no reconhecimento da base de comandos de direcionamento está associada à geração das amostras, mais especificamente em relação a não utilização do filtro passa baixas e ao posicionamento do microfone. Pois diferenças na amplitude das amostras de um mesmo locutor relativas ao mesmo comando, em algumas situações, apresentaram grande variação. Ainda assim, as taxas de reconhecimento obtidas demonstraram ser satisfatórias.

Algumas soluções que poderiam ser aplicadas para elevar as taxas de reconhecimento da base de comandos de direcionamento dizem respeito a reconstrução dessa base, com:

- utilização da placa *anti-alias*;
- um posicionamento fixo do microfone;
- utilização de um microfone de melhor qualidade;
- ampliação do número de usuários e consequentemente no número de amostras;
- maior controle sobre o ruído do ambiente de gravação das amostras;
- utilização de outras técnicas de processamento de sinais que caracterizassem melhor o sinal;
- utilização de outro método em conjunto com redes neurais, por exemplo HMM para a classificação dos dados.

6 Conclusão e Considerações Finais

Este trabalho cumpriu o seu propósito, o desenvolvimento em computador de uma aplicação para o reconhecimento de um vocabulário restrito de comandos de direcionamento provenientes da fala e independentes do locutor.

Para atingir esta meta, foi necessário um estudo detalhado do estado da arte das pesquisas que já tratavam esta tecnologia ainda recente, buscando uma visão, e por que não dizer a noção dos fundamentos do reconhecimento de voz.

Dando seqüência ao trabalho determinou-se quais técnicas e métodos seriam aplicados. No sentido de fazer uma aplicação compacta e ao mesmo tempo confiável, optou-se pela extração dos coeficientes cepstrais de amostras sonoras e o treinamento e classificação por redes neurais, em especial os modelos *Backpropagation* e *Fuzzy ARTMAP*.

Um ponto importante e interessante na realização deste trabalho foi a criação da base de treinamento e de amostras de teste, com uma relação direta com os usuários que não somente concederam sua voz para compor um banco de amostras de fundamental importância, mas que também opinaram e questionaram, demostrando interesse em aprender um pouco mais dos propósitos deste trabalho.

A utilização do *software* Matlab, recentemente adquirido e difundido entre os alunos desta instituição, foi de grande valia para o desenvolvimento do trabalho. Suas funções dedicadas, representadas por suas *Toolboxes*, mais especificamente as relacionadas ao processamento de sinais e às redes neurais artificiais, demostraram-se muito úteis, principalmente pela agilidade de manipulação, geração de um código fonte compacto e inteligível e pela velocidade de processamento da informação, em especial no treinamento das redes neurais que revela-se uma tarefa bastante custosa quando trata-se de uma grande quantidade de dados.

Os resultados obtidos por meio dos testes realizados foram considerados bons. Como era esperado, as amostras relativas ao conjunto de treinamento da base de comandos de direcionamento utilizadas no treinamento das redes foram melhores identificadas. Como é o caso da rede *Backpropagation* R2 que obteve uma taxa de 86% de reconhecimento para amostras de treinamento, contra 82,8% de acerto para amostras de teste desconhecidas ao modelo, demonstrando uma queda pouco significativa se comparada a variabilidade das informações. De modo geral, se comparado ao modelo *Fuzzy ARTMAP* nota-se que este é mais sensível à variação entre as amostras que o modelo BP que apresenta uma maior robustez.

A utilização de outras bases sonoras para validação do sistema proposto contribuiu no sentido de assegurar a efetividade das implementações realizadas. Embora o enfoque dos experimentos realizados com essas bases foi um pouco diferente do abordado neste trabalho, a comparação entre os resultados obtidos frente aos demais trabalhos que efetivaram o reconhecimento, demonstrou um bom índice de classificação dos dados pela metodologia aqui aplicada.

A questão da independência do locutor no reconhecimento de voz para palavras isoladas é um fator difícil de ser tratado, uma vez que as diversas locuções relativas a um determinado comando fornecidas por um único locutor não apresentam uma uniformidade desejada, o que torna o problema ainda mais complexo quando diz respeito ao desenvolvimento de sistemas deste tipo com a independência de locutores.

Outro tópico importante a ser relatado é o uso de usuários de ambos os sexos, o que causa uma considerável diferença na variação da freqüência dos sons emitidos entre homens e mulheres, ressaltando que dentre diversos trabalhos consultados, a maioria de suas bases de dados eram compostas exclusivamente por usuários masculinos.

As taxas de reconhecimento obtidas foram consideradas boas, mas não excelentes como sempre se almeja. Como algumas sugestões para incrementar os índices de acerto obtidos destacam-se: a utilização de um filtro passa baixas na aquisição dos dados; a imposição de algumas regras no processo de captura do som, assim como uma melhor qualidade do equipamento utilizado; ampliação do universo de locutores; e a utilização de técnicas mais sofisticadas que a obtenção dos coeficientes cepstrais. Também é importante destacar, que talvez, um treinamento diferenciado entre o sexo dos locutores poderia contribuir a elevar os atuais índices resultantes.

De uma forma geral, deseja-se que este trabalho venha a contribuir com as pesquisas relativas ao reconhecimento de voz, estendendo a utilização das redes neurais, que embora tenham seu uso bastante difundido em diversas áreas ainda ocupam uma discreta participação na resolução de problemas relacionados ao reconhecimento da fala.

Apesar de alguns obstáculos encontrados, o trabalho demostrou-se de uma validade inestimável para o autor. Embora apresentando diversas dificuldades aliadas à complexidade presente na maioria dos trabalhos que envolvem a pesquisa, o mesmo possibilitou ao autor desta dissertação a descoberta e o conhecimento de uma área muito interessante, que apresenta infinitas possibilidades de aplicação e grandes perspectivas para o futuro.

6.1 Trabalhos Futuros

Como trabalhos futuros é importante destacar a realização das modificações já sugeridas neste capítulo, o que deve propiciar uma elevação na taxa de reconhecimento, permitindo uma maior flexibilidade do sistema em relação ao universo de locutores.

Também é possível a extensão do vocabulário com a adição de novos comandos, enriquecendo o vocabulário e permitindo a sua utilização na realização de demais tarefas.

Um ponto que merece destaque e que é uma carência dos sistemas com o mesmo propósito deste trabalho, principalmente os elaborados no Brasil e que utilizam nossa língua nativa para treinamento e testes, é a inexistência de uma base de dados sólida que sirva como referência no desenvolvimento de sistemas desta natureza. Sua criação e utilização iria propiciar uma validação com melhor qualidade e consequentemente um maior grau de confiabilidade aos trabalhos, pois quando um determinado sistema de

reconhecimento de voz fosse proposto, seu idealizador teria melhores subsídios e garantia alegando que seu sistema foi validado com a referida base de dados. Isto serviria como um selo de qualidade aos novos sistemas de reconhecimento de voz.

Quanto a aplicações futuras é permitido sonhar um pouco, indicando aplicações que podem ser realizadas ou que já são realidade, como:

- controle de diversos utensílios e elétro-eletrônicos;
- controle de automóveis;
- manipulação de robôs;
- controle de ambientes;
- interagir com jogos de computador, permitindo ao usuário o desenvolvimento de um diálogo com demais personagens do jogo.

Todas as aplicações visam o bem estar e comodidade de seus usuários, assim como agilidade na realização de determinadas tarefas. Servindo como suporte as pessoas com impedimentos físicos tornando-os mais independentes em atividades pessoais.

A gama de facilidade que a voz pode ofertar encontra seu limite na imaginação de quem desenvolve sistemas desta natureza.

Anexo 1- Resultado da Saída da Rede R2

Este anexo apresenta alguns resultados provenientes da propagação dos dados de treinamento da base de amostras sonoras produzida por este trabalho sobre a rede neural *Backpropagation* referenciada como R2 nesta dissertação.

É apresentado um conjunto de amostras referente a um determinado comando. E a classificação apresenta o local e o nome do arquivo lido, os comando com o valor de saída correspondente seguida da classificação.

```
Apresentação de amostras do comando "DIREITA"

======================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_d1.wav
======================================================================
direita   esquerda  siga      pare      recue
0.9966    0.0000    0.0004    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
======================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_d2.wav
======================================================================
direita   esquerda  siga      pare      recue
0.9985    0.0000    0.0000    0.0002    0.0000

A locução reconhecida é da palavra "DIREITA"
======================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_d3.wav
======================================================================
direita   esquerda  siga      pare      recue
0.9998    0.0002    0.0000    0.0005    0.0000

A locução reconhecida é da palavra "DIREITA"
======================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_d4.wav
======================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0002    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
======================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_d5.wav
======================================================================
direita   esquerda  siga      pare      recue
0.9974    0.0001    0.0000    0.0065    0.0000

A locução reconhecida é da palavra "DIREITA"
======================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_d1.wav
======================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0005    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_d2.wav
========================================================================
    direita   esquerda  siga      pare      recue
    1.0000    0.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_d3.wav
========================================================================
direita   esquerda  siga      pare      recue
0.9956    0.0000    0.0035    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_d4.wav
========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0000    0.0000    0.0001

A locução reconhecida é da palavra "DIREITA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_d5.wav
========================================================================
direita   esquerda  siga      pare      recue
0.9956    0.0000    0.0045    0.0000    0.0001

A locução reconhecida é da palavra "DIREITA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_d1.wav
========================================================================
direita   esquerda  siga      pare      recue
0.9977    0.0000    0.0011    0.0001    0.0000

A locução reconhecida é da palavra "DIREITA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_d2.wav
========================================================================
direita   esquerda  siga      pare      recue
0.9998    0.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_d3.wav
========================================================================
direita   esquerda  siga      pare      recue
    0.0000    0.0000    0.0015    0.0000    0.0011

A locução não foi reconhecida
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_d4.wav
========================================================================
direita   esquerda  siga      pare      recue
0.9969    0.0000    0.0000    0.0000    0.0001

A locução reconhecida é da palavra "DIREITA"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_d5.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9970    0.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_d6.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9960    0.0000    0.0001    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_d7.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9967    0.0034    0.0000    0.0000    0.0006

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_d8.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_d9.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9970    0.0003    0.0000    0.0000    0.0038

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_d10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9998    0.0020    0.0000    0.0000    0.0031

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU13\U13_d1.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU13\U13_d2.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0002    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU13\U13_d3.wav
==========================================================================
direita   esquerda  siga      pare      recue
    0.2105    0.0017    0.0064    0.0000    0.0000

A locução não foi reconhecida
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU13\U13_d4.wav
==========================================================================
direita   esquerda  siga      pare      recue
    1.0000    0.0011    0.0000    0.0000    0.0002

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU13\U13_d5.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0427    0.0001    0.0004    0.0000    0.1963

A locução não foi reconhecida
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU13\U13_d6.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0259    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU13\U13_d7.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9954    0.0002    0.0002    0.0000    0.0001

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU13\U13_d8.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9997    0.0000    0.0017    0.0000    0.0001

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU13\U13_d9.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.3423    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU13\U13_d10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9982    0.0001    0.0049    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_d1.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0002    0.0000    0.0003

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_d2.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0015    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_d3.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_d4.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9982    0.0002    0.0020    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_d5.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0000    0.0000    0.0002

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_d6.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0009    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_d7.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_d8.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0037    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_d9.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9996    0.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_d10.wav
==========================================================================
direita   esquerda  siga      pare      recue
1.0000    0.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "DIREITA"
```

```
Apresentação de amostras do comando "ESQUERDA"


==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_e1.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_e2.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0000    0.0001    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_e3.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0000    0.0000    0.0001

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_e4.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9999    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_e5.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9999    0.0000    0.0004    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_e6.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_e7.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0000    0.0001    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_e8.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_e9.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0352    0.0024    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU1\U1_e10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_e1.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9927    0.2633    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_e2.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_e3.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9997    0.0006    0.0000    0.0002

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_e4.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_e5.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0021    0.9996    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_e6.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0002    0.9654    0.0001    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_e7.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0001    0.0348    0.8489    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_e8.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9985    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_e9.wav
========================================================================
direita   esquerda  siga      pare      recue
0.2177    0.3797    0.0000    0.0000    0.0000

A locução não foi reconhecida
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU2\U2_e10.wav
========================================================================
direita   esquerda  siga      pare      recue
0.9485    0.9729    0.0000    0.0000    0.0001

A locução reconhecida é da palavra "ESQUERDA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_e1.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9993    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_e2.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    1.0000    0.0000    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_e3.wav
========================================================================
direita   esquerda  siga      pare      recue
    0.0000    1.0000    0.0007    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_e4.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9961    0.0000    0.0000    0.5363

A locução reconhecida é da palavra "ESQUERDA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_e5.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0079    0.5078    0.0013    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_e6.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9980    0.0002    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_e7.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9969    0.9547    0.0000    0.0000

A locução reconhecida é da palavra "ESQUERDA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_e8.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0077    0.0134    0.0000    0.1489

A locução não foi reconhecida
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_e9.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0002    0.0134    0.0000    0.0002

A locução não foi reconhecida
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU7\U7_e10.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0008    0.1429    0.0001    0.0000    0.1160

A locução não foi reconhecida
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU9\U9_e1.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9798    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU9\U9_e2.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.9616    0.9933    0.0000    0.0001    0.0000

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU9\U9_e3.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9928    0.0000    0.0000    0.0019

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU9\U9_e4.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0028    0.8925    0.0000    0.0002

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU9\U9_e5.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9602    0.0018    0.0000    0.2566

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU9\U9_e6.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0019    0.0151    0.0353    0.0000    0.0000

A locução não foi reconhecida
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU9\U9_e7.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0001    0.3932    0.0000    0.0000    0.0107

A locução não foi reconhecida
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU9\U9_e8.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.1185    0.9468    0.0000    0.0094

A locução reconhecida é da palavra "SIGA"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU9\U9_e9.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.9985    0.0000    0.0000    0.0012

A locução reconhecida é da palavra "ESQUERDA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU9\U9_e10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0274    0.9977    0.0000    0.0000    0.2232

A locução reconhecida é da palavra "ESQUERDA"
```

```
Apresentação de amostras do comando "SIGA"


==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU10\U10_s8.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU10\U10_s9.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0001    0.3371    0.0442    0.0000    0.0000

A locução não foi reconhecida
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU10\U10_s10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    1.0000    0.0000    0.0001

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_s1.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.1342    0.1908    0.0000    0.0000    0.0065

A locução não foi reconhecida
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_s2.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0009    0.0000    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_s3.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_s4.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0039    0.0000    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_s5.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_s6.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.5794    0.9994    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_s7.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0011    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_s8.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0001    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_s9.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0048    0.9997    0.0003    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_s10.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU16\U16_s1.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0007    0.9999    0.0000    0.0001

A locução reconhecida é da palavra "SIGA"
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU16\U16_s2.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0008    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU16\U16_s3.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0002    0.0120    0.9913    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU16\U16_s4.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0506    0.9878    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU16\U16_s5.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0081    0.9954    0.0040    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU16\U16_s6.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0003    0.0001    0.9973    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU16\U16_s7.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0001    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU16\U16_s8.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0039    1.0000    0.0000    0.0039

A locução reconhecida é da palavra "SIGA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU16\U16_s9.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU16\U16_s10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0009    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_s1.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0006    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_s2.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0007    0.9992    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_s3.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0032    0.9974    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_s4.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0003    1.0000    0.0004    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_s5.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0002    0.9998    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_s6.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0011    0.0002    0.9935    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_s7.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0088    0.9977    0.0024    0.0000

A locução reconhecida é da palavra "SIGA"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_s8.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0430    0.9999    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_s9.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0009    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_s10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0010    1.0000    0.0000    0.0000

A locução reconhecida é da palavra "SIGA"
```

```
Apresentação de amostras do comando "PARE"


==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU5\U5_p1.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0006

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU5\U5_p2.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU5\U5_p3.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU5\U5_p4.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0001

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU5\U5_p5.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0048

A locução reconhecida é da palavra "PARE"
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU5\U5_p6.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0003

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU5\U5_p7.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0037

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU5\U5_p8.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU5\U5_p9.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU5\U5_p10.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU8\U8_p1.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU8\U8_p2.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0001    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU8\U8_p3.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU8\U8_p4.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0004

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU8\U8_p5.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.9993    0.0007

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU8\U8_p6.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU8\U8_p7.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0064    0.0000    0.0000    0.9483    0.0020

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU8\U8_p8.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU8\U8_p9.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU8\U8_p10.wav
========================================================================
direita   esquerda  siga      pare      recue
0.2534    0.1872    0.0000    0.0006    0.0000

A locução não foi reconhecida
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_p1.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0008    0.0287    0.0266    0.0091

A locução não foi reconhecida
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_p2.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0016

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_p3.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_p4.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_p5.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0001    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_p6.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.8592    0.6789

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_p7.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_p8.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0001    0.0003    0.0005    0.9953    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_p9.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0039    0.9982    0.0000

A locução reconhecida é da palavra "PARE"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU12\U12_p10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0003

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_p1.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_p2.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0021

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_p3.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_p4.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_p5.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_p6.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_p7.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_p8.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0001    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_p9.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU20\U20_p10.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    1.0000    0.0000

A locução reconhecida é da palavra "PARE"
```

```
Apresentação de amostras do comando "RECUE"


========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU3\U3_r1.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0024    0.0000    0.0000    0.9994

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU3\U3_r2.wav
========================================================================
direita   esquerda  siga      pare      recue
0.9831    0.0000    0.0000    0.0000    0.9606

A locução reconhecida é da palavra "DIREITA"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU3\U3_r3.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU3\U3_r4.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0014    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU3\U3_r5.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0001    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU3\U3_r6.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU3\U3_r7.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0019    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU3\U3_r8.wav
========================================================================
direita   esquerda  siga      pare      recue
0.1294    0.0000    0.0000    0.0000    0.0106

A locução não foi reconhecida
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU3\U3_r9.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0001    0.0000    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU3\U3_r10.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU6\U6_r1.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0001    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU6\U6_r2.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0006    0.0000    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU6\U6_r3.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0001    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU6\U6_r4.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0012    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU6\U6_r5.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0028    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU6\U6_r6.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0058    0.0010    0.0000    0.0000    0.9793

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU6\U6_r7.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0003    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU6\U6_r8.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0010    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU6\U6_r9.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.4193    0.0201    0.0000    0.9901

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU6\U6_r10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0004    0.0001    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU14\U14_r1.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0009    0.9989

A locução reconhecida é da palavra "RECUE"
```

```
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU14\U14_r2.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0000    0.9999

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU14\U14_r3.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0001    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU14\U14_r4.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0010    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU14\U14_r5.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0004    0.9972

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU14\U14_r6.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0013    0.9997

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU14\U14_r7.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU14\U14_r8.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU14\U14_r9.wav
========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0001    0.0000    0.0002    1.0000

A locução reconhecida é da palavra "RECUE"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU14\U14_r10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_r1.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0004    0.9999

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_r2.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0014    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_r3.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0003    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_r4.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0002    0.9996

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_r5.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0077    0.9933

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_r6.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_r7.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0003    1.0000

A locução reconhecida é da palavra "RECUE"
```

```
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_r8.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0007    0.0000    0.0000    1.0000

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_r9.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0005    0.9999

A locução reconhecida é da palavra "RECUE"
==========================================================================
Arquivo lido: e:\jvaliati\mat\Usersvoz\UserU19\U19_r10.wav
==========================================================================
direita   esquerda  siga      pare      recue
0.0000    0.0000    0.0000    0.0018    1.0000

A locução reconhecida é da palavra "RECUE"
```

Anexo 2- Tabela de Resultados

Este anexo apresenta uma tabela com os resultados da classificação das bases de dados: de comandos de direcionamento, dados parciais da base do projeto REVOX e base de dados do LaPSI. Os percentuais apresentados representam a taxa de reconhecimento geral que as redes *Backpropagation* e *Fuzzy ARTMAP* obtiveram na validação dos dados.

TABELA A2 – Taxas de reconhecimento das redes neurais sobre as bases de dados

Dados Aplicados		Taxa de Reconhecimento	
		Backpropagation	Fuzzy ARTMAP
Base de dados de comandos de direcionamento	Treino (1000 amostras)	86%	100%
	Teste (250 amostras)	82,8%	65,2%
Dados parciais do REVOX	Treino (300 amostras)	98,9%	100%
	Teste (150 amostras)	88%	90,7%
Base do LaPSI	Treino (900 amostras)	96,6%	100%
Metade da base do LaPSI	Treino (450 amostras)	97,1%	100%
	Teste (450 amostras)	64,9%	60,2%
Base do LaPSI	70% da base (Treino)	97,6%	100%
	30% da base (Teste)	64,4%	68,5%

Bibliografia

[BRA 98] BRAGA, A.P.; CARVALHO, A.; LUDERMIR, T.B. Fundamentos de Redes Neurais Artificiais. In: ESCOLA DE COMPUTAÇÃO, 9., 1998. **Anais...** Rio de Janeiro: DCC/IM, 1998.

[BEZ 94] BEZERRA, M.R. **Reconhecimento Automático do Locutor para Fins Forenses, Utilizando Técnicas de Redes Neurais**. Rio de Janeiro: IME, 1994. Dissertação de Mestrado.

[CAR 91] CARPENTER, G.A.; GROSSBERG, S.; ROSEN, D.B. Fuzzy ART: Fast stable learning and categorization of analog patterns by adaptive resonance system. **Neural Networks**, v. 4, p. 759-771, 1991.

[CAR 92] CARPENTER, G.A.; GROSSBERG, S.; MARKUZON, N.; REYNOLDS, J.H.; ROSEN, D.B. Fuzzy ARTMAP: A neural network architecture for incremental supervised learning af analog multidemensional maps. **IEEE Transactions on Neural Networks**, v. 3, n. 5, p. 698-713, Sept. 1992.

[DEL 93] DELLER J.R.; PROAKIS J.G.; HANSEN J.H.L. **Discrete-time Processing of Speech Signals**. New Jersey: Prentice-Hall, 1993. 908p.

[DIN 97] DINIZ, S.S. **Uso de Técnicas Neurais para o Reconhecimento de Comandos à Voz**. Rio de Janeiro: IME, 1997. Dissertação de Mestrado.

[ENG 99] ENGEL, P.M. **O Modelo ART**. [S.l.: s.n.], 1999. Lâminas de Aula da Disciplina de Redes Neurais do PPGC da UFRGS.

[FAN 86] FANZERES, A. **Fatores Subjetivos da Audição Humana e A Física e Psicologia do Nosso Ouvido**. [S.l.]: Nova Eletrônica, 1986.

[FOL 68] FOLMER, J.; TORE N.O. **Oscilações, Ondas, Acústica**. São Paulo: Nobel, 1968.

[FUR 89] FURUI, S. **Digital Speech Processing, Synthesis, and Recognition**. New York: Marcel Dekker, 1989, 390p.

[GUA 93] GUAZZELLI, A. **Do ART 1 ao Fuzzy ARTMAP: um estudo sobre modelos de redes neurais artificiais baseados na teoria da adaptação ressonante (ART)**. Porto Alegre: CPGCC da UFRGS, 1993. 61p.

[GUA 94] GUAZZELLI, A. **Aprendizagem em Sistemas Híbridos**. Porto Alegre: CPGCC da UFRGS, 1994. Dissertação de Mestrado.

[HAR 83] HART, J. et al. Manipulação de Sons da Fala. **Ciência e Técnica**, [S.l.], v.2, n. 6, p. 179-191, 1983.

[HAR 89] HARKINS, J. Voice Processing and Disabled People: A Simbiotic Relationship. **Internacional Voice System Review**, [S.l.], 1989.

[HEB 49] HEBB, D.O. **The Organization of Behavior**. New York: Wiley, 1949.

[HEC 90] HECTH-NIELSEN, R. **Neurocomputing**. Massachusetts: Addison Wesley, 1990.

[HEL 67] HELMS, H.D. Fast Fourier Transform Method of Computing Diferrence Equations and Simulating Filters. **IEEE Transactions on Audio and Eletroacoustics**, [S.l.], v.15, n.2, p. 85-90, 1967.

[HER 91] HERTZ, J.; KROGH, A.; PALMER, R.G. **Introduction to Theory of Neural Computation**. Redwood City, CA: Addison-Wesley, 1991.

[KOV 96] KOVÁCS, Z.L. **Redes Neurais:** Fundamentos e Aplicações. 2. ed. São Paulo: Ed. Acadêmica, 1996.

[LAB 95] LABORATÓRIO DE PROCESSAMENTO DE SINAIS E IMAGENS. **Implementação de um Sistema de Controle Vocal de Equipamentos de Automação Industrial**. Porto Alegre: LaPSI, 1995. 25p.

[LOE 96] LOESCH, C.; SARI, S.T. **Redes Neurais Artificiais Fundamentos e Modelos**. Blumenal: Ed. da FURB, 1996.

[LUF 91] LUFT, J.A. **Reconhecimento automático de voz para palavras isoladas e independentes do locutor**. Porto Alegre: PPGEMM da UFRGS, 1994. Dissertação de Mestrado.

[MAL 85] MALVINO, A.P. **Microcomputadores e Microprocessadores**. São Paulo: McGraw-Hill, 1985.

[MAR 96] MARKOWITZ, J. **Using Speech Recognition**. New Jersey: Prentice-Hall, 1996. 292p.

[MCC 43] McCULLOCH, W.S.; PITTS, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. **Bulletin of Mathematical Biophysics**, [S.l.], p. 115-133, 1943.

[MEN 70] MENDEL, J.M.; McLAREN, R.W. Reinforcement Learning Control and Pattern Recognition Systems. In: MENDEL, J. M.; FU, K. S. (Eds.). **Adaptative, Learning and Pattern Recognition Systems:** Theory and Applications. New York: Academic Press, 1970. p. 287-318.

[MIN 69] MINSKY, M.; PAPERT, S. **Perceptrons:** an introduction to computational geometry. Massachusetts: MIT Press, 1969.

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

[PET 99] PETRY, A.; BARONE, D.A.C. Sistema para Controle de Elevadores por Voz. In: CONFERENCIA LATINOAMERICANA DE INFORMÁTICA, CLEI, 25., 1999, Asuncion. **Memorias...**Asuncionn: Universidad Autonoma de Asuncion, 1999. v.1, p. 63-74.

[RAB 75] RABINER, L. R.; SAMBUR, M. R. An Algorithm for Determining the Endpoints of Isolated Utterances. **Bell System Journal**, [S.l.], v.54, n.2, p. 297-315, 1975.

[RAB 75a] RABINER, L.R.; GOLD, B. **Theory and Application of Digital Signal Processing**. Englewood Cliffs: Prentice-Hall, 1975. 762p.

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

[ROC 87] ROCHA, L. Processamiento de Voz. In: ESCOLA BRASILEIRO-ARGENTINA DE INFORMÁTICA, EBAI, 1., 1987. **Anais...** [S.l. : s.n.], 1987. 85p.

[RUM 86] RUMELHART, D.E.; HILTON, G.E.; WILLIAMS, J. Learning Representations by Error Propagation. In: RUMELHART, D.; McCLELLAND, J.(Eds.). **Parallel Distributed Processing Explorations in the Microstructure of Cognition**. Cambridge, MA: Mit Press, 1986, v.1, p. 318-362.

[SCO 94] SCOTT, N. Speech Recognition for Individuals with Disabilities. In: ADVANCED SPEECH APLICATIONS AND TECHNOLOGIES CONFERENCE, 1994. **Proceedings...** [S.l. : s.n.], 1994.

[SHA 82] SHALCK, T.; McMAHAN, M. Firmware-programmable C Aids Speech Recognition. **Eletronic Desing**, Cleveland, v.30, n.15, p. 143-147, July 1982.

[SIM 90] SIMPSON, P. K. **Artificial Neural Systems**. New York: Pergamon Press, 1990.

[STO 66] STOCKHAM, T.G. High-Speed Convulation and Correlation. In: SPRING JOINT COMPUTER CONFERENCE, 1966. **Proceedings...** [S.l. : s.n.], 1966. v.28, p. 229-233.

[TAF 95] TAFNER, M.A. **Reconhecimento de Palavras Faladas Isoladas Usando Redes Neurais Artificiais**. Florianópolis: PPGEP da UFSC, 1995. Dissertação de Mestrado.

[TAU 82] TAUB, H., SCHILLING, D. **Eletrônica Digital**. São Paulo: McGraw-Hill, 1982.

[TRI 94] TRINDADE, A.L. **Estudo de Técnicas de Processamento de Sinais para a Geração de Gráficos Espectrais de Sinais Digitais**. Porto Alegre: CPGCC da UFRGS, 1994. 55p.

[VAL 99] VALIATI, J.F. **Estudo de Técnicas de Processamento de Sinais de Voz**. Porto Alegre: CPGCC da UFRGS, 1999. 46p.

[VIZ 99] VIZZOTO, G.M.; PETRY, A.; BARONE, D.A.C. Utilização de um Algoritmo de Alinhamento Temporal Dinâmico para o Reconhecimento de Voz. In: CONFERÊNCIA CIENTÍFICA DA UFRGS, 2., 1999. **Rumos da Pesquisa**: multíplas trajetórias. Porto Alegre: UFRGS, PROPESQ, 1999.

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

Informática  **UFRGS**

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

"Reconhecimento de Voz para Comandos de Direcionamento por Meio de Redes Neurais"

por

João Francisco Valiati

Dissertação apresentada aos Senhores:

Prof. Dr. Dante Augusto Couto Barone

Prof. Dr. Luigi Carro

Prof. Dr. Fernando Santos Osório (UNISINOS)

Vista e permitida a impressão.
Porto Alegre, 09/01/2004.

Prof. Dr. Paulo Martins Engel,
Orientador.

Prof. Carlos Alberto Heuser
Coordenador do Programa de Pós-Graduação em Computação - PPGC
Instituto de Informática - UFRGS