

UNIVERSIDADE ESTADUAL DE CAMPINAS  
FACULDADE DE ENGENHARIA ELÉTRICA  
DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E AUTOMAÇÃO INDUSTRIAL

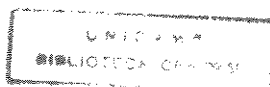
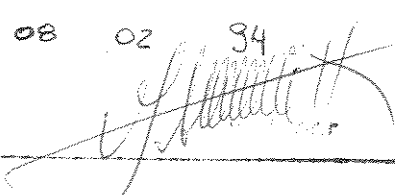
# Reconhecimento de Fonemas da Língua Portuguesa pelo Uso de Redes Neurais do Tipo “Perceptron” Multi-Camadas

por Eng. Luiz Eduardo Roncato Cordeiro<sup>zt</sup>  
orientador Prof. Dr. Márcio Luiz de Andrade Netto<sup>t</sup>

Dissertação submetida à Faculdade de Engenharia Elétrica da Universidade Estadual de Campinas, como parte dos requisitos para obtenção do Título de Mestre em Engenharia Elétrica.

21 de março de 1994

Este exemplar contém a versão final da tese  
defendida por LUIZ EDUARDO RONCATO CORDEIRO  
à Comissão de Avaliação da Comissão  
Julgadora em 08 02 94



*Este trabalho contou com o apoio financeiro do CNPq.*

*No dia 8 de fevereiro de 1994 às 14:00 horas ocorreu a defesa deste trabalho cuja banca contou com a presença dos seguintes professores:*

*Prof. Márcio Luiz de Andrade Netto,  
da Universidade Estadual de Campinas;*

*Prof. Fernando Antônio Campos Gomide,  
da Universidade Estadual de Campinas;*

*Prof. Gilberto Arantes Carrijo,  
da Universidade Federal de Uberlândia.*

# Sumário

<b>Lista de Figuras</b>	<b>iv</b>
<b>Lista de Tabelas</b>	<b>vi</b>
<b>Resumo</b>	<b>vii</b>
<b>Abstract</b>	<b>viii</b>
<b>Agradecimentos</b>	<b>x</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Modelo Clássico de Reconhecimento de Voz . . . . .	2
1.2 Histórico das Redes Neurais . . . . .	4
1.3 Modelo Proposto para o Reconhecimento de Fonemas . . . . .	5
<b>2 Aspectos Fisiológicos da Audição</b>	<b>8</b>
2.1 Localização dos Componentes do Sistema Auditivo . . . . .	8
2.2 Ouvido Externo . . . . .	10
2.3 Ouvido Médio . . . . .	10
2.3.1 Transmissão da Membrana Timpânica à Cóclea . . . . .	11
2.3.2 Transmissão de Sons através do Crânio . . . . .	13
2.4 Ouvido Interno—Cóclea . . . . .	13
2.4.1 Anatomia Funcional . . . . .	13
2.4.2 Transmissão das Ondas Sonoras na Cóclea . . . . .	15
2.4.3 Função do Órgão de Corti . . . . .	17
2.4.4 Reconhecimento do Tom . . . . .	20
<b>3 Modelo Computacional da Cóclea</b>	<b>23</b>
3.1 Implementação Cascata-Paralelo . . . . .	25
3.1.1 Parâmetros da Implementação Cascata-Paralelo . . . . .	27
3.1.2 Construção dos Filtros . . . . .	30
3.2 Implementação Cascata . . . . .	32
3.2.1 Estágio de Pré-Ênfase . . . . .	34
3.2.2 Filtros dos Estágios em Cascata . . . . .	38

3.3	Detecção na Cóclea . . . . .	39
3.4	Compressão e Adaptação na Cóclea . . . . .	40
3.4.1	Parâmetros e Implementação dos CAGs . . . . .	41
3.5	Adições ao Modelo . . . . .	44
<b>4</b>	<b>Redes Neurais e Reconhecimento de Padrões</b>	<b>47</b>
4.1	Rede “Perceptron” Simples . . . . .	49
4.2	Simplificação do Modelo do Neurônio Artificial . . . . .	51
4.3	Redes Neurais “Perceptron” Multicamadas . . . . .	52
4.4	Algoritmos de Aprendizagem . . . . .	55
4.4.1	“Backpropagation” . . . . .	55
4.4.2	Mínimos Quadrados Recursivo . . . . .	58
4.5	Exemplos de Redes . . . . .	64
<b>5</b>	<b>Implementação do Classificador de Fonemas</b>	<b>71</b>
5.1	Extração dos Padrões de Voz . . . . .	71
5.1.1	Comentários sobre Notação utilizada para os Fonemas . . . . .	72
5.1.2	Características dos Fonemas da Língua Portuguesa . . . . .	73
5.1.3	Amostragem dos Padrões no Tempo . . . . .	80
5.1.4	Obtenção dos Padrões para as Redes Neurais . . . . .	82
5.2	Especificação das Redes Neurais . . . . .	87
5.2.1	Redes para os Padrões da FFT . . . . .	88
5.2.2	Redes para os Padrões da Cóclea . . . . .	89
5.3	Comentários sobre a Implementação . . . . .	89
<b>6</b>	<b>Resultados Experimentais</b>	<b>92</b>
6.1	Padrões Correspondentes às Vogais . . . . .	93
6.2	Padrões Correspondentes aos Fonemas Explosivos . . . . .	95
6.3	Padrões Correspondentes aos Fonemas Fricativos . . . . .	97
6.4	Testes com Outros Locutores . . . . .	99
6.5	Testes com Palavras Inteiras . . . . .	101
6.6	Comentários . . . . .	103
<b>7</b>	<b>Conclusões</b>	<b>104</b>
	<b>Bibliografia</b>	<b>107</b>
<b>A</b>	<b>Análise LPC</b>	<b>111</b>
<b>B</b>	<b>Filtragem Homomórfica de Sinais</b>	<b>114</b>
<b>C</b>	<b>Transformada Discreta de Hartley</b>	<b>116</b>

# Lista de Figuras

1.1	Modelo Clássico de Reconhecimento de Voz . . . . .	2
1.2	Modelo de Reconhecimento de Voz Proposto . . . . .	6
2.1	Ouvido Humano . . . . .	9
2.2	Ouvido Médio e Interno . . . . .	11
2.3	Corte Transversal da Cóclea . . . . .	14
2.4	Movimento de Líquidos na Cóclea . . . . .	14
2.5	Padrões de Vibração na Cóclea . . . . .	16
2.6	Amplitudes Máximas de Vibração para Várias Frequências . . . . .	17
2.7	Órgão de Corti . . . . .	18
2.8	Posição das Células Ciliadas na Lâmina Reticular e suas Ineruações . . . . .	18
2.9	Estimulação das Células Ciliadas . . . . .	19
3.1	Esquematisação do Modelo da Cóclea . . . . .	25
3.2	Implementação Cascata-Paralelo . . . . .	26
3.3	Pólos e Zeros da Implementação Cascata-Paralelo . . . . .	26
3.4	Resposta do Filtro $AR_{10}$ . . . . .	30
3.5	Resposta da Cascata até o Filtro $AR_{10}$ . . . . .	31
3.6	Resposta do Filtro $R_{10}$ . . . . .	31
3.7	Resposta do Canal 10 . . . . .	32
3.8	Pólos e Zeros da Implementação Cascata . . . . .	33
3.9	Implementação Cascata . . . . .	34
3.10	Resposta dos Ouvidos Externo e Médio . . . . .	35
3.11	Resposta do Filtro Compensador . . . . .	36
3.12	Resposta Conjunta do Compensador e Ouvidos Externo e Médio . . . . .	37
3.13	Resposta do Filtro de Pré-Ênfase . . . . .	37
3.14	Resposta Isolada do Canal 25 - Modelo Cascata . . . . .	39
3.15	Resposta de 4 Canais do Modelo da Cóclea . . . . .	40
3.16	Esquema do CAG . . . . .	42
3.17	Função Característica do CAG ( $a = 1$ ) . . . . .	43
3.18	Função Característica do CAG ( $a = 0,1$ ) . . . . .	43
3.19	Banco de CAGs . . . . .	45
3.20	Diagrama Completo de um Canal da Cóclea . . . . .	46

4.1	Simbologia de um Neurônio Artificial . . . . .	48
4.2	Funções de Ativação do Neurônio . . . . .	49
4.3	Regiões de Decisão para um Neurônio de Duas Entradas . . . . .	50
4.4	Rede Neural Multicamadas . . . . .	53
4.5	Tipos de Regiões de Decisão para Redes Multicamadas . . . . .	54
4.6	Rede Emuladora da Função XOR . . . . .	66
4.7	Mapa de Saída da Rede XOR (Calculada) . . . . .	66
4.8	Mapa de Saída da Rede XOR (“Backpropagation”) . . . . .	68
4.9	Mapa de Saída da Rede XOR (MQR, Filtro de Kalman) . . . . .	69
5.1	Esquematização Acústica do Trato Vocal . . . . .	72
5.2	Sinal no Tempo da Vogal /á/ . . . . .	74
5.3	Sinal no Tempo para a Vogal /é/ . . . . .	74
5.4	Sinal no Tempo para a Vogal /i/ . . . . .	75
5.5	Sinal no Tempo para a Vogal /ó/ . . . . .	75
5.6	Sinal no Tempo para a Vogal /u/ . . . . .	76
5.7	Sinal no Tempo para a Sílabas /sá/ . . . . .	77
5.8	Sinal no Tempo para a Sílabas /fá/ . . . . .	77
5.9	Sinal no Tempo para a Sílabas /pá/ . . . . .	78
5.10	Sinal no Tempo para a Sílabas /tá/ . . . . .	78
5.11	Sinal no Tempo para a Sílabas /vá/ . . . . .	79
5.12	Sinal no Tempo para a Sílabas /bá/ . . . . .	80
5.13	Extração de Padrões por Espectro Segmentado . . . . .	83
5.14	Padrão Neural para a Vogal /á/, FFT . . . . .	84
5.15	Extração de Padrões pelo Modelo da Cóclea . . . . .	85
5.16	Padrão Neural para a Vogal /á/, Cóclea . . . . .	86

# Lista de Tabelas

3.1	Parâmetros dos Bancos CAG . . . . .	44
4.1	Algoritmo de Aprendizagem, “Backpropagation” . . . . .	59
4.2	Algoritmo de Aprendizagem, Mínimos Quadrados Recursivo . . . . .	65
4.3	Tabela Verdade para a Rede <i>XOR</i> . . . . .	65
5.1	Configurações das Redes para FFT . . . . .	88
5.2	Configurações das Redes para Cóclea . . . . .	89
6.1	Acertos dos Testes—Vogais/FFT . . . . .	94
6.2	Acertos dos Testes—Vogais/Cóclea . . . . .	94
6.3	Erros nos Testes—Vogais/FFT . . . . .	94
6.4	Erros nos Testes—Vogais/Cóclea . . . . .	95
6.5	Acerto dos Testes—Explosivos/FFT . . . . .	96
6.6	Acertos dos Testes—Explosivos/Cóclea . . . . .	96
6.7	Erros nos Testes—Explosivos/FFT . . . . .	96
6.8	Erros nos Testes—Explosivos/Cóclea . . . . .	97
6.9	Acertos dos Testes—Fricativos/FFT . . . . .	98
6.10	Acertos dos Testes—Fricativos/Cóclea . . . . .	98
6.11	Erros nos Testes—Fricativos/FFT . . . . .	98
6.12	Erros dos Testes—Fricativos/Cóclea . . . . .	98
6.13	Acertos dos Testes—Vogais/FFT—Locutor A . . . . .	100
6.14	Acertos dos Testes—Vogais/FFT—Locutor B . . . . .	100
6.15	Acertos dos Testes—Vogais/Cóclea—Locutor A . . . . .	100
6.16	Acertos dos Testes—Vogais/Cóclea—Locutor B . . . . .	100
6.17	Acertos dos Testes—Números/Cóclea . . . . .	102

# Resumo

Neste trabalho propõe-se a construção de um sistema de reconhecimento de fonemas da língua portuguesa por intermédio de redes neurais do tipo “perceptron” multicamadas. Este sistema é constituído por um modelo matemático do ouvido humano e por um modelo de redes neurais. O modelo do ouvido humano, neste caso, efetua um pré-processamento no sinal sonoro, muito parecido com a transformada de Fourier, gerando os dados para excitação dos neurônios de entrada das redes neurais. Tais redes são construídas de modo que, dados os sinais provenientes do modelo do ouvido, possam indicar o fonema correspondente ao sinal sonoro de entrada. Deve-se salientar que foi realizada neste trabalho uma comparação entre o modelo do ouvido e a transformada rápida de Fourier como pré-processadores do sinal de áudio, exatamente para ilustrar as semelhanças entre ambos.



# Abstract

The purpose of this work is the construction of a recognition system for the portuguese idiom phonemes by the use of multi layered perceptron neural networks. This system is constituted by a mathematical model of the human ear and of a model of neural networks. The human ear model does a pre-processing of the sound signal, very similar to Fourier transform, generating data for excitation of neural network input neurons. The network is such that, given the signals from the ear's model, it may classify the corresponding phonem from the input sound signal. It was done also a comparison between the ear's model and the Fourier transform as a pre-processor to sound signal, to illustrate the similarities between both processes.

Dedico este trabalho aos meus pais,  
Orlando e Maria José  
e à minha namorada  
Vanderlene  
pelo apoio dado em todas as horas.

# Agradecimentos

Quero agradecer, em primeiro lugar, ao meu orientador Prof. Márcio Luiz de Andrade Netto pelo incentivo, apoio e amizade durante o curso de mestrado e pelos conselhos nos momentos difíceis.

Gostaria de agradecer, em especial, a duas pessoas, que desde a minha graduação têm-se mostrado grandes amigos, Waldemar Scudeller Jr. e Mauro Jorge Atalla, que até este momento está se divertindo em um doutorado nos Estados Unidos.

Meus agradecimentos ao pessoal do SIFEE, Prof. Akebo Yamakami, Gorgonio B. Araújo e Luiz Antonio Sales Monteiro, que durante um ano possibilitaram a mim um ambiente de trabalho diferente e divertido, mostrando-me as gratificações e azares da vida de um administrador de sistema.

Agradeço ao pessoal do DPM, em especial ao Prof. Loir Afonso Moreira, que me cedeu algumas horas preciosas nas estações de trabalho para o desenvolvimento da minha tese. E aos amigos que fiz por lá, Prof. Marco Lúcio Bittencourt, Prof. Ilmar Ferreira Santos, Prof. Humberto Camargo Picolli e ao Prof. Vicente Lopes Jr.

Não poderia deixar de lado um grande amigo, Prof. Janito Vaqueiro Ferreira, que, apesar de ser um engenheiro mecânico, mostrou ser um perito em circuitos elétricos digitais. Pena que este se encontra na terra da Revolução Industrial, Inglaterra, investindo na finalização de seu doutorado.

Por último, gostaria de mostrar minha gratidão a todos os amigos que fiz na Unicamp. Em especial aos amigos da RPV.

*The need for learning arises whenever available a priori information is incomplete. The type of learning depends on the degree of completeness of this a priori information. In learning with supervision, it is assumed that at each instant of time we know in advance the desired response of the learning system, and we use the difference between the desired and actual response, that is, the error of the learning system, to correct its behaviour. In learning without supervision, we do not know the desired response of the learning system.*

**Ya. Z. Tsypkin**

Foundations of the Theory of Learning Systems

# Capítulo 1

## Introdução

Um dos grandes desafios desta época é a construção de máquinas capazes de compreender a voz humana. Atualmente, os grandes avanços conseguidos com as pesquisas concentram-se em máquinas capazes de reproduzir com razoável perfeição a voz humana. As máquinas capazes de reconhecê-la, no entanto, tiveram um avanço tecnológico modesto.

A primeira máquina capaz de reconhecer com um certo sucesso a pronúncia de determinadas palavras data de 1952 [2]. Tratava-se de um sistema capaz de distinguir os dez dígitos da língua inglesa falados ao telefone. Esta máquina atingiu percentuais de acerto de quase 100% quando calibrada e usada por uma mesma pessoa, no entanto o índice de acerto caía para até 50% quando outra pessoa a utilizava.

Muitos trabalhos se sucederam na década de 60 [10], a nível de laboratório, em reconhecimento de palavras, sílabas, letras e fonemas isolados. Todos baseados nas descobertas de algumas das propriedades da voz através de *espectrógrafos* [12] e pelas facilidades introduzidas pelos computadores digitais.

Duas linhas de atividades foram então iniciadas. De um lado, sistemas que procuravam distinguir um conjunto grande de palavras, geralmente de 30 a 50 palavras, pronunciadas por apenas uma pessoa e, de outro lado, tentativas de reconhecimento de poucas palavras, como dígitos, para um grupo de 5 a 25 pessoas. Esta divisão ocorreu devido à dificuldade em se conseguir criar uma máquina capaz de reconhecer um vocabulário genérico e vasto de um grupo grande de pessoas (sistemas independentes do locutor ou multilocutor) [34].

Outra dificuldade, encontrada ainda hoje em reconhecimento de voz, é conhecida como *coarticulação* que, numa fala natural, corresponde à aglutinação que ocorre entre o fim

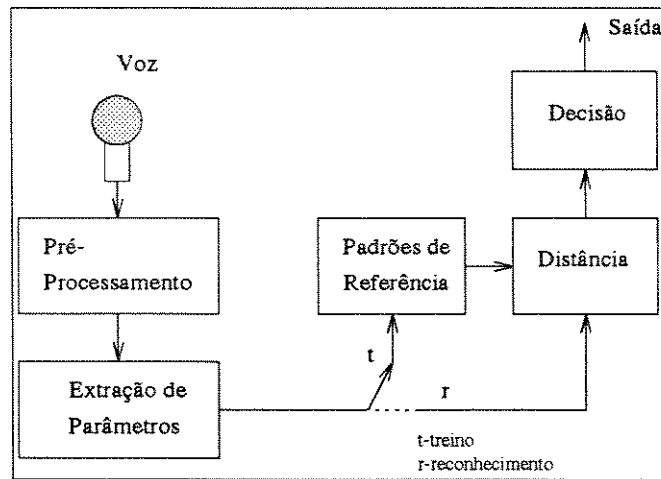


Figura 1.1: Diagrama de um modelo clássico de um sistema capaz de reconhecer voz.

de uma palavra e o início da palavra seguinte formando padrões acústicos bem diferentes dos padrões das palavras pronunciadas isoladamente, por exemplo, como acontece com o número 32 que é normalmente pronunciado como /trinteidois/ transformando o /a/ no final da palavra /trinta/ e o /e/ no ditongo /ei/. Esta dificuldade causada por coarticulação dividiu os sistemas de voz no que diz respeito à pronúncia da palavra. Existem sistemas de reconhecimento de *palavras isoladas* que exigem um pequeno intervalo entre as palavras; sistemas para *fala natural* que permitem a pronúncia como em uma conversa normal e sistemas de *fala conectada* onde não há a necessidade de pausas, porém é exigida uma pronúncia clara das palavras.

Deve-se salientar neste ponto que o reconhecimento da fala natural é extremamente complicado e exige grandes esforços e conhecimentos na área de *lingüística*.

## 1.1 Modelo Clássico de Reconhecimento de Voz

A estrutura de um *sistema clássico* de reconhecimento de voz pode ser dada pela figura 1.1, consistindo de 5 etapas fundamentais.

Numa primeira etapa de *pré-processamento*, determinadas componentes da voz são filtradas, compensadas ou extraídas em computadores digitais onde é realizada a amostragem do sinal. O sinal passa então por uma segunda etapa, a de *extração de parâmetros* que na maioria das vezes consiste na segmentação do sinal através de um janelamento temporal. Este janelamento geralmente consiste na retirada de uma porção de 10 a 30ms do sinal so-

bre o qual se efetua uma análise a *curto prazo*, uma vez que o movimento articulatorio dos lábios, dentes e língua podem ser considerados quase estacionários neste intervalo. A esta porção retirada do sinal de voz é aplicado algum tipo de processamento que deverá fornecer os parâmetros que formarão os padrões a serem utilizados pelo sistema de reconhecimento. Dentre os padrões extraídos do sinal de voz o mais comum é o espectro de frequência obtido através da *transformada rápida de Fourier*. Ainda no domínio da frequência, pode-se tentar isolar a resposta em frequência do trato vocal<sup>1</sup> através de técnicas como a filtragem homomórfica de sinais [23].

Outros tipos de análises a curto prazo podem ser realizadas no sinal de voz. Diretamente do segmento extraído do sinal de voz podem ser determinados por exemplo: a energia contida no sinal de voz do segmento; a contagem de picos positivos e/ou negativos do sinal e o número de cruzamentos por zero. Ainda no domínio do tempo pode ser realizada uma análise com o intuito de se obter a função de transferência correspondente ao trato vocal através da predição linear (LPC).

Tanto a filtragem homomórfica de sinais quanto a metodologia da predição linear são detalhadas exaustivamente nas bibliografias sobre processamento de sinais (ou voz) tais como [23] ou [25]. Entretanto, resumos sobre a análise LPC e sobre a filtragem homomórfica podem ser encontrados nos apêndices A e B, respectivamente.

Deve-se salientar, neste ponto, que a maioria dos parâmetros extraídos do sinal de voz são obtidos através da modelagem do sistema fonador, o qual produz os sons usados na comunicação humana.

As etapas seguintes são caracterizadas pelas técnicas básicas de *reconhecimento de padrões*. Uma vez extraídas as características, ou parâmetros, dos sinais é formado um conjunto de *padrões de referência* durante a fase de *treinamento* que são utilizados como base de conhecimento sobre as palavras ou sons *ensinados*.

Durante a fase de *reconhecimento*, os parâmetros de um sinal de voz desconhecido até então são comparados com os parâmetros de cada *padrão de referência* e, a partir desta comparação, são realizados cálculos das distâncias entre o novo sinal e os padrões de referência. A *decisão* é tomada, em uma última etapa, escolhendo-se o padrão de referência que mais se aproximou do novo padrão.

Como pôde ser observado, a decisão no modelo clássico é obtida através de grande

---

<sup>1</sup>Conjunto de órgãos e cavidades capazes de “modular” os pulsos emitidos pelas cordas vocais para produzir os fonemas e sons.

esforço computacional, uma vez que uma medida de distância é calculada entre o sinal sendo testado e cada um dos padrões armazenados como referência. Este tipo de procedimento requer ainda um elevado grau de paralelismo computacional, no caso de se desejar sistemas rápidos. Além disto, estes modelos requerem grande quantidade de memória para armazenar os padrões de referência.

Com o surgimento das *redes neurais* ou *sistemas conexionistas* conseguiu-se minimizar parte destes problemas. A rede neural é formada por unidades computacionais simples, altamente interconectadas e com a característica de um ambiente altamente paralelo, uma vez que o processamento nestas unidades é independente. O *conhecimento* nas redes neurais é representado apenas pelos pesos dados às *conexões* entre as unidades e pela natureza do funcionamento das mesmas. Estes modelos de redes neurais tiveram sua origem nos estudos dos *sistemas nervosos biológicos* e nos estudos dos modelos elétricos e matemáticos dos *neurônios*.

## 1.2 Histórico das Redes Neurais

Na década de 40 dois pesquisadores da área biológica (McCulloch & Pitts [20]) propuseram o primeiro modelo matemático do funcionamento de um neurônio. Este modelo, apesar de simples, trouxe uma grande contribuição para as discussões sobre a natureza da *inteligência humana*, estimulando especulações sobre a estrutura de um cérebro e permitindo a criação dos primeiros modelos matemáticos de dispositivos artificiais que buscavam analogias biológicas.

Uma grande atividade ocorreu nesta área emergente provocando um grande incentivo por parte de órgãos financiadores, pois as promessas eram no mínimo fantásticas. Este apoio durou até meados da década de 60 quando as instituições financiadoras retiraram o apoio devido principalmente ao não cumprimento das metas propostas nos projetos. A publicação do livro *Perceptrons* de Minsk e Papert [21] onde prova-se que as estruturas utilizadas nos dispositivos da época não eram capazes de aprender regras lógicas tão simples quanto a do *ou-exclusivo*, selou as atividades relacionadas às redes neurais.

Pouquíssimos pesquisadores continuaram as pesquisas sobre o assunto, entre eles destacam-se Teuvo Kohonen (Finlândia), Edoardo Caianiello (Itália), Stephen Grossberg, Bernard Widrow e James Anderson (E.U.A.) e Kunihiko Fukushima (Japão).

O renascimento do interesse científico sobre este assunto teve início através dos



físicos com a aplicação dos conceitos conexionistas ao problema de modelamento de materiais para-magnéticos como o *vidro de spin*, realizado por Hopfield [8] e publicado em 1982. A idéia básica era da analogia entre o *spin magnético* de um átomo influenciando a todos os demais existentes no vidro e, evidentemente, sofrendo a influência de todos na determinação de sua própria orientação magnética. Esta analogia gerou um modelo dinâmico não linear que se provou estável desde que a matriz de pesos que medem a força de interação entre cada par de *spins* fosse simétrica. Esta condição é naturalmente verificada no problema físico não causando, portanto, dificuldades para a aplicação.

A partir daí, diversas possibilidades de aplicação foram exploradas, como, por exemplo, em otimização e em reconhecimento de padrões, com resultados bastante encorajadores.

Em 1986 ocorreu o fato que efetivamente colocou a área de *Redes Neurais* como um das prioritárias na obtenção de recursos. Este fato foi a explicitação por Rumelhart, Hinton e Williams [26] de um algoritmo de aprendizado (*backpropagation*) para as redes do tipo Perceptron com estrutura multi-camadas, fazendo cair definitivamente a alegação de que estes tipos de redes resolviam apenas problemas triviais. Cabe aqui observar que Minsky e Papert estavam absolutamente corretos no que afirmaram, pois os algoritmos de aprendizagem até então eram capazes de aprender apenas a solução para problemas simples de classificação. O algoritmo de *backpropagation* foi desenvolvido anteriormente por outros pesquisadores (Werbos em 1974, Parker em 1975 e Le Cun em 1975) que o estavam aplicando em problemas não relacionados diretamente com as redes neurais.

### 1.3 Modelo Proposto para o Reconhecimento de Fonemas

Este trabalho tem como objetivo substituir o modelo clássico de reconhecimento baseado nas medidas de *distâncias* com relação aos *padrões de referência* por *redes neurais "perceptron" multi-camadas* de acordo com a estrutura apresentada na figura 1.2.

Em geral os sistemas clássicos de análise e reconhecimento de voz [25, 34] baseiam-se em modelos do sistema fonador,<sup>2</sup> para realizar a extração dos parâmetros do sinal de voz, raramente são utilizados, nos trabalhos relacionados à área, modelos do ouvido humano para esta função. De fato, a principal contribuição deste trabalho deve-se à utilização de um modelo computacional do ouvido humano na extração dos parâmetros do sinal de voz.

---

<sup>2</sup>Responsável pela produção dos sons dos fonemas.

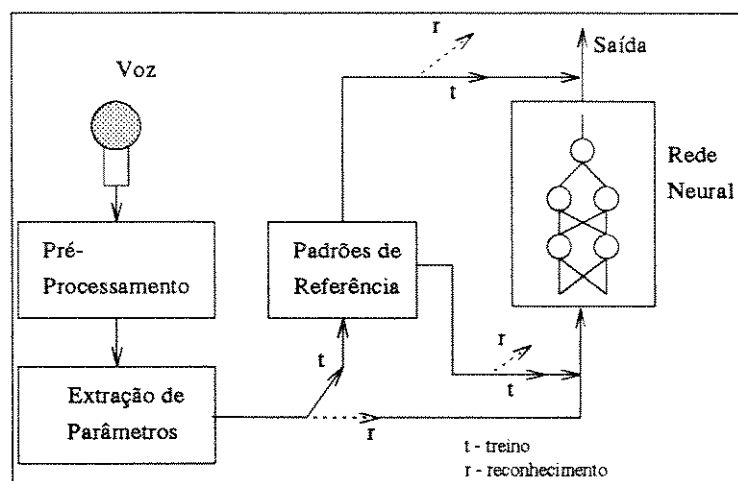


Figura 1.2: Diagrama do modelo proposto de um sistema capaz de reconhecer fonemas.

Para tanto foi escolhido, pela simplicidade de implementação, o *modelo dinâmico do ouvido* apresentado por Richard F. Lyon em 1982 [16] composto por uma cascata de filtros capazes de isolar as componentes de frequência do sinal de voz e de modelar algumas das características encontradas no ouvido humano. Este modelo foi usado para extrair as características dos sinais sonoros e para gerar os padrões usados no treinamento e teste das redes neurais. Foi usado, a título de comparação de desempenho, o espectro em frequência obtido pela *transformada rápida de Fourier* de segmentos de tempo extraídos do sinal sonoro digitalizado. Estes espectros formaram um segundo grupo de padrões para o ensino e testes das redes neurais.

Não é objetivo deste trabalho a criação de um sistema capaz de reconhecer fonemas em *tempo real*, tampouco de sistemas multilocutores, e sim de definir bases para a possível construção de sistemas mais complexos, inclusive com capacidades de operação a tempo-real ou multilocutores.

Após esta rápida introdução sobre os sistemas de reconhecimento de voz e sobre as redes neurais será feita uma descrição da organização do texto desta dissertação.

Antes de se passar para o modelo computacional do ouvido humano, o capítulo 2 entrará em detalhes a respeito da fisiologia e funcionamento das várias partes do ouvido humano. Esta introdução servirá para facilitar o entendimento da construção do modelo do ouvido interno, desenvolvido por Richard F. Lyon [16]. Este modelo será abordado no capítulo 3.

De uma maneira generalizada, o capítulo 4 discutirá o uso das redes neurais “perceptron” multi-camadas em reconhecimento de padrões.

No capítulo 5 serão apresentadas as informações a respeito da implementação das redes, das metodologias de extração dos parâmetros da voz e da amostragem dos sons. O capítulo 6 entra em detalhes a respeito dos resultados obtidos pelas redes neurais com os padrões extraídos via *transformada rápida de Fourier* e via *modelo da cóclea*.

E finalmente, no capítulo 7 é feita a avaliação dos resultados e das perspectivas de desenvolvimentos futuros a partir deste trabalho.

## Capítulo 2

# Aspectos Fisiológicos da Audição

Este capítulo destina-se ao estudo da fisiologia do sistema auditivo, voltando a análise para como o ouvido processa os sinais sonoros provenientes do meio ambiente e os transforma em impulsos nervosos para, então, enviá-los ao cérebro.

A audição, como a maioria dos sentidos, é um sentido composto por receptores mecânicos, pois o ouvido responde às vibrações mecânicas das ondas sonoras no ar. Será mostrado neste capítulo o modo pelo qual o ouvido capta as ondas, discrimina as frequências e transmite as informações ao sistema nervoso central.

### 2.1 Localização dos Componentes do Sistema Auditivo

O ouvido humano costuma ser subdividido em três regiões distintas denominadas, *ouvido externo*, *ouvido médio* e *ouvido interno*, tendo cada uma delas funções específicas [5, 22]. A última subdivisão é um órgão específico que será detalhado na seção 2.4. Este órgão realiza todo o trabalho de conversão das vibrações sonoras em impulsos nervosos e é chamado *cóclea*. A figura 2.1 mostra a localização de cada umas das partes do ouvido.

O ouvido externo compreende o *pavilhão auditivo*<sup>1</sup> e o *duto auditivo*, que termina na *membrana timpânica*, ou *tímpano*. O ouvido médio é constituído pelo tímpano e pelo *sistema ossicular* formado por três ossículos denominados *bigorna*, *martelo* e *estribo*.<sup>2</sup>

O ouvido interno, *cóclea*, é quem cuida da transdução das vibrações sonoras em impulsos nervosos, correspondendo na figura 2.1 ao órgão em forma de caracol. Os anéis

---

<sup>1</sup>Chama-se de orelha somente o pavilhão auditivo e de ouvido todo o sistema auditivo.

<sup>2</sup>Possuem esses nomes por se assemelharem aos respectivos objetos.

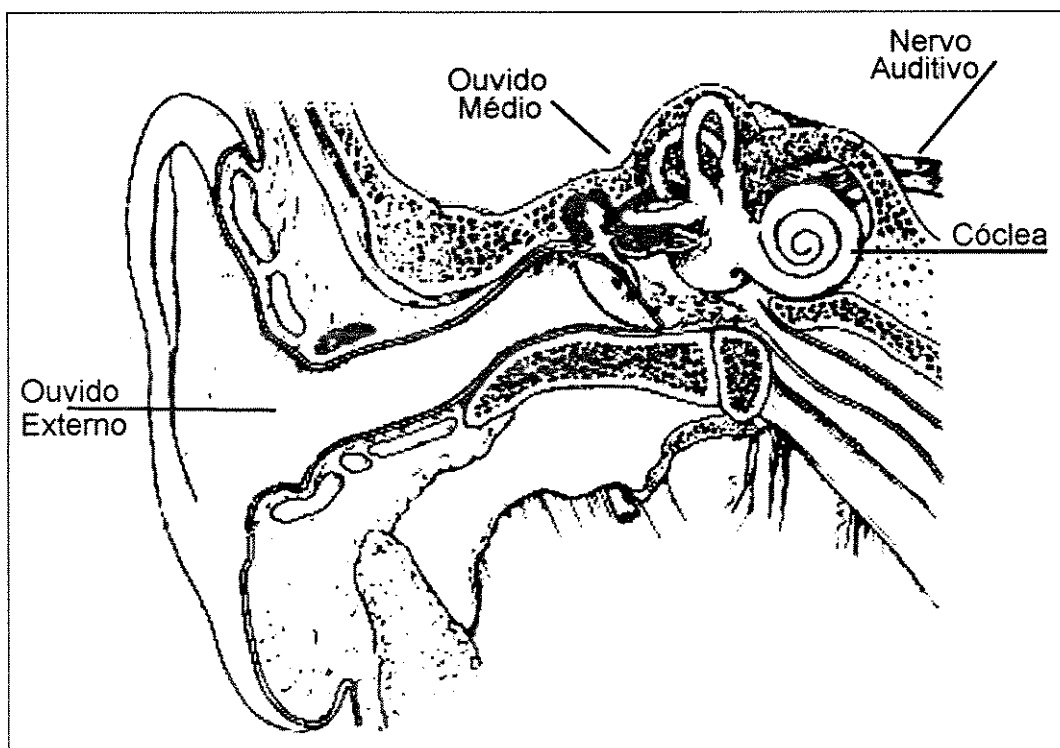


Figura 2.1: Subdivisões do ouvido humano [29].

circulares ligados à cóclea não fazem parte do sistema auditivo e correspondem ao órgão de equilíbrio, às vezes chamado de *labirinto*,<sup>3</sup> enviando sinais sobre a posição da cabeça ao cerebelo. Sabe-se no entanto que estes órgãos possuem certa sensibilidade às vibrações sonoras de frequências muito baixas, mas no entanto as sensações produzidas no cérebro somente causam perda de equilíbrio no caso de sons de amplitude elevada [22].

A seguir será explicado o funcionamento de cada uma das partes do ouvido e suas influências nos sinais sonoros.

## 2.2 Ouvido Externo

O ouvido externo tem como função a captação dos sons através do pavilhão auditivo sendo que o duto auditivo somente isola a membrana timpânica de choques mecânicos que poderiam causar danos permanentes (perfurações), pois é extremamente delgada e frágil. A forma do duto auditivo insere no sistema uma ressonância em torno de 3 000Hz [5, 22]. No entanto o grau de ressonância é pequeno e não evidencia muito o som desta frequência [5]. Esta parte do ouvido poderá ser vista com maiores detalhes na figura 2.1.

## 2.3 Ouvido Médio

A figura 2.2 mostra com maior detalhe a região onde estão os ouvidos médio e interno. O ouvido médio tem como função principal a transmissão das vibrações sonoras de um meio gasoso (ar) para um meio líquido (interior da cóclea), de uma maneira eficiente. Esta seção do ouvido faz o casamento das impedâncias destes meios. Através de pequenos músculos, o ouvido médio é capaz, ainda, de proporcionar um sistema de segurança contra sons de intensidade elevada.

O ouvido médio é conectado à faringe através da trompa de Eustáquio e esta conexão garante a equalização das pressões em ambos os lados do tímpano. Sem essa equalização, a transmissão do som seria atenuada. Esta conexão encontra-se normalmente fechada, abrindo-se durante a deglutição, a mastigação,<sup>4</sup> o bocejo e o espirro [22].

<sup>3</sup>Doenças nesses órgãos provocam perda de equilíbrio e mal-estar chamada de Labirintite. Semelhante, talvez, ao enjôo causado pelas oscilações dos navios em alto mar.

<sup>4</sup>Nos tempos pioneiros da aviação e nos casos de aviões militares, eram oferecidas gomas de mascar aos passageiros para evitar os estalos característicos e a má sensação provocada pela súbita mudança na pressão ambiente.

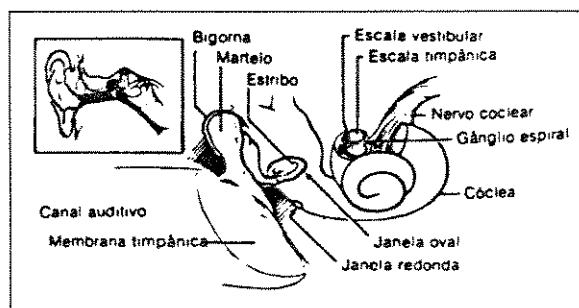


Figura 2.2: A membrana timpânica, a cadeia de ossículos do ouvido médio e o ouvido interno [5].

### 2.3.1 Transmissão da Membrana Timpânica à Cóclea

A figura 2.2 mostra a *membrana timpânica* e a *cadeia de ossículos*, que transmite o som através do ouvido médio. A membrana do tímpano tem forma cônica e unido ao seu centro está o *cabo do martelo*. O martelo acha-se firmemente unido, em sua outra extremidade à *bigorna*; assim, sempre que o martelo se move a bigorna se move em sincronia. O extremo oposto da bigorna articula-se com a cabeça do *estribo* e a base do mesmo se apóia na abertura da janela oval, onde as ondas sonoras são transmitidas ao ouvido interno [5, 22].

O movimento dos ossículos do ouvido médio fazem com que as vibrações da membrana timpânica sejam transmitidas até a base da cóclea pelo estribo, através do deslocamento para dentro e para fora da cóclea ao nível da janela oval [5].

O cabo do martelo é puxado constantemente para dentro do ouvido médio por ligamentos e pelo músculo tensor do tímpano que mantém a membrana timpânica constantemente tensa. Isto garante a transmissão de qualquer vibração que ocorra na membrana [22]. As análises da transmissão tímpano-cóclea são geralmente feitas com frequências abaixo de 3 000Hz, pois nessas frequências os ossículos se movem como um corpo rígido e o estribo adquire somente um movimento de vai-vem na cóclea. Nessas frequências os modos de vibração da cóclea são mais simples, facilitando a análise da transmissão que não leva em conta, por exemplo, as propriedades mecânicas da caixa do ouvido médio e das janelas [22].

Esse sistema mecânico introduz no ouvido médio duas ressonâncias, uma principal em torno de 1 200Hz que é aparentemente causada por elementos da membrana timpânica e da junta bigorna-estribo, e uma secundária em 800Hz aparentemente relacionada a elementos existentes além da articulação bigorna-estribo.

Combinando os efeitos da ressonância dos ouvidos externo e médio, a transmissão das ondas sonoras do ar para a cóclea é mais eficiente para as ondas de 600 a 6 000Hz, diminuindo acima e abaixo desta faixa [5].

### Casamento de Impedâncias

A amplitude de movimento da base do estribo com cada vibração sonora representa apenas três quartos da amplitude do movimento do martelo, aumentando a força do movimento por um fator de 1,3 aproximadamente. Contudo a membrana timpânica é bem maior do que a superfície do estribo ( $55\text{mm}^2$  e  $3,2\text{mm}^2$ , respectivamente), o que produz uma pressão sobre o líquido da cóclea cerca de 22 vezes maior do que a pressão sonora contra o tímpano [5]. De fato, a diferença de impedâncias entre o ar e a água são tais que somente 0,1% da energia incidente é transmitida através da interface, sendo o resto refletido. Experimentalmente, a análise da perda em um ouvido de um gato foi de aproximadamente 34dB <sup>(5)</sup> sendo que a cadeia de ossículos consegue recuperar quase que completamente o sinal sonoro [22].

### Atenuação do Som—Sistema de Proteção

Outra função interessante do ouvido médio é a capacidade do mesmo de atenuar o som transmitido à cóclea através de pequenos músculos, o *músculo tensor do tímpano* e o *músculo estapédio*, este último ligado ao estribo.

Sons intensos de duração maior do que 40ms sendo transmitidos para o sistema nervoso produzem um reflexo que origina a contração dos músculos tensor (que puxa o martelo e tímpano para dentro do ouvido médio) e estapédio (que puxa o estribo para dentro, porém sem nenhum efeito na membrana timpânica). Esta contração aumenta a rigidez do sistema fazendo com que as frequências abaixo de 1 000Hz sofram uma atenuação de 30 a 40dB [5, 22]. Este reflexo é inexistente para sons com duração menores que 40ms.

A função deste mecanismo muscular parece ser dupla [5]:

- Proteger a cóclea de lesões por ruídos excessivamente intensos. Os sons capazes de provocar sérias lesões na cóclea são os de baixa frequência, no entanto o tempo de resposta desta proteção é baixo o que pode provocar danos à cóclea no caso de sons súbitos;

---

<sup>5</sup>Corresponde à redução de uma voz forte a um simples cochicho quase inaudível.



- Mascarar os sons de baixa frequência em lugares muito barulhentos. Assim se suprime grande parte do ruído ambiental.

No caso de barulhos de alta intensidade e duração curta, menores que 40ms, o modo de vibração da cadeia dos ossículos provoca uma diminuição no volume do fluido deslocado na cóclea [22]. Este mecanismo de ação rápida suplementa os reflexos do ouvido médio, permitindo uma proteção maior.

O sistema nervoso central também é capaz de ativar este sistema muscular e isso ocorre nos seguintes casos: em estado de alta concentração mental, durante o período de sono de movimentos oculares rápidos [22] e quando falamos. Quando falamos, os sinais colaterais do sistema nervoso central ativam estes músculos ao mesmo tempo em que envia sinais para o mecanismo da fala, evitando que a pessoa ouça a própria voz<sup>6</sup> [5].

### 2.3.2 Transmissão de Sons através do Crânio

Por a cóclea estar inserida em uma cavidade do osso temporal, as vibrações sonoras no crânio podem causar vibrações no líquido dentro da mesma. Por este motivo as pessoas podem “sentir” as vibrações sonoras no caso de um diapasão colocado em contato com o osso do crânio, se a intensidade sonora produzida pelo mesmo for suficiente<sup>7</sup> [5].

## 2.4 Ouvido Interno—Cóclea

### 2.4.1 Anatomia Funcional

A cóclea é um órgão composto por um tubo de 35mm de comprimento enrolado em forma de caracol com três subdivisões internas. Este tubo está inserido em uma estrutura óssea que impede deslocamentos no sentido radial e quando é feita uma incisão no mesmo o corte não se alarga, não estando, portanto, sob tensão (figura 2.3). As subdivisões internas a este tubo são denominadas de *escala vestibular*, *escala média* e *escala timpânica*. A escala vestibular e escala média são separadas pela *membrana vestibular* e as escalas média e timpânica pela *membrana basilar*. Na superfície da membrana basilar existe uma estrutura

<sup>6</sup>Talvez isso seja percebido quando ouve-se a própria voz gravada, que nunca é identificada como aquela “ouvida” pela própria pessoa.

<sup>7</sup>Infelizmente a energia disponível não é suficiente para fazer com que a pessoa possa ouvir pelo osso (mesmo para sons aéreos intensos), o que ajudaria os deficientes auditivos.

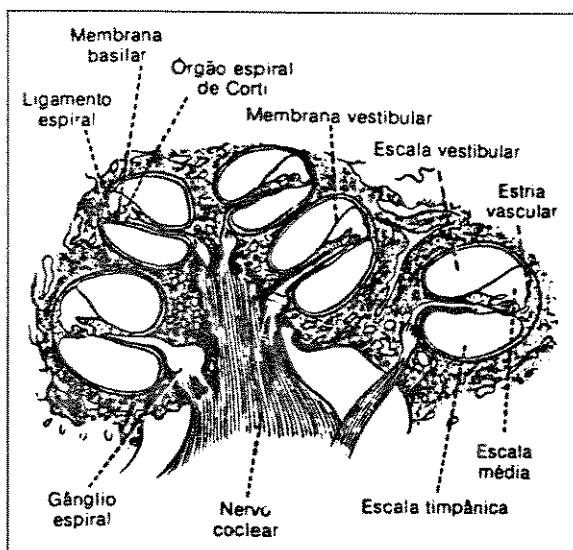


Figura 2.3: Corte transversal da cóclea [5].

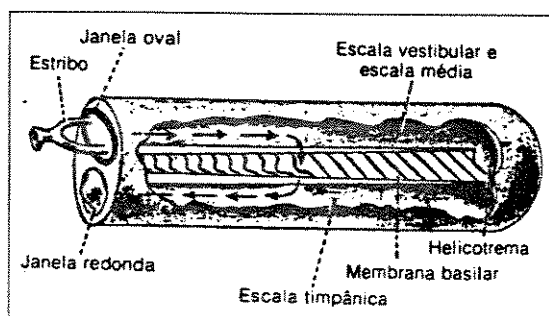


Figura 2.4: O movimento de líquido na cóclea após um deslocamento brusco do estribo para dentro da mesma [5].

denominada *órgão de Corti*, que contém uma série de células ciliadas. São esses os órgãos receptores que geram impulsos nervosos em resposta às vibrações sonoras [5, 19, 22].

Na figura 2.4 estão ilustradas esquematicamente as partes funcionais da cóclea para transmissão das vibrações sonoras. Observa-se que a membrana vestibular não está sendo representada, pois é extremamente delgada e não interfere na condução das vibrações mecânicas. Sua função é conservar um líquido especial na escala média necessário ao funcionamento das células ciliadas receptoras.

As vibrações sonoras penetram na escala vestibular pela base do estribo ao nível da janela oval, esta base está ligada à janela oval de tal modo que possibilita o movimento do estribo para dentro e para fora de acordo com as vibrações. Observa-se ainda, na figura 2.4, a

comunicação entre as escalas vestibular e timpânica por um orifício denominado *helicotrema*. Se o movimento do líquido for vagaroso, haverá tempo para o líquido da escala vestibular se deslocar para a timpânica, provocando uma protusão da janela redonda. No entanto se este movimento for rápido não haverá tempo para a acomodação do líquido, fazendo com que a onda sonora tome um “atalho” pela membrana basilar, fazendo-a oscilar, vide figura 2.4 [5]. Portanto, devido ao fato das paredes da cóclea estarem em contato com o osso e do meio líquido ser praticamente incompressível, qualquer movimento do estribo resultará em um movimento na membrana basilar [19].

### A Membrana Basilar e a Ressonância na Cóclea

A membrana basilar contém pelo menos 20 000 *fibras basilares* que se projetam a partir do centro ósseo da cóclea até a parede externa, esta membrana não está sob tensão e pode ser comparada a uma folha gelatinosa coberta por camadas de fibras homogêneas e finas. Essas fibras estão incluídas na membrana basilar mas possuem a extremidade livre e por serem rijas podem vibrar como as palhetas de uma harmônica [5].

O comprimento e espessura das fibras basilares variam com relação à distância da membrana a partir da base, sendo as fibras mais finas e compridas próximo ao helicotrema. Estas diferenças provocam uma variação na rigidez das fibras de um fator de 100, sendo assim, as fibras mais curtas e rijas próximas à base tendem a vibrar com frequências maiores do que as longas próximas ao helicotrema [5, 19, 22].

#### 2.4.2 Transmissão das Ondas Sonoras na Cóclea

Se a base do estribo se move instantaneamente para dentro, a janela redonda deve mover-se para fora também instantaneamente, pois a cóclea é rodeada por parede ósseas. Como a inércia da onda líquida é alta, o líquido não terá tempo de percorrer todo o caminho até o helicotrema e de volta para a janela redonda, o efeito inicial é que a membrana basilar próxima à base faz proeminência na direção da janela redonda. Contudo, as tensões elásticas exercidas pelas fibras basilares quando elas se inclinam para a janela redonda iniciam uma onda que “viaja” ao longo da membrana basilar, conforme ilustra a figura 2.5, para três frequências diferentes.

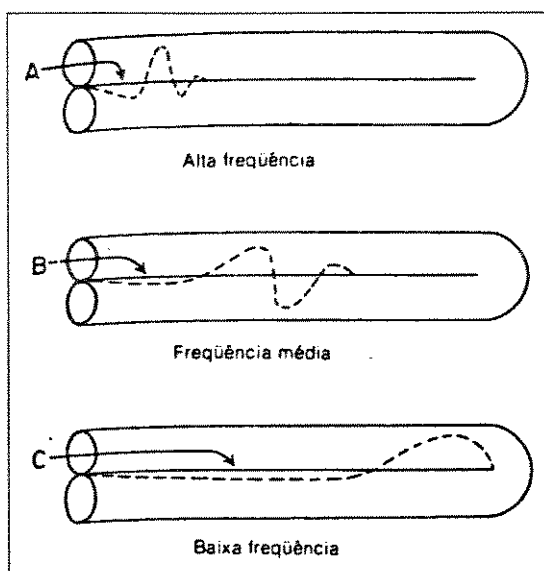


Figura 2.5: Representação esquemática das “ondas propagadas” ao longo da membrana basilar para frequências sonoras alta, média e baixa. Nota-se as regiões de maior amplitude [5].

### Padrão de Vibração da Membrana Basilar para Sons de Frequências Diferentes

Observa-se na figura 2.5 que existem tipos diferentes de transmissão para as ondas sonoras de frequências diferentes. Cada onda é relativamente débil no início, porém aumenta em potência nas porções da membrana basilar que tem frequência ressonante natural igual à frequência do som correspondente. Neste ponto a membrana basilar vibra de tal maneira que quase toda a energia da onda se dissipa por completo. Conseqüentemente a onda termina neste ponto e já não se propaga pelo restante da membrana basilar [5, 19, 22].

### Padrão da Amplitude de Vibração da Membrana Basilar

As curvas da figura 2.6A mostram a posição de uma onda sonora na membrana quando: (a) o estribo se encontra na posição mais interna; (b) o estribo no ponto neutro; (c) o estribo no ponto mais externo e (d) o estribo novamente no ponto neutro.

As envoltórias desses sinais fornecem a amplitude máxima de vibração da membrana basilar durante o ciclo vibratório completo para uma frequência específica. A figura 2.6B apresenta tipos de amplitudes de vibração para diferentes frequências. Observa-se que o terminal basal (perto do estribo) vibra pelo menos debilmente para todas as frequências,

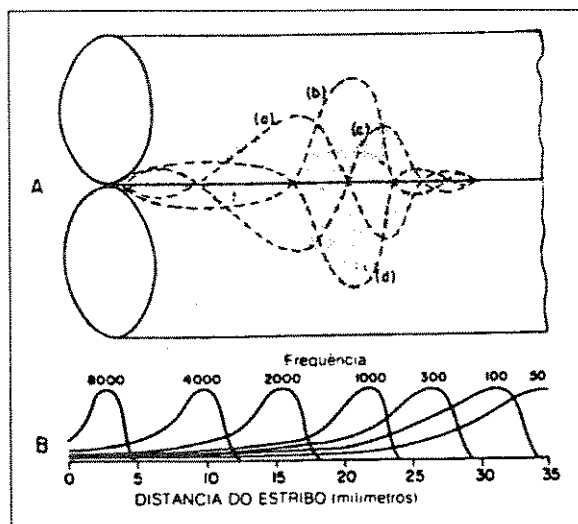


Figura 2.6: (A) Padrão da amplitude de vibração da membrana basilar para um som de frequência média. (B) Padrões de amplitude para sons de todas as frequências entre 50 e 8 000Hz, mostrando os pontos de máxima amplitude na membrana basilar para as diversas frequências [5].

contudo, além da área de ressonância para uma determinada frequência, a vibração mecânica da membrana basilar desaparece rapidamente [5, 19, 22].

### 2.4.3 Função do Órgão de Corti

O órgão de Corti, ilustrado nas figuras 2.3 e 2.7, é o órgão receptor que gera os impulsos nervosos em resposta à vibração da membrana basilar. Os receptores sensitivos reais do órgão de Corti consistem de dois tipos de *células ciliadas*, uma fileira única de *células ciliadas internas* totalizando cerca de 3 500 e medindo  $12\mu\text{m}$  de diâmetro e três a quatro fileiras de *células ciliadas externas* totalizando cerca de 20 000 com diâmetros de apenas  $8\mu\text{m}$  [5, 19, 22]. A figura 2.8 ilustra a disposição destas células na *lâmina reticular*, onde estas células encontram-se fixadas, para a cóclea de um gato<sup>8</sup> [22]. A base e os lados das células ciliadas estão incluídos numa rede de terminações do nervo coclear. Essas fibras nervosas vão para o *gânglio espiral de Corti*. O gânglio espiral, por sua vez, envia axônios ao *nervo coclear* e daí ao sistema nervoso central.

<sup>8</sup>A maioria dos animais possuem estruturas auditivas muito semelhantes ao do homem, sendo as principais diferenças a nível do córtex auditivo [19].

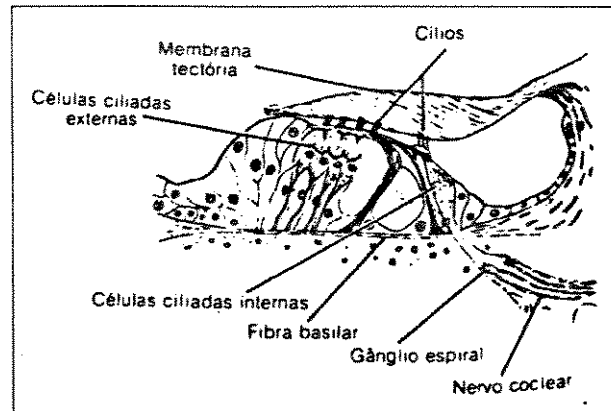


Figura 2.7: Órgão de Corti apresentando especialmente as células ciliadas e a membrana tectória contra a qual os cílios se projetam.

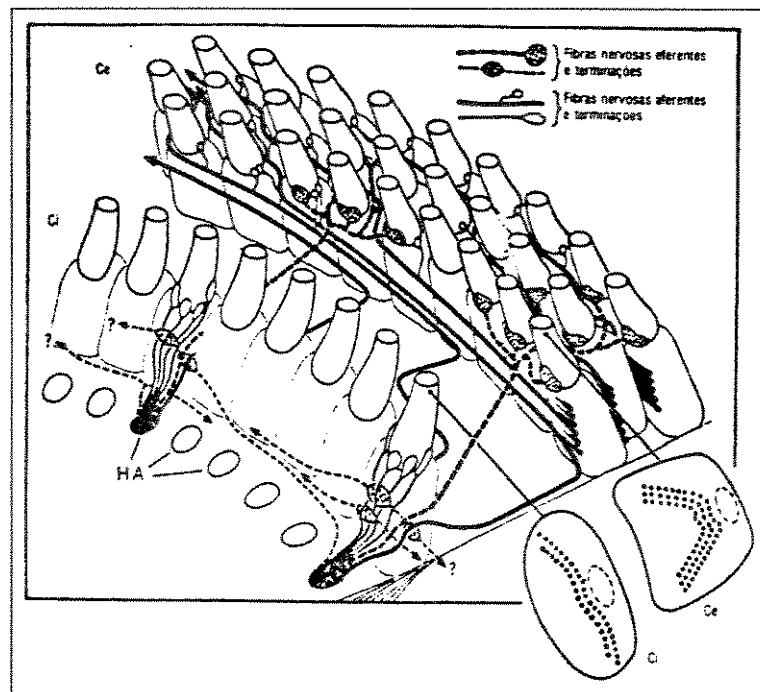


Figura 2.8: Esquema básico da posição e enervação das células ciliadas de um gato. Na figura Ci são as células ciliadas internas e Ce as externas. As terminações nervosas as fibras nervosas aferentes (para o cérebro) estão representadas pelas linhas contínuas e as eferentes (do cérebro) em linhas tracejadas. O canto inferior direito mostra a vista superior das células ciliadas [22].

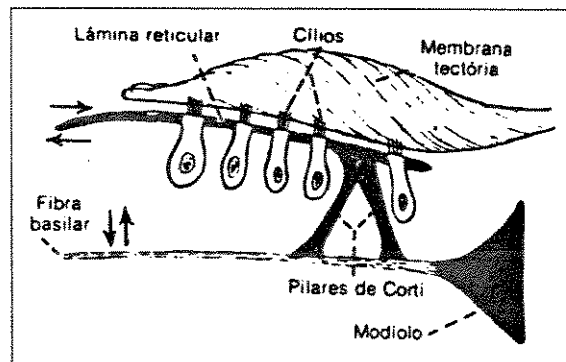


Figura 2.9: Estimulação das células ciliadas pelo movimento de vai-vem dos cílios na membrana tectória.

### Excitação das Células Ciliadas

Observa-se na figura 2.7 que pequenos pêlos ou cílios projetam-se para cima a partir das células ciliadas e, ou tocam, ou são incluídos na camada de gel da superfície da *membrana tectória*, que se encontra acima dos cílios na escala média [5]. A membrana tectória devido à sua composição é mais rígida não respondendo às vibrações, porém responde a deslocamentos estáticos vagarosos [22]. O movimento de inclinação dos cílios excita as células e estas, por sua vez, excitam as fibras nervosas que circundam a base.

A figura 2.9 ilustra o mecanismo pelo qual as vibrações da membrana basilar excitam os terminais dos cílios. Nota-se que os extremos superiores das células ciliadas estão fixados a uma estrutura denominada *lâmina reticular*. Além disso, a lâmina reticular é bastante rígida sendo contínua com uma estrutura triangular também rígida, denominada *pilares de Corti*, que se fixam nas fibras basilares.

O movimento para cima das fibras da membrana basilar move a lâmina reticular para cima e *para dentro*. A seguir quando a membrana basilar move-se para baixo, a membrana tectória se move para baixo e *para fora*. Estes últimos movimentos fazem com que os cílios oscilem em um ou em outro sentido na membrana tectória, o que excita as fibras do nervo coclear quando a membrana basilar vibra [5].

### Mecanismo pelo qual as Células Ciliadas Excitam as Fibras Nervosas

O deslocamento dos cílios em um ou outro sentido produz variações alternadas do potencial elétrico através da membrana celular das células que se encontram em movimento.

Esse potencial alternado é o *potencial gerador* das células ciliadas, e, por sua vez, estimula os terminais nervosos do nervo coclear que se encontram ao nível das células ciliadas [5, 22].

Quando as fibras da membrana basilar se deslocam no sentido da escala vestibular (figura 2.9) as células ciliadas se despolarizam e essa despolarização provoca uma quantidade crescente de potenciais de ação na fibra nervosa, aumentando a quantidade de pulsos nos neurônios. Mas, se a fibra basilar se move na direção oposta, as células ciliadas se hiperpolarizam e a quantidade de potenciais de ação caem, fazendo com que os neurônios cessem os disparos [5]. No entanto, se a membrana for mantida em uma mesma posição por um certo período de tempo, na qual os neurônios estariam emitindo pulsos, os neurônios passarão a uma nova condição onde cessarão os disparos. Esta característica destas células, que provocam os disparos quando movimentadas em uma única direção, são equivalentes a um retificador de meia-onda, apresentando ainda um ajuste de zero automático [19].

Observa-se na figura 2.8 que as células ciliadas externas não possuem fibras nervosas aferentes, que emitem pulsos em direção ao sistema nervoso, em contato com sua membrana, isto significa que as informações enviadas ao cérebro são obtidas somente pelas células ciliadas internas que possuem esse tipo de fibra. No entanto, os dois conjuntos de células possuem fibras eferentes, que recebem pulsos do sistema nervoso. De fato, os impulsos nervosos recebidos pelas células internas provocam a inibição lateral das células vizinhas à que se encontra em atividade, esse fenômeno causa um aumento da acuidade na classificação das freqüências. As células ciliadas externas não participam na classificação das freqüências, mas funcionam como pequenos músculos, que, quando recebem impulsos do sistema nervoso provocam um aumento ou diminuição da atividade das células ciliadas internas. Este sistema funciona como um *controle de ganho automático* e/ou um sistema de mascaramento, capaz de diminuir a influência de uma determinada componente de freqüência do som [19].

#### 2.4.4 Reconhecimento do Tom

Existe uma confusão entre o conceito de altura do som (tonalidade) e a sua freqüência, a diferença entre esses dois conceitos são mínimas. A tonalidade é a percepção consciente da freqüência sonora e pode não ser a mesma que a freqüência sonora real.

Por um lado, os sons de tom baixo (baixa freqüência) produzem um efeito máximo sobre a membrana basilar ao nível do vértice da cóclea e os sons de tom mais alto (freqüência



maior) ao nível da base da cóclea; para as frequências intermediárias, a membrana vibra em níveis intermediários desses dois extremos. Por outro lado, existe uma organização espacial das fibras do nervo auditivo, entre a cóclea e os núcleos do oitavo par no tronco cerebral; as fibras de cada zona respectiva da membrana basilar terminam numa região correspondente desses núcleos. Quando se registram os sinais das vias auditivas no tronco cerebral, e dos campos receptores auditivos no córtex, vê-se que certos neurónios são ativados por nervos específicos. Portanto, um método primário que o sistema nervoso utiliza para reconhecer diferentes tons é a identificação do ponto da membrana basilar no qual a vibração é máxima. A isto dá-se o nome de *hipótese tónica* do reconhecimento do tom, ou ainda *teoria posicional* [5, 19].

### Determinação da Intensidade Sonora

A intensidade sonora é determinada pelo sistema auditivo no mínimo de três maneiras diferentes [5]:

1. Quando o som aumenta de intensidade aumenta também a amplitude de vibração da membrana basilar e dos cílios, de modo que as células ciliadas excitam as terminações nervosas com mais intensidade.
2. A amplitude de vibração aumentando, faz com que aumente mais e mais o número de células ciliadas nas bordas da porção vibrante da membrana basilar, produzindo-se uma *soma espacial* de impulsos.
3. Algumas células ciliadas não são estimuladas até que a vibração da membrana basilar alcance intensidade elevada e, de alguma forma, a estimulação dessas células adverte o sistema nervoso de que o som é intenso.

### Detecção de Alterações na Intensidade Sonora

O ouvido pode discriminar alterações de intensidade de som variando desde um sussurro mais baixo até o mais intenso ruído, que tem aproximadamente *um trilhão de vezes* mais energia sonora. Contudo, o ouvido interpreta essa grande diferença de intensidade sonora como uma alteração de aproximadamente 1 para 10 000. Assim, a escala de intensidade encontra-se muito “comprimida” pelos mecanismos de percepção sonora do sistema auditivo. Isto, evidentemente, permite que uma pessoa interprete diferenças de intensidades

de som numa amplitude muito grande, muito maior do que seria possível se não existisse a compressão de escala.

Como pôde ser visto neste capítulo, o ouvido humano é um órgão extremamente complexo em sua estrutura fisiológica e em seu funcionamento. A cóclea, encarregada da tarefa mais importante e complicada do ouvido, a transdução dos sons em impulsos nervosos para o sistema nervoso central, é um órgão de modelamento trabalhoso devido às características não lineares e, principalmente, à sua capacidade de adaptar-se tanto no domínio do tempo quanto no da frequência. Este órgão é objeto de estudos tanto a nível da propagação das ondas sonoras pelos dutos e pela membrana basilar quanto a nível de sistemas capazes de emulá-la.

O capítulo seguinte entrará em detalhes sobre a construção de um modelo computacional simplificado capaz de emular algumas das características principais encontradas na cóclea.

## Capítulo 3

# Modelo Computacional da Cóclea

Apresenta-se neste capítulo o modelo computacional do ouvido interno (cuja fisiologia foi estudada no capítulo 2) desenvolvido por Richard F. Lyon [16, 17, 18]. Tal modelo é uma simplificação das funções da cóclea que converte sinais sonoros em representações neurais que consistem de impulsos nervosos propagados pelos neurônios até o sistema nervoso central. O modelo computacional da cóclea tenta manter as características de separabilidade em frequência dos sons e parametrização da voz realizada pelo ouvido interno. A resposta deste modelo, no entanto, consiste de um vetor cujos elementos constituem-se de valores de “intensidade” e não de impulsos nervosos como os fornecidos pela cóclea biológica.

Tal modelo foi obtido a partir de estudos realizados na cóclea sobre a propagação de ondas mecânicas nos fluídos internos de seus dutos. Este trabalho não entrará em detalhes sobre o modelo de propagação das ondas restringindo-se apenas ao modelo final. Referências sobre este assunto em específico podem ser obtidas diretamente através dos trabalhos de Lyon [16, 17, 18, 19] ou através de trabalhos sobre a mecânica da cóclea como de Shroeder [28], Zweig et al. [36] e Lighthill [14].

A capacidade de adaptação do ouvido humano aos sons de determinadas línguas é uma de suas características mais marcantes, podendo causar dificuldades na comunicação entre povos de línguas diferentes, que certamente possuirão um conjunto diferente de sons para a fala. Esta dificuldade pode ser observada mesmo em se tratando de línguas de mesma origem. Pode-se dar como exemplo o fato de que pessoas de língua espanhola não entendem completamente (ou não conseguem ouvir) todos os fonemas utilizados pela língua portuguesa (brasileira). Curiosamente a dificuldade que um brasileiro tem em compreender os fonemas espanhóis é bem menor. O modelo do ouvido desenvolvido por Lyon é relati-

vamente complexo e possui uma grande quantidade de parâmetros. Neste trabalho, foram utilizados os parâmetros sugeridos por Lyon e por Slaney [29], ambos relacionados à língua inglesa. Estes parâmetros podem não estar perfeitamente adequados para os sons da língua portuguesa devendo ser alterados futuramente para um melhor ajuste do modelo.

Conforme foi visto no capítulo 2, as ondas sonoras entram, no ouvido interno, pela *janela oval*, propagam-se pelo interior da cóclea e, devido às diferenças de elasticidade apresentada pela *membrana basilar*, uma determinada porção desta membrana irá ressoar a uma determinada frequência. Tal movimento é detectado pelas *células ciliadas internas* que convertem os movimentos em impulsos detectáveis pelo *sistema nervoso central*. É importante observar que as ondas sonoras são separadas em frequência e que um determinado ponto no interior da cóclea responde melhor a uma determinada frequência. Em essência a cóclea mapeia o conteúdo de frequência da onda sonora em um *sistema posicional*. Perto da base (janela oval) a cóclea é sensível às altas frequências e, à medida que os sons se propagam pelo interior do ouvido interno, a sensibilidade desloca-se para as frequências baixas [29].

A simplificação realizada por Lyon consiste em modelar a cóclea e a membrana basilar por uma série de filtros lineares, independentes e em cascata, diferente da cóclea onde a membrana é essencialmente interativa. Estes filtros correspondem a discretizações de determinadas regiões (secções longitudinais) da cóclea, caracterizando o *sistema posicional*. Toda não-linearidade do ouvido interno foi modelada nas camadas de detecção e compressão para facilitar a implementação deste modelo. As camadas de detecção e compressão correspondem a *retificadores de meia-onda*, que na cóclea é realizada pelas *células ciliadas internas* [5, 22], e um conjunto de *controladores automáticos de ganho*, (*CAG*), realizados pelas *células ciliadas externas* [19]. O diagrama de blocos do modelo pode ser observado na figura 3.1.

Quanto à construção da camada dos filtros, existem duas formulações propostas [16, 18]: na primeira, cada estágio de filtragem é implementado como dois filtros separados e na segunda implementação estes dois filtros são combinados em apenas um filtro de segunda ordem. Na primeira implementação o som passa por uma cascata de filtros que modelam a propagação da onda pelo interior da cóclea e seguem por um banco de filtros em paralelo com os anteriores que modelam o movimento da membrana basilar. Esta versão é conhecida como *implementação cascata-paralelo*. A segunda versão é conhecida por *implementação cascata*. Malcolm Slaney, em seu relatório técnico [29], descreve ambas implementações

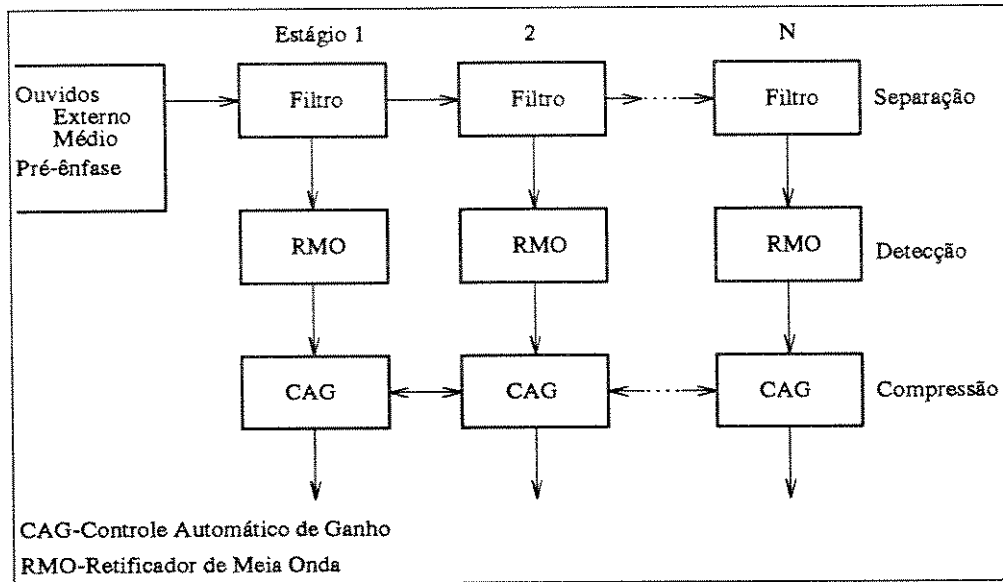


Figura 3.1: Esquematização do modelo da cóclea [29].

detalhadamente.

### 3.1 Implementação Cascata-Paralelo

A implementação cascata-paralelo, descrita por Richard F. Lyon [16] e detalhada por Malcolm Slaney [29] arranja uma série de *filtros anti-ressonantes*,<sup>1</sup> que modelam a propagação da onda pela cóclea, com *filtros ressonantes*<sup>2</sup> em paralelo aos filtros anti-ressonantes, que convertem as ondas em movimentos da *membrana basilar*, segundo a estrutura mostrada na figura 3.2. Os filtros anti-ressonantes são constituídos por um par de pólos complexos com baixo fator de qualidade e por um par de zeros complexos com um fator de qualidade alto (cerca de 5 vezes maior). Os filtros ressonantes possuem um zero na origem e um par de pólos complexos exatamente na mesma posição dos pólos do filtro anti-ressonante do canal seguinte (frequência um pouco abaixo dos pólos do filtro anti-ressonante), com o mesmo fator de qualidade. O modelo obtido com o uso destes filtros é tanto melhor quanto menor for a secção longitudinal utilizada. A disposição dos pólos e zeros dos filtros pode ser vista na figura 3.3. Detalhes da construção dos mesmos serão apresentados a seguir.

O efeito conjunto da cascata de filtros anti-ressonantes equivale a uma cascata de

<sup>1</sup>Filtros rejeita-faixa.

<sup>2</sup>Filtros passa-faixa.

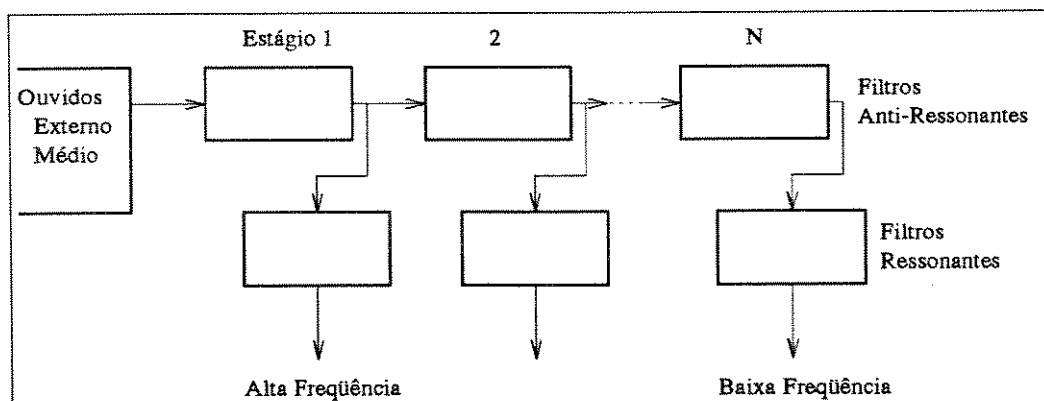


Figura 3.2: Esquema da implementação cascata-paralelo dos filtros da cóclea [16].

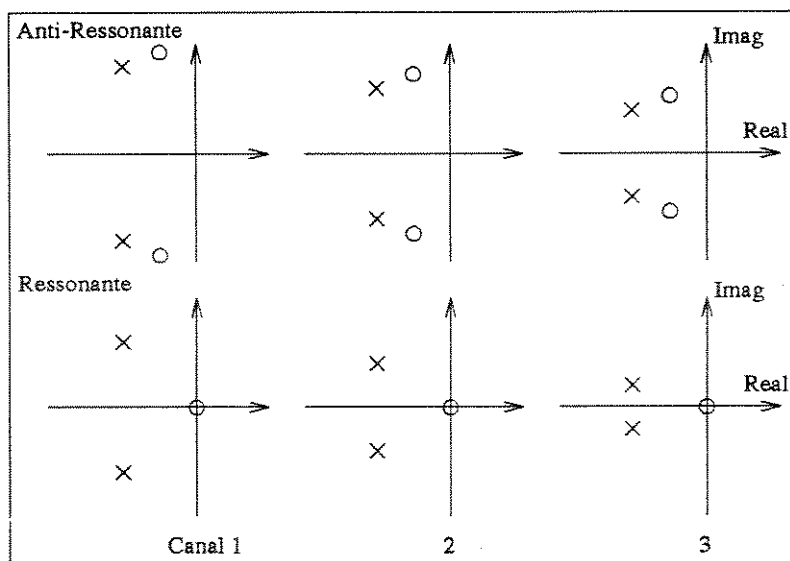


Figura 3.3: Posicionamento dos polos ( $\times$ ) e zeros ( $\circ$ ) da implementação cascata-paralelo dos filtros da cóclea [29], dispostos no plano  $s$ .

filtros passa-baixa, devido aos mesmos operarem em frequências cada vez mais baixas, retirando gradualmente da onda sonora as componentes das frequências altas até as frequências mais baixas.

### 3.1.1 Parâmetros da Implementação Cascata-Paralelo

Os parâmetros para a implementação dos filtros do modelo cascata-paralelo foram obtidos experimentalmente através de medições realizadas na cóclea e pelas equações que descrevem a propagação das ondas sonoras pelo seu interior [16, 28]. Para uma maior facilidade de entendimento será usado o domínio da transformada de Laplace ( $s$ ).

Considera-se nesta implementação a propriedade de *escalabilidade* encontrada na cóclea. O meio por onde as ondas se propagam é dito ser *escalável* se as propriedades da resposta em um determinado ponto do meio à propagação das ondas é similar à resposta encontrada em qualquer outro ponto com uma mudança na escala de tempo. Portanto, em um sistema escalável, a resposta para cada secção pode ser especificada por uma única função de transferência  $H(\omega/\omega_n)$ , onde  $\omega/\omega_n$  é uma frequência adimensional normalizada, e  $\omega_n$  é alguma frequência natural convenientemente definida.

Para se obter as frequências naturais  $\omega_n$  ao longo da cóclea geralmente supõe-se que a variação de  $\omega_n$  ao longo da distância a partir da base da cóclea seja geométrica, isto é,  $\omega_n$  começa com um valor máximo de frequência perto da base e decai exponencialmente à medida que a distância a partir da base aumenta. Esta característica é observada na cóclea principalmente para as frequências acima de 1kHz [28]. Este tipo de variação da frequência ao longo da cóclea é utilizada para se obter uma solução analítica para as equações de propagação da onda sonora [28, 36]. Para frequências abaixo de 1kHz Slaney [29] utiliza um decaimento linear da frequência natural.

Os pólos e zeros dos filtros nesta implementação são sempre pares de raízes complexas conjugadas podendo ser caracterizados por um polinômio em  $s$  do tipo  $P(s) = s^2 + (2\pi F/Q)s + 4\pi^2 F^2$  onde  $F$  é a frequência natural não amortecida em Hertz e  $Q$  é o fator de qualidade associado às raízes do polinômio. O fator de qualidade relaciona-se com o amortecimento  $\xi$  pela igualdade  $\xi = (2Q)^{-1}$ . Portanto, para se obter os filtros, é necessário encontrar apenas os valores das frequências naturais não amortecidas ( $F$ ) e os fatores de qualidade ( $Q$ ) dos pólos e zeros.

Numerando cada estágio de filtros a partir da base da cóclea<sup>3</sup> pode-se definir a frequência central de operação dos estágios (canais) da cascata de filtros. Os efeitos da anti-ressonância e da ressonância de cada canal da cóclea ocorre nas proximidades da frequência central do canal.

Iniciando-se com uma frequência arbitrária qualquer, que define a maior frequência que o modelo da cóclea pode “ouvir”, a frequência central de cada canal da cóclea decrescerá por uma fração da banda de passagem do canal anterior. A frequência central ( $F_c$ ) para cada canal pode ser obtida recursivamente através da seguinte equação:

$$\begin{aligned} F_{c_0} &= 10\,000\text{Hz} \\ F_{c_i} &= F_{c_{(i-1)}} - P \cdot B_{(i-1)}, \quad i = 1, \dots, N, \end{aligned} \quad (3.1)$$

onde  $P$  define o número de estágios sobrepostos à banda de passagem de cada estágio de filtros e  $B_i$  é a banda de passagem, em Hertz, do  $i$ -ésimo canal. Existe um limite inferior para a frequência central  $F_c$  em torno de 63Hz que será detalhado a seguir. A constante  $P$  é arbitrária e determina, junto com  $F_{c_0}$ , o número de estágios ( $N$ ) de filtragem possíveis de serem implementados. No decorrer deste trabalho foram utilizados 5 ( $P = 0,2$ ) filtros sobrepostos a banda de passagem de cada canal (para a implementação cascata com uma frequência de amostragem de 8kHz a constante  $P$  produz 80 estágios de filtragem).

A banda de passagem em Hertz,  $B_i$ , é obtida pela equação

$$B_i = \frac{\sqrt{F_{c_i}^2 + F_q^2}}{Q_m}, \quad (3.2)$$

onde  $F_q = 1\text{kHz}$  determina o ponto de quebra entre a variação linear (abaixo de 1kHz) e exponencial (acima de 1kHz) da frequência central  $F_c$ .  $Q_m = 8$  é o maior valor para o fator de qualidade dos pólos dos filtros e é adimensional. Pode-se observar que, para frequências acima de 1kHz,  $B_i \approx F_{c_i}/Q_m$  que substituído na equação (3.1) produz a variação exponencial das frequências centrais ( $F_{c_i} \approx C^i F_{c_0}$ ). De maneira similar, para as frequências abaixo de 1kHz,  $B_i \approx F_q/Q_m$ , produzindo uma variação linear das frequências centrais ( $F_{c_i} \approx F_{c_0} - C \cdot i$ ).

As frequências naturais não amortecidas dos pólos dos filtros anti-ressonantes, no modelo de Lyon, estão localizadas exatamente nas frequências centrais dos estágios, ou seja,

$$F_{p_i} = F_{c_i}, \quad (3.3)$$

---

<sup>3</sup>A base da cóclea está localizada perto da janela oval onde o ossículo estribo se conecta.



onde  $F_{p_i}$  é dado em Hertz. O fator de qualidade  $Q_{p_i}$ , dos pólos dos filtros podem ser obtidos pela equação:

$$Q_{p_i} = \frac{F_{p_i}}{B_i}, \quad (3.4)$$

onde  $Q_{p_i}$  é adimensional e determina o grau de ressonância dos pólos. Estes pólos produzem uma pequena amplificação na frequência central da cada canal.

Uma vez que este modelo da cóclea utiliza-se de pares complexos de pólos e zeros, vê-se pelo polinômio  $P(s) = s^2 + (2\pi F/Q)s + 4\pi^2 F^2$  que para raízes complexas conjugadas é necessário que o fator de qualidade  $Q$  seja maior que  $\frac{1}{2}$ . Pela equação (3.4) isto ocorre quando  $F_c$  é maior que aproximadamente 63Hz. Como  $F_c$  deve ser sempre maior que 63Hz e com  $F_{c_0} = 10\text{kHz}$  e  $P = 0,2$  obtém-se o número máximo de canais  $N = 116$  da cóclea.

As frequências naturais não amortecidas dos zeros dos filtros anti-ressonantes estão localizadas um pouco acima das frequências dos pólos, para provocar a rejeição de banda. As frequências naturais dos zeros são obtidas através de

$$F_{z_i} = F_{p_i} + D \cdot P \cdot B_i, \quad (3.5)$$

onde  $D = 0,5$  é uma constante que especifica o quão distante estão os zeros com relação aos pólos. Observa-se que a variação das frequências naturais dos zeros é função da variação das frequências centrais ( $P \cdot B_i$ ) de filtro a filtro.

O fator de qualidade dos zeros do filtro anti-ressonante é dado por

$$Q_{z_i} = S \cdot \frac{F_{z_i}}{B_i}, \quad (3.6)$$

onde<sup>4</sup>  $S = 5$  é uma constante que especifica o quão acentuado é o efeito da anti-ressonância (zeros) comparada ao efeito da ressonância (pólos).

Nota-se que a resposta em frequência de um filtro anti-ressonante será deslocada para uma frequência um pouco menor que a do filtro imediatamente anterior a este.

Na implementação realizada por Lyon [29] a posição dos pólos do filtro ressonante estão localizados na mesma posição dos pólos do filtro ressonante, deslocados de um canal (figura 3.3). Assim, as equações (3.3) e (3.4) podem ser utilizadas para a obtenção dos filtros ressonantes.

---

<sup>4</sup>Do inglês Sharpness.

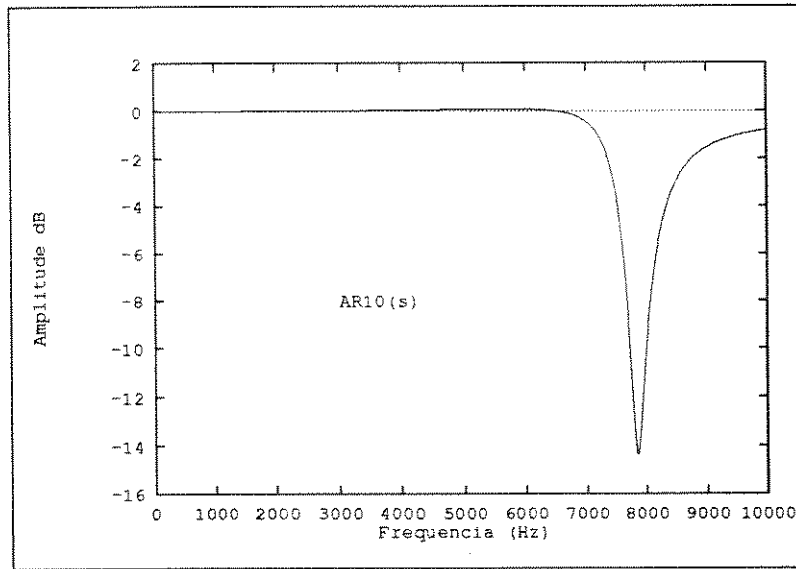


Figura 3.4: Resposta em frequência do filtro anti-ressonante do décimo canal da cóclea.

### 3.1.2 Construção dos Filtros

Com base nos parâmetros obtidos, pode-se construir os filtros de cada canal do banco de filtros da cóclea. As equações (3.3) e (3.5) fornecem, respectivamente, os valores das frequências naturais não amortecidas dos pólos e zeros dos filtros anti-ressonantes e as equações (3.4) e (3.6) os valores dos fatores de qualidade associados aos pólos e zeros, respectivamente.

Assim, um filtro anti-ressonante poderá ser caracterizado pela seguinte função de transferência:

$$AR_i(s) = K \cdot \frac{s^2 + (2\pi F_{z_i}/Q_{z_i})s + 4\pi^2 F_{z_i}^2}{s^2 + (2\pi F_{p_i}/Q_{p_i})s + 4\pi^2 F_{p_i}^2}, \quad (3.7)$$

onde  $K$  é uma constante tal, que o ganho DC do filtro seja igual a 1,0. O filtro ressonante correspondente poderá ser obtido através da seguinte função de transferência:

$$R_i(s) = K \cdot \frac{s}{s^2 + (2\pi F_{p_{i+1}}/Q_{p_{i+1}})s + 4\pi^2 F_{p_{i+1}}^2}, \quad (3.8)$$

onde  $K$  é obtido de tal forma que na frequência  $F_{c_{i+1}}$  o ganho do filtro seja unitário.

A figura 3.4 mostra a resposta em frequência do filtro anti-ressonante do décimo canal da cóclea ( $i = 10$ ), onde pode ser observado o alto fator de qualidade do par de zeros. Na figura 3.5 têm-se o efeito cascata do primeiro ao décimo filtros anti-ressonantes onde nota-se o resultado global como um filtro passa-baixa. A figura 3.6 mostra a resposta em

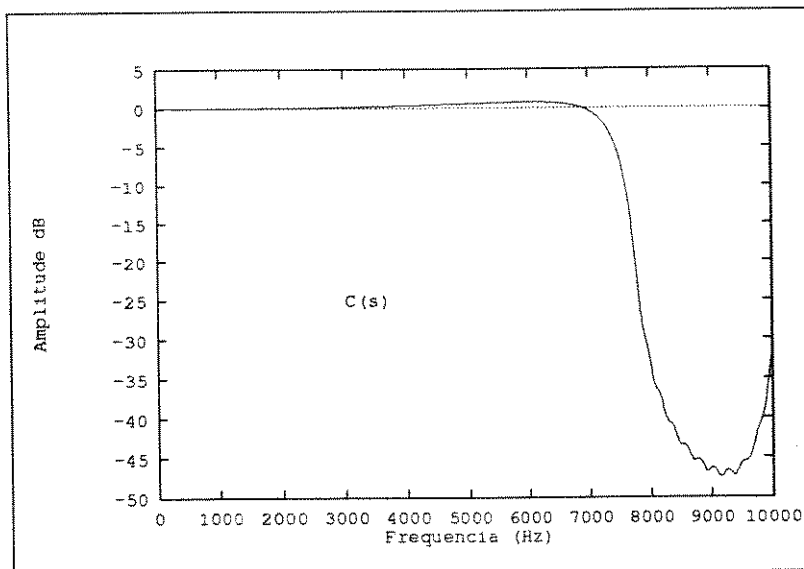


Figura 3.5: Resposta em frequência da cascata de filtros anti-ressonantes do primeiro até o décimo canal da cóclea.

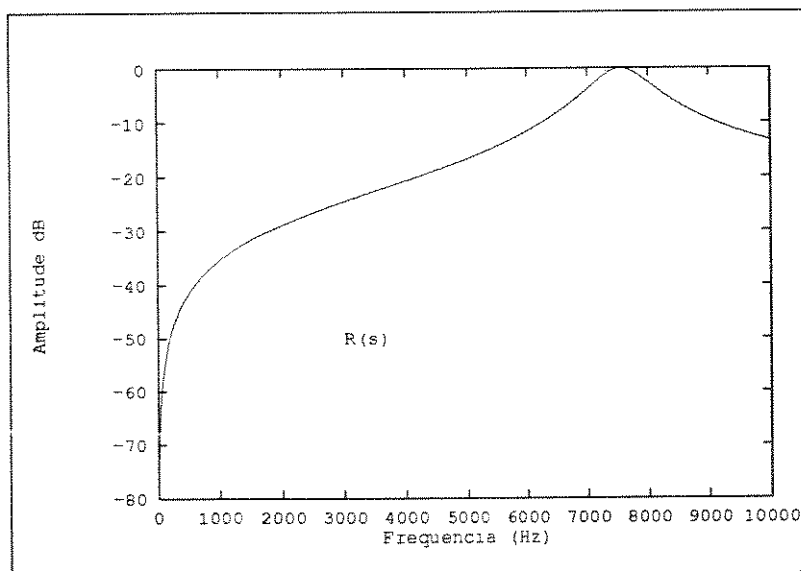


Figura 3.6: Resposta em frequência do filtro ressonante do décimo canal da cóclea.

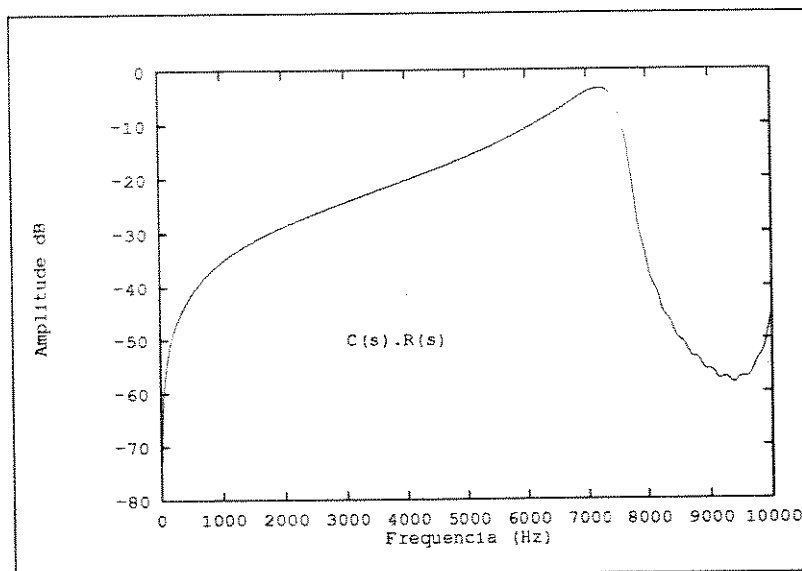


Figura 3.7: Resposta em frequência do décimo canal da cóclea.

frequência do filtro ressonante, mostrando os efeitos do zero na origem e do par de pólos com baixo fator de qualidade em  $F_{c11}$ . E a figura 3.7 mostra a resposta em frequência da saída do décimo canal da cóclea. Nestes gráficos não foram considerados os efeitos do ouvido externo e do médio. Pode-se observar ainda que, a partir de aproximadamente 9kHz, há um aumento no ganho do canal, este efeito pode ser eliminado adicionando-se um filtro passa-baixa com frequência de corte em 10kHz, pois foi suposto que este ouvido não “escutará” nada além de 10kHz, conforme foi dito anteriormente.

## 3.2 Implementação Cascata

Esta implementação combina os filtros anti-ressonantes e ressonantes em um único filtro, com uma disposição de pólos e zeros conforme é dada na figura 3.8 (os pólos e zeros estão localizados nas mesmas posições dos da figura 3.3). Tal alteração não afeta o comportamento dos filtros na cascata, pois os pólos do filtro ressonantes estão dispostos nas mesmas posições dos pólos do filtro anti-ressonante do estágio seguinte, como pôde ser notado anteriormente na implementação dos filtros. A única vantagem desta implementação é reduzir o esforço computacional necessário para a execução do modelo [29] e por este motivo esta implementação foi utilizada neste trabalho.

O diferenciador do filtro ressonante (zero na origem) é deslocado para um estágio

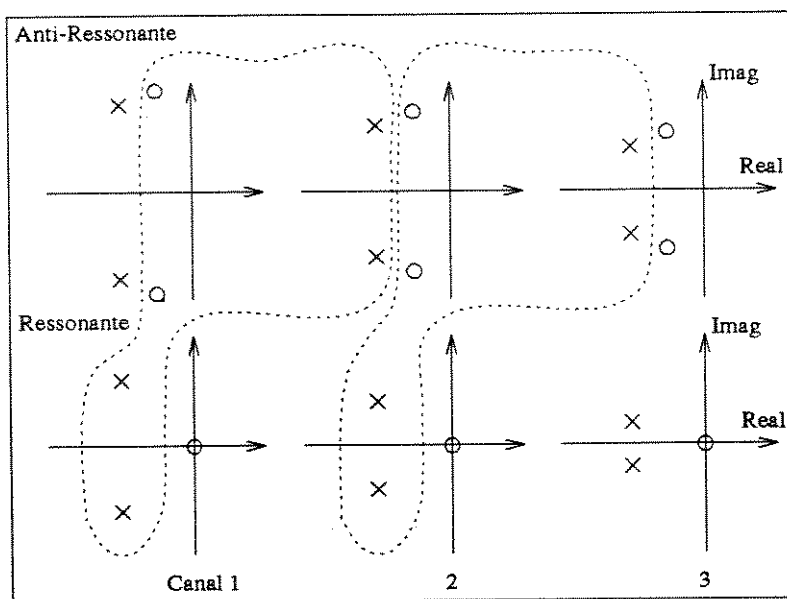


Figura 3.8: Posicionamento e combinação dos pólos ( $\times$ ) e zeros ( $\circ$ ) na implementação cascata dos filtros da cóclea [29], dispostos no plano  $s$ .

de pré-ênfase colocado antes do início do banco de filtros [17, 29]. O deslocamento é possível devido aos filtros serem lineares. Este estágio de pré-ênfase inclui também os efeitos dos ouvidos externo e médio e o do par de pólos do primeiro estágio da cascata.

A saída dos filtros agora alimentarão diretamente as camadas de detecção e os estágios seguintes de filtragem. O modelo do ouvido assume então o formato mostrado na figura 3.9.

Os parâmetros usados por este modelo são em essência os mesmos usados no modelo cascata-paralelo, devido aos pólos e zeros estarem nas mesmas posições, portanto pode-se usar as mesmas equações apresentadas na seção 3.1.1. No entanto a constante  $D$  usada na equação (3.5) que estava ajustada em 0,5 foi elevada para 1,5 avançando os zeros com relação aos pólos em um canal [29]. O motivo do avanço dos zeros com relação aos pólos em um canal pode ser observado mais facilmente na figura 3.8 onde nota-se que o único filtro do canal é formado pelos pólos do filtro ressonante e pelos zeros do filtro anti-ressonante da implementação cascata-paralelo. Na implementação cascata-paralelo os pólos do filtro ressonante são obtidos através da equação (3.3) ( $F_{p_i} = F_{c_i}$ ) e estão deslocados de um canal com relação aos pólos do filtro anti-ressonante. Como os zeros são obtidos através da equação 3.5 ( $F_{z_i} = F_{p_i} + D \cdot P \cdot B_i$ ) faz-se necessário o uso de  $D = 1,5$ .

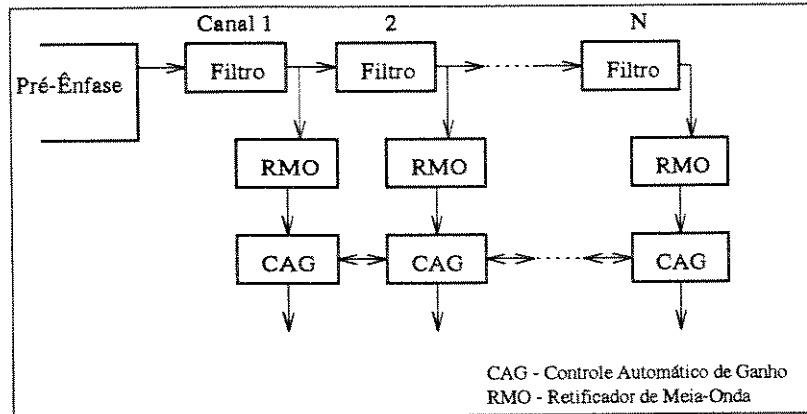


Figura 3.9: Esquematização da implementação cascata dos filtros da cóclea [29].

Considerando a discretização no tempo desta implementação, a frequência central de cada filtro deverá ser calculada de modo diferente, sendo a maior frequência possível dada por

$$F_{c_0} = \frac{F_a}{2} - \left[ F_z \left( \frac{F_a}{2} \right) - \frac{F_a}{2} \right] + P \cdot B \left( \frac{F_a}{2} \right), \quad (3.9)$$

onde  $F_a$  é a frequência de amostragem em Hertz;  $B(F_c)$ ,  $F_z(F_c)$  são as equações (3.2) e (3.5), respectivamente, modificadas para tornarem-se função da frequência central  $F_{c_i}$  e  $P$  é o mesmo parâmetro da implementação cascata-paralelo. Observa-se que  $F_{c_0}$  é função da frequência de Nyquist ( $F_a/2$ ).

O cálculo da frequência central de cada filtro é o mesmo da implementação anterior e é dado por

$$F_{c_i} = F_{c_{(i-1)}} - P \cdot B_{(i-1)}, \quad (3.10)$$

onde  $i$  assume valores de 1 até um número máximo  $N$  de canais possíveis, que com  $P = 0,2$  e com uma frequência de amostragem de 8kHz será de 80 canais ou estágios de filtragem.

### 3.2.1 Estágio de Pré-Ênfase

Os efeitos dos ouvidos médio e externo são modelados por Malcolm Slaney, em [29], como um único estágio de pré-ênfase. Este modelo é bastante rudimentar e desconsidera os mecanismos de proteção e atenuação do ouvido médio, porém, para uso em processamento de voz, mostra-se suficientemente eficiente e foi adotado neste trabalho.

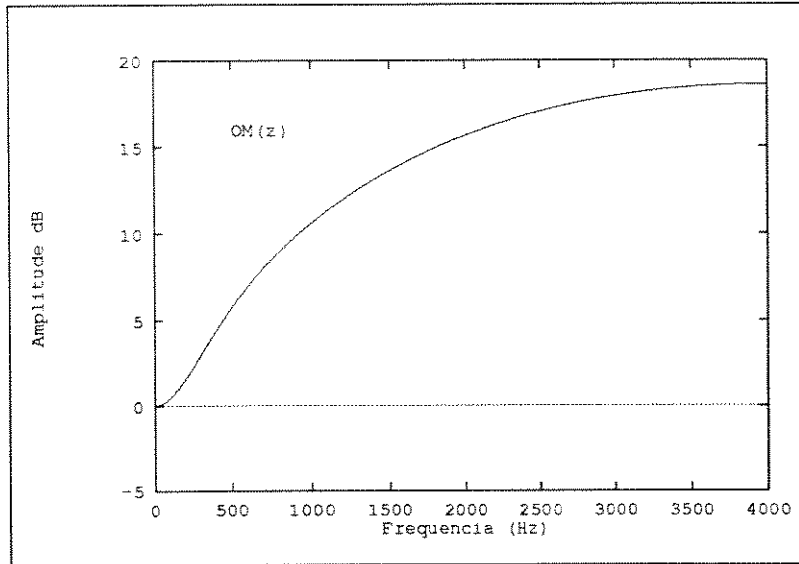


Figura 3.10: Resposta em frequência do modelo dos ouvidos externo e médio para uma amostragem de 8kHz.

Os ouvidos médio e externo introduzem na cóclea um ganho suave nas altas frequências, ajudando a normalizar a entrada para o ouvido interno e tornando mais simples o processamento da cóclea [29]. Os ouvidos externo e médio são modelados por um único estágio de filtragem correspondente a um filtro passa-alta com frequência de corte de 300Hz.

A função de transferência deste filtro é dada por

$$OM(z) = K \cdot \frac{1 - \exp(-2\pi 300/F_a) \cdot z}{z} \quad (3.11)$$

onde  $K$  é uma constante tal que o ganho DC do filtro seja unitário e  $F_a$  é a frequência de amostragem. A resposta em frequência deste filtro é mostrada na figura 3.10 para  $F_a = 8000\text{Hz}$ .

Cada estágio de filtro ressonante da implementação cascata-paralela descrita por Richard F. Lyon em [16] usa um diferenciador para converter ondas de pressão em movimentos da membrana basilar. Esta implementação, no entanto, desloca o diferenciador  $(1 - z)$  para o estágio de pré-ênfase e adiciona ainda um diferenciador na frequência de Nyquist  $(1 + z)$  para compensar a proximidade entre os pólos perto de  $z = -1$  nas altas frequências. Os ganhos dos diferenciadores da implementação cascata-paralelo serão compensados na construção dos filtros de cada canal.

Este estágio “compensador” fica caracterizado pela seguinte função de transferên-

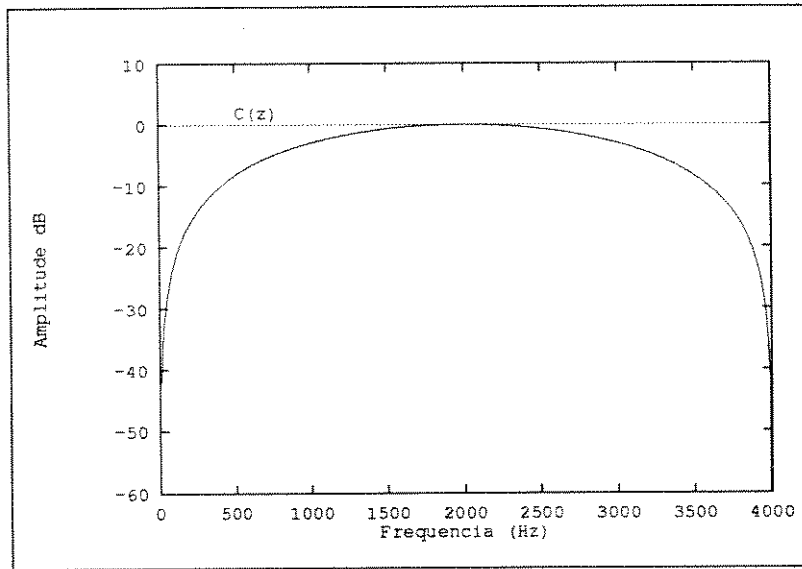


Figura 3.11: Resposta em frequência do filtro compensador do estágio de pré-ênfase da cóclea para uma amostragem de 8kHz.

cia,

$$C(z) = K \frac{(1+z)(1-z)}{z^2}, \quad (3.12)$$

onde  $K$  é ajustado de modo a  $C(z)$  apresentar ganho unitário em  $F_a/4$ . A figura 3.11 ilustra a resposta em frequência deste filtro, que combinado com o filtro anterior ( $OM(z)$ ), produz a resposta em frequência dada pela figura 3.12.

Finalmente, adiciona-se a este estágio o par de pólos do primeiro estágio da cascata. Os dois primeiros pólos podem ser modelados pela seguinte função de transferência [29],

$$FP(z) = \frac{1}{z^2 - (2\rho \cos \theta)z + \rho^2} \quad (3.13)$$

$$\rho = \exp\left(-\pi \frac{F_{p0}}{F_a \cdot Q_{p0}}\right)$$

$$\theta = 2\pi \frac{F_{p0}}{F_a} \sqrt{1 - \frac{1}{4 \cdot Q_{p0}^2}}$$

que combinado com os filtros anteriores fornece o estágio de pré-ênfase desta implementação, dado pela função de transferência:

$$F(z) = OM(z) \cdot C(z) \cdot FP(z), \quad (3.14)$$

cuja resposta em frequência pode ser vista na figura 3.13 para  $F_a = 8\text{kHz}$ .



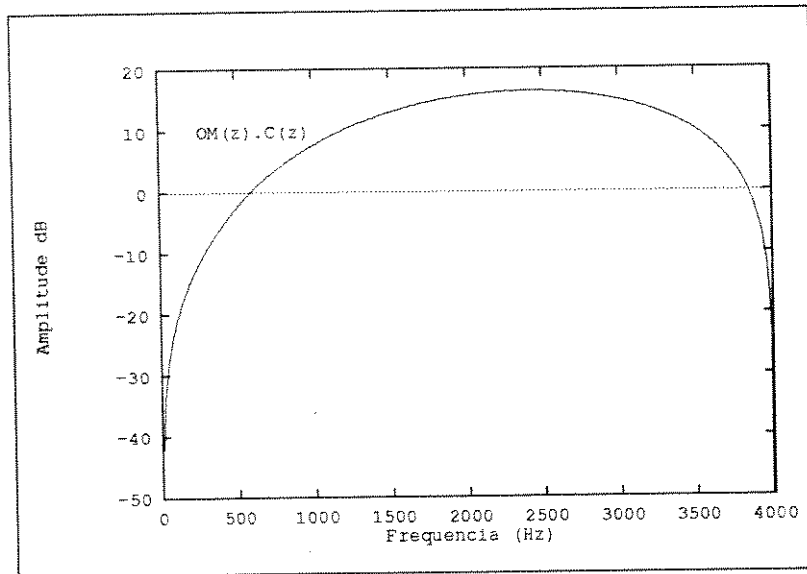


Figura 3.12: Resposta em freqüência dos filtro compensador e do modelo do ouvido externo e médio para uma amostragem de 8kHz.

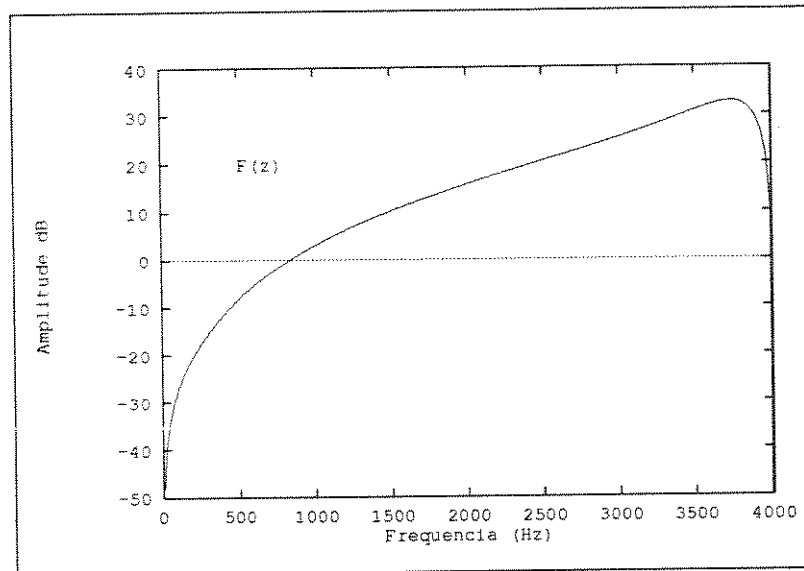


Figura 3.13: Resposta em freqüência do filtro de pré-ênfase para uma amostragem de 8kHz.

### 3.2.2 Filtros dos Estágios em Cascata

Os filtros dos estágios da cascata agora são modelados como uma combinação de um par de pólos e um par de zeros. Cada um desses filtros pode ser representado pela seguinte função de transferência [29],

$$EF_i = K \cdot \frac{z^2 - (2\rho_{z_i} \cos(\theta_{z_i}))z + \rho_{z_i}^2}{z^2 - (2\rho_{p_i} \cos(\theta_{p_i}))z + \rho_{p_i}^2}, \quad (3.15)$$

onde  $K$  é uma constante tal que o ganho do filtro para níveis DC seja unitário e  $\rho_{z_i}$ ,  $\theta_{z_i}$ ,  $\rho_{p_i}$  e  $\theta_{p_i}$  são dados pelas seguintes equações,

$$\begin{aligned} \rho_{z_i} &= \exp\left(-\pi \frac{F_{z_i}}{F_a \cdot Q_{z_i}}\right), \\ \rho_{p_i} &= \exp\left(-\pi \frac{F_{p_i}}{F_a \cdot Q_{p_i}}\right), \\ \theta_{z_i} &= 2\pi \frac{F_{z_i}}{F_a} \sqrt{1 - \frac{1}{4 \cdot Q_{z_i}^2}}, \\ \theta_{p_i} &= 2\pi \frac{F_{p_i}}{F_a} \sqrt{1 - \frac{1}{4 \cdot Q_{p_i}^2}}. \end{aligned}$$

Um ajuste no ganho deverá ser realizado nestes filtros para compensar o ganho fornecido pelo diferenciador no estágio de pré-ênfase. O ganho de um diferenciador ideal é proporcional a frequência, assim, fornecendo um ganho dependente da frequência de cada estágio de filtro da cóclea compensa-se este problema.

Malcolm Slaney em [29] usa como compensação de ganho a seguinte relação,

$$g_i = \frac{F_{c_{i-1}}}{F_{c_i}}, \quad (3.16)$$

cujos efeitos é normalizar o diferenciador para mantê-lo com ganho unitário na frequência central do filtro.

A função de transferência dos filtros dos estágios da cascata com a compensação de ganho passa a ser,

$$EF_i(z) = g_i \cdot K \cdot \frac{z^2 - (2\rho_{z_i} \cos(\theta_{z_i}))z + \rho_{z_i}^2}{z^2 - (2\rho_{p_i} \cos(\theta_{p_i}))z + \rho_{p_i}^2}, \quad (3.17)$$

que têm os mesmos valores de  $K$  e dos parâmetros  $\rho_{z_i}$ ,  $\theta_{z_i}$ ,  $\rho_{p_i}$  e  $\theta_{p_i}$  da equação (3.15).

A resposta em frequência para o vigésimo quinto canal da cóclea, que possui a frequência central perto de 2kHz, pode ser visto na figura 3.14, independente do restante

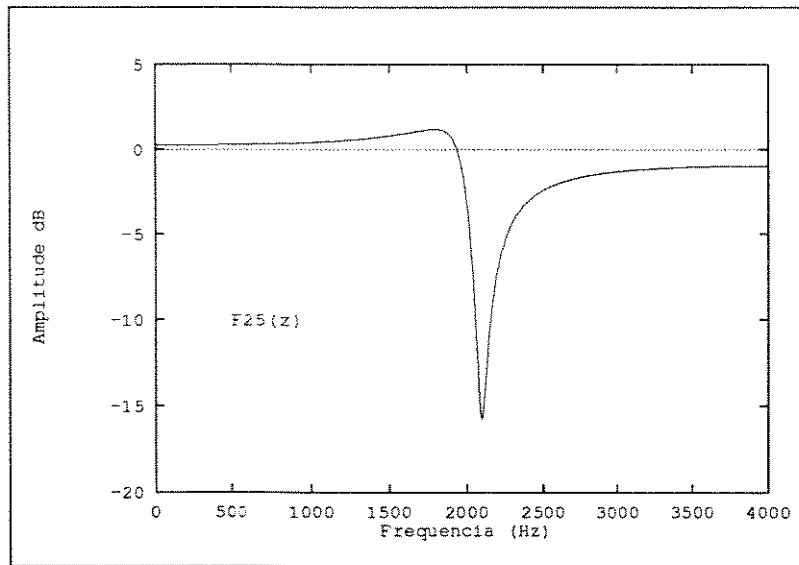


Figura 3.14: Resposta em frequência do vigésimo quinto (isolado dos demais) canal do modelo da cóclea, sem considerar os efeitos do estágio de pré-ênfase, para uma frequência de amostragem de 8kHz.

dos filtros da cascata. Observa-se a semelhança com a resposta de um canal do modelo cascata-paralelo.

Na figura 3.15 estão representadas as respostas em frequência de quatro canais do modelo da cóclea incluindo nestas respostas o filtro de pré-ênfase que engloba o modelo dos ouvidos externo e médio, o filtro compensador e o par de pólos do primeiro canal. Pode-se observar na figura que o efeito do modelo do ouvido externo e médio começa a se salientar nos ganhos dos canais de índices maiores, que correspondem as baixas frequências.

### 3.3 Detecção na Cóclea

A detecção na cóclea é realizada pelas *células ciliadas internas*, localizadas no *órgão de Corti* na *membrana basilar* [5, 22], que possuem a característica de responder somente quando os cílios são deslocados em uma certa direção [22]. Este tipo de comportamento assemelha-se ao do *retificador de meia-onda (RMO)*, sendo que as características desta retificação não são óbvias. Existem diversas propostas de modelos sendo alguns deles caracterizados por uma retificação “suave” e outros baseados em funções exponenciais [16].

Neste trabalho optou-se por utilizar o mesmo tipo de retificador do modelo da

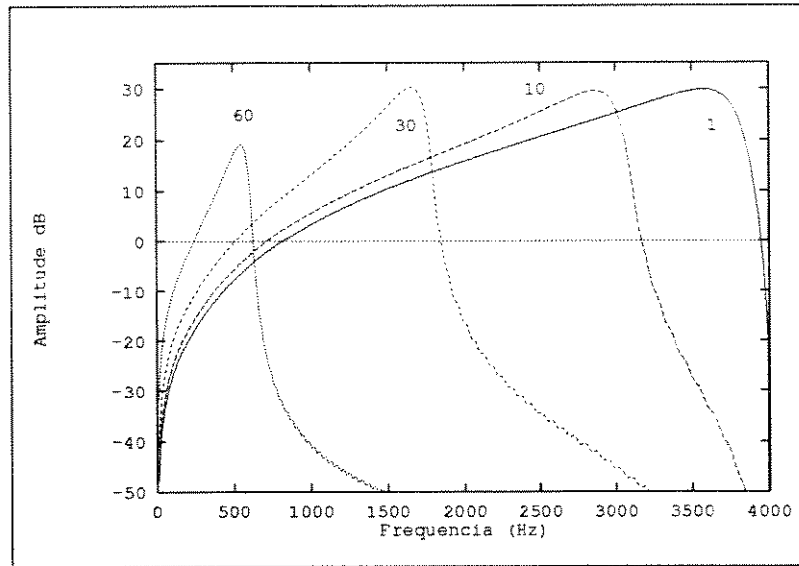


Figura 3.15: Resposta em frequência de quatro canais do modelo da cóclea, considerando os efeitos adicionados pelo estágio de pré-ênfase, para uma frequência de amostragem de 8kHz.

cóclea utilizado por R. Lyon em [16], que consiste de um retificador de meia-onda ideal mais simples de ser implementado e analisado. Outra vantagem com relação aos outros retificadores é que este não apresenta variação de “ganho” com a amplitude do sinal. O retificador ideal é implementado pela seguinte equação,

$$RMO(x) = \begin{cases} 0, & \text{se } x \leq 0 \\ x, & \text{se } x > 0. \end{cases} \quad (3.18)$$

A retificação fornece sinais positivos que serão processados pelos controladores automáticos de ganho da camada de compressão.

### 3.4 Compressão e Adaptação na Cóclea

A compressão realizada pela cóclea nos sinais sonoros é uma das características mais marcantes e de difícil implementação deste órgão. A capacidade da cóclea de comprimir uma faixa de cerca de 12 ordens de grandeza, entre o limiar da audição e o limiar da dor, em uma faixa de cerca de quatro ordens de grandeza [16] torna a implementação computacional ou elétrica uma tarefa complexa.

Esta compressão no ouvido é realizada em parte pelas *células ciliadas externas*, que na cóclea aparentemente se comportam como músculos que devolvem parte da energia para a membrana basilar, possuindo as características de um *controlador automático de ganho* (CAG) [19]. No entanto tais controladores não são capazes de uma compressão desta ordem de grandeza sem provocar algum tipo de distorção no sinal [16].

Esta ordem de grandeza pode ser obtida, entretanto, combinando-se várias camadas de CAGs com diferentes parâmetros. Sendo os primeiros mais lentos e com alvos<sup>5</sup> maiores e os seguintes gradativamente mais rápidos e com alvos menores. Tal combinação foi sugerida por R. Lyon em [18].

Outra característica introduzida nestas camadas é a capacidade de mascarar determinados canais. Na cóclea, quando uma determinada célula ciliada interna responde, há uma tendência em se atenuar as células vizinhas realçando o sinal da primeira, este fenômeno é observado principalmente nos sensores ópticos da retina no olho. Esta “competição” é implementada no modelo computacional acoplando-se os CAGs de cada banco com seus vizinhos mais próximos [18]. Com este acoplamento cada CAG afetará todos os demais do mesmo banco, entretanto o efeito cairá exponencialmente com a distância [29].

### 3.4.1 Parâmetros e Implementação dos CAGs

Cada CAG de cada banco é implementado segundo o diagrama de blocos da figura 3.16, onde nota-se o acoplamento entre os CAG do mesmo banco. O comportamento do CAG é dado pelas seguintes equações,

$$y_i(k) = [1 - s_i(k)] \cdot x_i(k) \quad (3.19)$$

$$s_i(k) = \frac{1 - \epsilon}{3} [s_{i-1}(k-1) + s_i(k-1) + s_{i+1}(k-1)] + \frac{\epsilon}{a} y_i(k-1), \quad (3.20)$$

onde  $y_i(k)$  é a saída do CAG do  $i$ -ésimo canal no instante de tempo  $k$ ,  $x_i(k)$  é a entrada do CAG,  $s_i(k)$  é o estado do CAG,  $s_{i-1}(k)$  o estado do canal imediatamente à esquerda e  $s_{i+1}(k)$  o estado do canal imediatamente a direita. Assume-se neste trabalho que  $s_0(k) = s_1(k)$  e  $s_{N+1}(k) = s_N(k)$ , como condições de contorno para o primeiro e último canais [29]. O parâmetro  $a$  é o alvo do CAG e  $\epsilon$  é dado por

$$\epsilon = 1 - \exp\left(\frac{-1}{\tau \cdot F_a}\right),$$

onde  $\tau$  é a constante de tempo do CAG e  $F_a$  a frequência de amostragem em Hertz.

<sup>5</sup>Alvo corresponde ao valor da saída do CAG quando a entrada tende a infinito,  $x \rightarrow \infty$ .

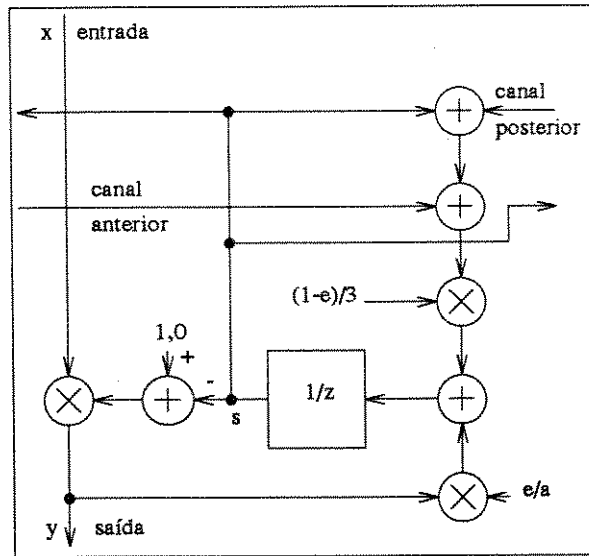


Figura 3.16: Esquema da implementação de um Controle Automático de Ganho da cóclea [18].

Em regime permanente, o valor de  $s_i(k)$  fornecido pela equação (3.20), para uma entrada constante  $x_i(k)$  em todos os canais, será igual a saída  $y_i(k)$  dividida pelo alvo  $a$  do CAG. A saída  $y_i(k)$  em situação de regime permanente poderá ser obtida substituindo-se

$$s_i(k) = \frac{y_i(k)}{a}$$

na equação (3.19), fornecendo a função característica do CAG quando este se encontra em regime,

$$y_i(k) = \frac{a \cdot x_i(k)}{a + x_i(k)}. \tag{3.21}$$

Assumindo  $a = 1$  pode-se notar pela figura 3.17 que o limite para valores grandes de entrada é igual ao alvo  $a$ . Outra característica deste CAG é que a saída do CAG é escalável, isto é, as respostas do mesmo são similares, havendo mudança apenas na escala, como pode ser visto na figura 3.18 com  $a = 0,1$ .

No entanto, a equação (3.20) apresenta o inconveniente de tornar o CAG instável quando a entrada torna-se excessivamente grande. Malcolm Slaney em [29] introduziu uma não-linearidade na equação para compensar este efeito e estabilizar o CAG evitando oscilações indesejadas. A modificação consiste em substituir  $s_i$  por  $s_i/(1 + s_i)$  na equação (3.19). Esta mudança evita que  $s_i$  torne-se grande e previne que o mesmo vá a zero rapi-

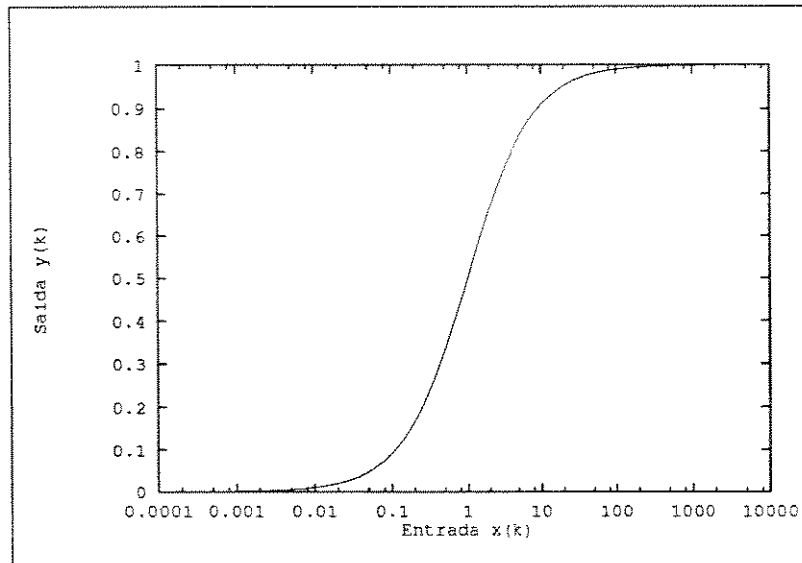


Figura 3.17: Função característica da saída com relação à entrada do CAG para o alvo  $a = 1$ .

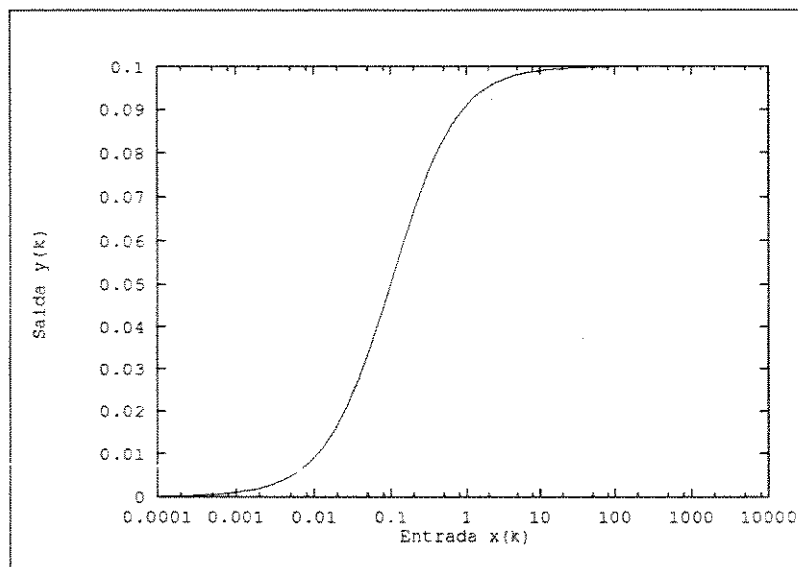


Figura 3.18: Função característica da saída com relação à entrada do CAG para o alvo  $a = 0,1$ .

Parâmetros dos Bancos CAG		
Banco	Constante de Tempo $\tau$	Alvo ( $\times 10^{-4}$ )
1	640ms	32
2	160ms	16
3	40ms	8
4	10ms	4

Tabela 3.1: Parâmetros experimentais dos bancos CAG do modelo computacional da cóclea.

damente. A equação (3.19) torna-se então,

$$y_i(k) = \left[ 1 - \frac{s_i(k)}{1 + s_i(k)} \right] \cdot x_i(k). \quad (3.22)$$

Utilizando a equação (3.22) a saída não mais será limitada pelo alvo  $a$ . Entretanto o fato desta equação eliminar as oscilações indesejadas e retirar a instabilidade provocada por um valor excessivamente grande da entrada  $x_i(k)$  faz desta equação uma boa opção uma vez que são mantidas as características da compressão dos sinais realizada pela cóclea. Neste trabalho optou-se por utilizar as equações (3.22) e (3.20) nos modelos dos CAGs.

Utiliza-se neste modelo quatro bancos de CAGs cada vez mais rápidos e com alvos gradativamente menores para modelar as características de compressão da cóclea. Richard F. Lyon em [18] determinou experimentalmente os parâmetros destes quatro bancos de CAGs que são utilizados no modelo deste trabalho, estes parâmetros estão listados na tabela 3.1. O primeiro banco de CAGs tem como entrada a saída do banco de retificadores e a saída de cada CAG alimenta o próximo banco conforme a esquematização da figura 3.19.

### 3.5 Adições ao Modelo

Richard Lyon em [17] adiciona ainda mais 3 camadas. Uma correspondente a diferença entre as saídas de dois canais e outras duas que correspondem a um Filtro/Dizimador. O modelo completo de um canal pode ser visto na figura 3.20.

A camada de diferença efetua a subtração da saída do canal  $i$  pelo canal  $i - 1$  eliminando os efeitos da fase e provoca um aumento na resposta de dois canais adjacentes na proporção da diferença de fase entre eles. O canal deverá passar por um retificador de meia-onda novamente para eliminar os valores negativos. Esta diferença pode ser interpretada como o efeito de um neurônio excitador-inibidor em um dos núcleos de nervos do sistema



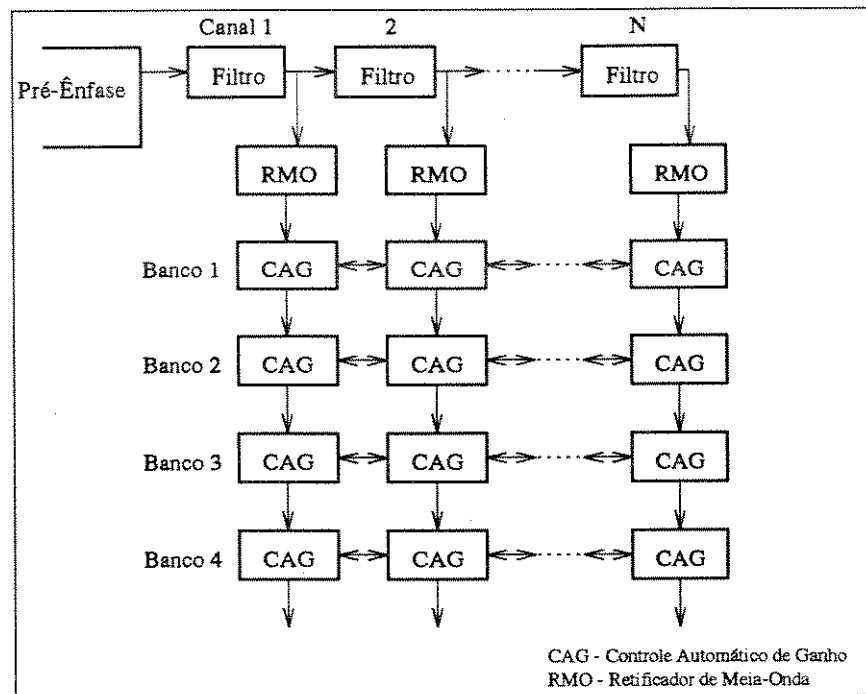


Figura 3.19: Esquema da implementação dos bancos de controle automático de ganho do modelo da cóclea.

nervoso auditivo. A diferença provoca uma melhora significativa das saídas do modelo da cóclea com relação às saídas das camadas dos CAG [18].

Os dois blocos de filtragem/dizimação deveriam eliminar as frequências altas (filtragem) e reamostrar o sinal a frequências sucessivamente mais baixas (dizimador), mas, devido a implementação em computador digital, os dois blocos foram mantidos apenas para as implementações futuras deste modelo sendo que ambos apenas efetuam a função de um filtro de primeira ordem, passa-baixa, com frequência de corte em torno de 30Hz [29] quando a taxa de amostragem é de 8kHz.

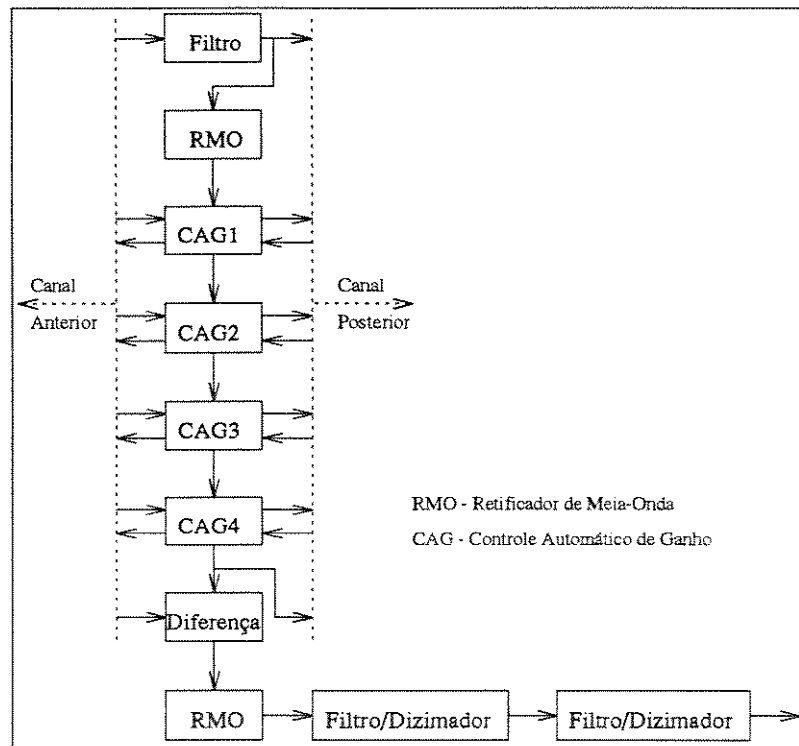


Figura 3.20: Diagrama completo dos blocos funcionais de um canal qualquer da cóclea [18].

## Capítulo 4

# Redes Neurais e Reconhecimento de Padrões

Modelos de redes neurais artificiais têm sido estudados com o intuito de se obter sistemas computacionais com o mesmo desempenho, em reconhecimento de padrões, encontrado em sistemas nervosos biológicos. Tenta-se atingir este desempenho através de uma densa interconectividade entre elementos computacionais simples que operam em paralelo. Estes modelos de redes inspiram-se no atual conhecimento de sistemas nervosos biológicos e sobre o próprio cérebro.

Este tipo de metodologia teve avanço significativo graças a estudos sobre reconhecimento de voz e imagens, onde as redes provam-se eficientes. Para se ter uma boa classificação de voz e imagem é necessário, dados os padrões a serem reconhecidos, testar um grande número de possibilidades sobre o padrão analisado com relação a determinados padrões de referência. Ao invés de um programa com instruções seqüenciais como em um computador de von Neumann, os modelos de redes neurais exploram várias possibilidades concorrentes simultaneamente usando redes altamente paralelas compostas por vários elementos computacionais ligados entre si por conexões com pesos variáveis.

Os elementos computacionais, aqui chamados simplesmente de *neurônios*, correspondem aos nós da rede e são modelos simplificados dos neurônios biológicos. Estes modelos foram obtidos a partir de estudos realizados sobre a geração e propagação de impulsos elétricos pela membrana celular dos neurônios [7]. Os neurônios artificiais usados nos modelos de redes neurais são não-lineares, tipicamente analógicos e podem ser lentos

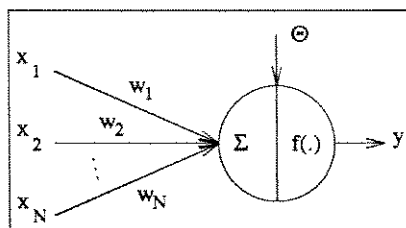


Figura 4.1: Simbologia de um neurônio artificial, sua função de transferência é dada pela equação (4.1).

comparados com os circuitos digitais atuais.

O modelo de neurônio mais simples e que engloba as principais características de uma rede neural biológica, paralelismo e alta conectividade, foi o proposto por McCulloch e Pitts [13, 26]. Este modelo efetua a soma algébrica ponderada das entradas de um neurônio e passa este resultado através de uma função não linear, sendo o neurônio caracterizado, ainda, por um limiar interno ( $\Theta$ ). A figura 4.1 mostra a simbologia usada para um neurônio com  $N$  entradas, com a saída dada pela equação

$$y = f \left( \sum_{i=1}^N x_i w_i + \Theta \right), \quad (4.1)$$

onde  $x_i$  corresponde a uma das entradas,  $w_i$  corresponde ao peso dado à esta entrada,  $\Theta$  ao limiar do neurônio e  $f(\cdot)$  é, geralmente, uma função não linear que define a característica digital/analgica do neurônio. Por exemplo, para neurônios digitais usa-se a função *sinal*,

$$f(x) = \text{sinal}(x) = \begin{cases} -1, & \text{se } x < 0 \\ +1, & \text{se } x \geq 0 \end{cases} \quad (4.2)$$

e para implementações analógicas, a função *sigmóide*,

$$f(x) = \text{sigmóide}(x) = \frac{1}{1 + \exp(-x)} \quad (4.3)$$

usa-se também a função *tangente hiperbólica*<sup>1</sup> para este fim. A equação (4.3) e a função tangente hiperbólica são as mais usadas nos algoritmos de aprendizagem disponíveis. Para ilustração, as funções estão representadas na figura 4.2.

A rede neural é caracterizada pelo tipo de neurônio, pela sua topologia e pelo algoritmo de aprendizagem utilizados. Este algoritmo especifica as condições iniciais da rede (valores iniciais dos pesos e limiares) e a maneira como os pesos serão ajustados durante o

<sup>1</sup>Tem comportamento similar à sigmóide, porém sua saída é limitada à  $[-1, 1]$ .

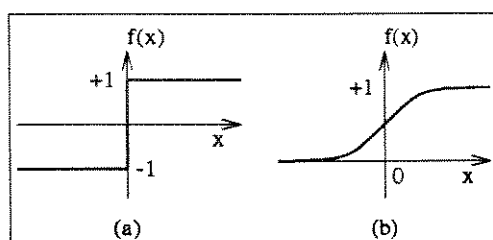


Figura 4.2: Funções de ativação do neurônio, sinal (a) e sigmóide (b).

treino para atingir-se o desempenho desejado. Tanto a topologia da rede quanto o tipo de aprendizagem usada são ainda objetos de estudos.

As redes neurais possuem outras características importantes além da alta capacidade computacional devido ao denso paralelismo. Redes neurais geralmente provém um alto grau de robustez à tolerância a falhas, comparada aos computadores seqüenciais, porque há uma grande quantidade de nós processadores cada um com seu conjunto primário de conexões. Defeitos em alguns dos nós ou nas conexões não irão afetar significativamente o desempenho global da rede [15]. Alguns tipos de rede, também, são capazes de adaptar os pesos das conexões em “tempo real” para otimizar seu desempenho utilizando-se dos resultados correntes. Não é objetivo deste trabalho aprofundar-se em demasia nos diversos tipos de modelos de redes neurais nem na matemática envolvida no assunto.

Com relação aos algoritmos de aprendizagem, ou de ajuste de pesos, estes podem ser classificados em *supervisionados* e *não supervisionados*. O primeiro tipo depende de interferência externa para o ajuste dos pesos e o segundo decide por si só. Como este trabalho necessita de uma resposta precisa da rede aos padrões de entrada serão usados algoritmos supervisionados.

A seguir é estudada uma rede do tipo *perceptron* simples.

## 4.1 Rede “Perceptron” Simples

A *rede perceptron simples* caracteriza-se por apresentar apenas uma camada de neurônios e podem ser usadas tanto com entradas analógicas quanto digitais (binárias). Este tipo de rede gerou muito interesse quando foi inicialmente desenvolvida por sua habilidade em aprender padrões simples [15]. No caso de um único neurônio, figura 4.1, este pode ser descrito apenas pelas equações (4.1) e (4.2). Assim sendo, se a saída assumir o valor  $+1,0$  teremos por, exemplo, um padrão de entrada pertencente a uma classe **A** e se o valor for

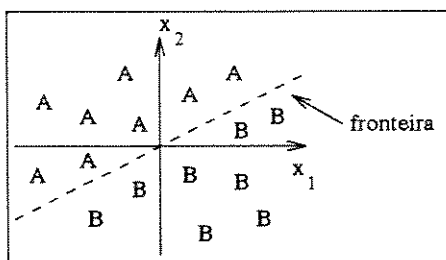


Figura 4.3: Regiões de decisão para um neurônio com duas entradas,  $x_1$  e  $x_2$  são as entradas do neurônio.

$-1,0$ , o padrão pertenceria a uma outra classe **B** ou, simplesmente, não pertenceria a classe **A**. A *fronteira de decisão* será dada, então, por

$$\sum_{i=1}^N x_i w_i + \Theta = 0 \tag{4.4}$$

que corresponde a equação de um hiperplano que separa o espaço  $\mathcal{R}^N$  em duas regiões. Estas regiões resultam na classificação das entradas como pertencente a uma classe ou outra e serão chamadas de *regiões de decisão*.

Para o caso bidimensional, a fronteira será uma reta separando o plano de entrada em duas regiões de decisão, conforme mostrado na figura 4.3, sendo a fronteira definida pela equação

$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{\Theta}{w_2}. \tag{4.5}$$

Como pode-se perceber pela figura 4.3, uma rede perceptron simples terá um bom desempenho se as classes de padrões puderem ser separadas por um hiperplano, ou seja, se as regiões de decisão puderem ser dispostas em lados opostos de algum hiperplano. No caso de classes de padrões que não podem ser separadas, ou aquelas cuja distribuição espacial se sobrepõe a alguma outra classe, a rede perceptron simples não alcançará um desempenho satisfatório. Pode-se notar ainda que, uma vez definido o modelo matemático do neurônio, o “conhecimento” adquirido pelo neurônio estará representado pelos pesos dados às conexões e pelo limiar.

Os pesos e o limiar em um neurônio podem ser ajustados por um grande número de métodos, no entanto Rosenblatt em [15, 26] desenvolveu uma metodologia básica para a aprendizagem em uma rede perceptron, denominado *algoritmo de convergência perceptron* ou *regra delta*.

O algoritmo consiste dos seguintes passos:

1. inicializar os pesos  $w_i$  e  $\Theta$  com valores numéricos aleatórios pequenos;
2. adquirir os valores das entradas  $\mathbf{x} = \{x_1, x_2, \dots, x_N\}^T$  e a saída desejada para estas entradas  $d(\mathbf{x})$ ;
3. calcular a saída da rede através das equações (4.1) e (4.3)

$$y(t) = \text{sinal} \left( \sum_{i=1}^N w_i(t)x_i(t) + \Theta(t) \right);$$

4. adaptar os pesos e limiar usando as equações

$$\begin{aligned} w_i(t+1) &= w_i(t) + \eta[d(t) - y(t)]x_i(t), \quad i = 1, 2, \dots, N \quad \text{e} \\ \Theta(t+1) &= \Theta(t) + \eta[d(t) - y(t)] \quad \text{e} \end{aligned}$$

5. repetir os passos 2, 3 e 4 até que o erro  $\epsilon(t) = d(t) - y(t)$  atinja valores desejados.

Este algoritmo inclui o parâmetro  $\eta$ , denominado *taxa de adaptação*, que corresponde a um ganho positivo tal que  $\eta \in \mathcal{R}^+$ . A ele deve ser atribuído um valor que faça com que o algoritmo obtenha a convergência o mais rapidamente possível não provocando mudanças excessivamente bruscas nos pesos e limiar e nem provocando oscilações nos mesmos. Infelizmente não há maneira de se obter um valor ótimo para  $\eta$  devendo este ser obtido empiricamente.

O desempenho deste tipo de rede aumenta consideravelmente com a construção de estruturas mais complexas formadas por camadas de neurônios, a qual chama-se de redes neurais “perceptron” multicamadas. A seguir será mostrada uma topologia comum deste tipo de rede e seu comportamento.

## 4.2 Simplificação do Modelo do Neurônio Artificial

Cabe aqui uma pequena modificação no neurônio de McCulloch e Pitts, figura 4.1 e equação (4.1). Esta modificação consiste em acrescentar uma entrada falsa ao neurônio correspondente a um neurônio cuja saída seja sempre igual a 1 e cujo peso seja igual a  $\Theta$ . Deste modo, com

$$\begin{aligned} x_0 &= 1 \\ w_0 &= \Theta \end{aligned}$$

a equação (4.1) reduz-se a

$$y = f\left(\sum_{i=0}^N w_i x_i\right), \quad x_0 = 1. \quad (4.6)$$

Com esse tipo de modificação torna-se mais simples tanto a análise da rede multicamadas “perceptron” quanto o desenvolvimento dos algoritmos computacionais, pois o argumento da função  $f(\cdot)$  passa a ser representado como um produto interno de dois vetores. A simbologia para esta modificação continuará a mesma da figura 4.1 eliminando-se a representação do limiar ( $\Theta$ ) o que torna mais simples a diagramação da rede.

### 4.3 Redes Neurais “Perceptron” Multicamadas

As redes “perceptron” multicamadas caracterizam-se pela associação de neurônios artificiais de tal forma que eles se dispõem em camadas seguindo algumas regras básicas:

- um neurônio de uma camada inferior faz conexões com todos os neurônios da camada superior;
- neurônios de uma mesma camada não se interconectam e
- a camada mais inferior corresponderá a entrada e a mais superior a saída da rede.

A qualquer camada entre a camada de saída e a de entrada denomina-se *camada escondida*.

Esquemáticamente, para uma rede genérica de  $M$  camadas tem-se o diagrama da figura 4.4, note que os limiares não estão sendo mostrados.

Neste trabalho não será colocada a camada de entrada como uma camada composta de neurônios, em princípio porque esta camada geralmente seria formada por *células sensoriais* que apenas apresentariam os sinais à rede, cuidando, portanto, da transdução do sinal externo para sinais da rede. Eventualmente haveria algum tipo de pré-processamento nesta camada, tal como uma normalização dos dados, por exemplo.

A saída desta rede é dada por  $\mathbf{x}_M = \{x_{M1}, x_{M2}, \dots, x_{MN_M}\}$  e do diagrama tem-se:  $x_{ki}$  como a saída do  $i$ -ésimo neurônio da camada  $k$  e  $w_{kij}$  como o peso da conexão do  $i$ -ésimo neurônio da camada  $k$  com o  $j$ -ésimo neurônio da camada  $k - 1$ . Conforme foi definido na seção 4.2:  $w_{ki0}$  é o limiar do  $i$ -ésimo neurônio da camada  $k$  e a saída dos neurônios “sempre-ativos”  $x_{k0}$  é sempre 1.  $N_k$  é o número de elementos computacionais na



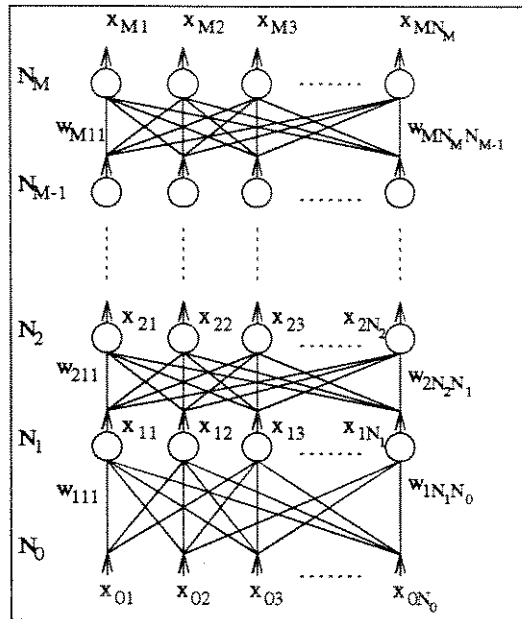


Figura 4.4: Estrutura de uma rede neural “perceptron” multi-camadas genérica.

camada  $k$ . Em todos os casos  $k = 1, 2, \dots, M$ . Do diagrama,  $\mathbf{x}_0 = \{x_{01}, x_{02}, \dots, x_{0N_0}\}$  é o sinal de entrada apresentado à rede e  $x_{00} = 1$ .

Assim, para se propagar um sinal  $\mathbf{p} = \{p_1, p_2, \dots, p_{N_0}\}$  deve-se aplicar a equação (4.6) recursivamente. Sendo as entradas da rede dadas por

$$\begin{aligned} x_{0i} &= p_i, \quad i = 1, 2, \dots, N_0, \\ x_{00} &= 1 \end{aligned}$$

e usando a equação (4.6) com uma pequena modificação, a fim de torná-la recursiva,

$$\begin{aligned} x_{ki} &= f_{ki} \left( \sum_{j=0}^{N_{k-1}} w_{kij} x_{k-1,j} \right) \\ x_{k0} &= 1, \quad i = 0, 1, \dots, N_k, \quad k = 1, \dots, M \end{aligned} \tag{4.7}$$

obtêm-se a saída  $\mathbf{x}_M = \{x_{M1}, x_{M2}, \dots, x_{MN_M}\}$  correspondente ao sinal  $\mathbf{p}$ .

Esta topologia é capaz de compor os diversos hiperplanos que cada um de seus neurônios gera, eliminando-se os inconvenientes da rede perceptron de camada simples. No entanto, até algum tempo atrás não existiam algoritmos de aprendizagem capazes de adaptarem os pesos dos neurônios das camadas escondidas. Este tipo de rede começou a ser usada após o desenvolvimento de um algoritmo denominado *backpropagation* baseado na técnica do gradiente, utilizada em problemas de otimização [6, 26].

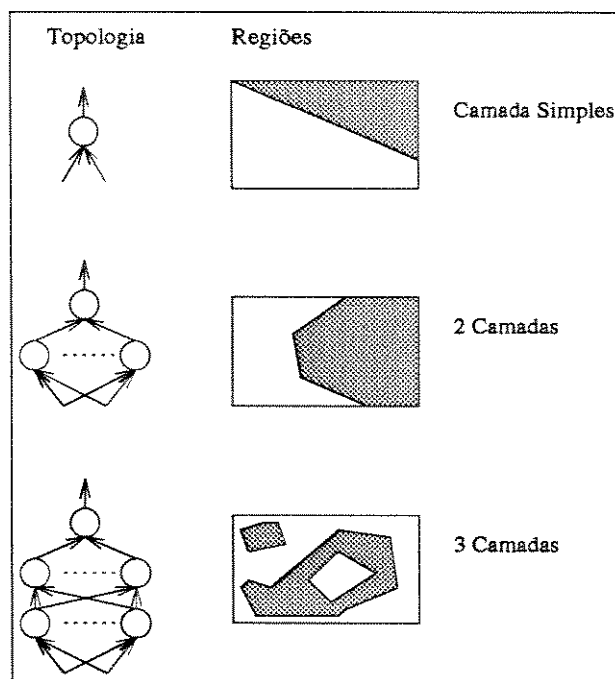


Figura 4.5: Tipos de regiões de decisão para diversas topologias de redes perceptron multicamadas.

Esta rede compõe vários hiperplanos no espaço de entrada o que fornece uma gama maior de regiões possíveis. Lippmann em [15] mostra os tipos de regiões que algumas estruturas de redes multicamadas podem gerar, estes tipos de regiões estão descritas na figura 4.5, para o caso de 2 entradas que corresponde a um espaço de entrada  $\mathcal{R}^2$ . Nesta mesma figura os neurônios têm a função de ativação,  $\text{sin}(x)$ , descrita pela equação (4.2).

Vê-se pela figura 4.5 que uma rede perceptron simples forma as regiões de decisão separando-as através de um único hiperplano e que uma rede perceptron de duas camadas forma regiões convexas quaisquer, podendo ser regiões abertas, no espaço formado pelas entradas. A rede de duas camadas, no entanto, não é capaz de formar regiões desconectadas no espaço formado pelas entradas. A complexidade da fronteira que divide as regiões de decisão da rede de duas camadas depende do número de neurônios da camada escondida [15].

Uma rede de 3 camadas pode formar arbitrariamente qualquer tipo de região de decisão no espaço formado pelas entradas, incluindo regiões desconectadas (figura 4.5). De fato, prova-se que uma rede perceptron de 3 camadas pode gerar qualquer tipo arbitrário de regiões de decisão [15]. A complexidade e números das regiões formadas por uma rede de

3 camadas depende do número de neurônios existentes nas duas camadas escondidas [15].

Usando uma função não linear do tipo sigmóide, equação (4.3), as fronteiras entre uma região de decisão e outra terão uma transição mais suave que aquela mostrada na figura 4.5. No entanto, esse tipo de função de ativação possibilitou o desenvolvimento de algoritmos de aprendizagem, entre eles o “backpropagation” [26] e mais recente os algoritmos derivados do filtro adaptativo de Kalman [1, 11, 27].

## 4.4 Algoritmos de Aprendizagem

É desejável que a rede responda de uma certa maneira a um determinado padrão de entrada  $\mathbf{p}$  ou, de outra forma, uma vez apresentado este padrão à camada de entrada, a rede deverá fornecer uma saída  $\mathbf{x}_M$  que é aquela escolhida para este mesmo padrão.

Os pesos e limiares,  $w_{ijk}$ , desta rede deverão ser ajustados a fim de se obter, para o padrão  $\mathbf{p}$ , a saída  $\mathbf{x}_M$ . No entanto, é extremamente difícil ajustar estes pesos de forma a obter um desempenho desejável sem se utilizar de algoritmos de aprendizagem. Neste caso, necessitam-se de algoritmos que minimizem o erro entre a saída desejada e aquela que a rede fornece para o padrão  $\mathbf{p}$ .

Tendo em vista estas necessidades, desenvolveram-se alguns métodos para se adaptar os pesos. Dois destes algoritmos serão descritos a seguir.

### 4.4.1 “Backpropagation”

Este algoritmo de aprendizagem é uma generalização e adaptação do algoritmo *mínimos quadrados* (LMS — Least Mean Square). Desenvolvido por Rumelhart, Hinton e Williams em [26], o algoritmo caracteriza-se por usar o método do gradiente para minimização do erro quadrático entre a saída fornecida pela rede e a saída realmente desejada.

Uma das exigências do algoritmo é que a função de ativação  $f(\cdot)$  deve ser diferenciável e monotônica-crescente ( $0 \leq f'(\cdot) < \infty$ ). Geralmente usam-se funções do tipo sigmóide, equação (4.3), ou tangente hiperbólica. O desenvolvimento desta regra de aprendizagem [26] será apresentado a seguir.

Para um padrão  $\mathbf{p} = \{p_1, p_2, \dots, p_{N_0}\}$ , apresentado à entrada da rede, deseja-se uma saída  $\mathbf{d} = \{d_1, d_2, \dots, d_{N_M}\}$ . Como condição inicial, os pesos das camadas  $w_{kij}$ ,  $k = 1, 2, \dots, M$ ,  $i = 0, 1, \dots, N_k$  e  $j = 0, 1, \dots, N_{k-1}$  serão inicializados com valores aleatórios pequenos. Pelas condições iniciais, a saída da rede,  $\mathbf{x}_M = \{x_{M1}, \dots, x_{MN_M}\}$ , durante a

apresentação do padrão não corresponderá a desejada  $\mathbf{d}$ . Assim define-se o erro quadrático entre a saída  $\mathbf{x}_M$  e a saída desejada  $\mathbf{d}$  como sendo

$$\varepsilon_{\mathbf{p}} = \frac{1}{2} \sum_{i=1}^{N_M} (d_i - x_{Mi})^2. \quad (4.8)$$

A medida do erro com relação a um conjunto de padrões  $\mathcal{D} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_Q\}$  pode ser dada por

$$\varepsilon = \sum_{i=1}^Q \varepsilon_{\mathbf{p}_i}. \quad (4.9)$$

Ainda, para o desenvolvimento desta regra, desmembra-se a equação (4.7) em,

$$x_{ki} = f_{ki}(a_{ki}) \quad (4.10)$$

$$a_{ki} = \sum_{j=0}^{N_{k-1}} w_{kij} x_{k-1,j} \quad (4.11)$$

$i = 0, 1, \dots, N_k$  e  $k = 1, 2, \dots, M$ .

Os pesos são ajustados a fim de minimizar  $\varepsilon_{\mathbf{p}}$  para o padrão  $\mathbf{p}$ . A quantidade a ser ajustada para cada peso será proporcional a variação que este peso provoca em  $\varepsilon_{\mathbf{p}}$ ,

$$\Delta_{\mathbf{p}} w_{kij} \propto -\frac{\partial \varepsilon_{\mathbf{p}}}{\partial w_{kij}}. \quad (4.12)$$

Para calcular-se esta derivada, usa-se a regra da cadeia repetidas vezes. É útil, a princípio, separar esta derivada em duas partes, uma relacionando a variação do erro como função da entrada de um neurônio ( $a_{ki}$ ) e outra representando o efeito da variação de um peso em particular na entrada deste neurônio ( $a_{ki}$ ). Então pode-se escrever,

$$\frac{\partial \varepsilon_{\mathbf{p}}}{\partial w_{kij}} = \frac{\partial \varepsilon_{\mathbf{p}}}{\partial a_{ki}} \frac{\partial a_{ki}}{\partial w_{kij}}. \quad (4.13)$$

Pela equação (4.11) observa-se que o segundo fator é

$$\frac{\partial a_{ki}}{\partial w_{kij}} = \frac{\partial}{\partial w_{kij}} \sum_{h=0}^{N_{k-1}} w_{kih} x_{k-1,h} = x_{k-1,j}. \quad (4.14)$$

Definindo-se  $\delta_{ki}$  como sendo,

$$\delta_{ki} = -\frac{\partial \varepsilon_{\mathbf{p}}}{\partial a_{ki}}, \quad (4.15)$$

a equação (4.13) assumirá a forma,

$$-\frac{\partial \varepsilon_{\mathbf{p}}}{\partial w_{kij}} = \delta_{ki} x_{k-1,j}. \quad (4.16)$$

Comparando-se com a equação (4.12), pode-se afirmar que o ajuste no peso será proporcional a  $\delta_{ki}$  e a  $x_{k-1,j}$ ,

$$\Delta_{\mathbf{p}} w_{kij} = \eta \delta_{ki} x_{k-1,j}, \quad (4.17)$$

similar à regra delta apresentada na seção 4.1. Deve-se, então, encontrar os valores de  $\delta_{ik}$  para cada neurônio da rede. Um fato interessante nesta regra é que vai-se propagar  $\delta_{ki}$  pela rede no sentido da saída para entrada, contrário ao fluxo normal dos dados (*backpropagation*).

Para o cômputo dos valores de  $\delta_{ki}$ , aplica-se a regra da cadeia em (4.15), separando-a em dois fatores, um relacionando as variações no erro como função da saída e a outra relacionando as variações na saída como função das variações na entrada ( $a_{ki}$ ). Tem-se, então,

$$\frac{\partial \varepsilon_{\mathbf{p}}}{\partial a_{ki}} = \frac{\partial \varepsilon_{\mathbf{p}}}{\partial x_{ki}} \frac{\partial x_{ki}}{\partial a_{ki}}. \quad (4.18)$$

Vê-se que o segundo fator pode ser obtido pela equação (4.10), ou

$$\frac{\partial x_{ki}}{\partial a_{ki}} = f'_{ki}(a_{ki}), \quad (4.19)$$

que é simplesmente a derivada de  $f(\cdot)$  da  $i$ -ésima unidade da camada  $k$  cujo argumento é a função de entrada, equação (4.11), deste elemento. Consideram-se dois casos para encontrar o primeiro fator de (4.18), um para os neurônios da camada de saída e outro para os neurônios das camadas escondidas.

Assumindo-se que o neurônio pertença à camada de saída, neste caso, segue da definição de  $\varepsilon_{\mathbf{p}}$ , equação (4.8), que

$$\frac{\partial \varepsilon_{\mathbf{p}}}{\partial x_{Mi}} = \frac{\partial}{\partial x_{Mi}} \left\{ \frac{1}{2} \left[ \sum_{i=0}^{N_M} (d_i - x_{Mi})^2 \right] \right\} = -(d_i - x_{Mi}) \quad (4.20)$$

que é o mesmo resultado da regra delta. Substituindo-se os dois fatores da equação (4.18) têm-se

$$\delta_{Mi} = (d_i - x_{Mi}) f'_{Mi}(a_{Mi}), \quad i = 0, 1, \dots, N_M, \quad (4.21)$$

para toda unidade de saída.

No entanto, se o neurônio não pertencer à camada de saída, usa-se a regra da cadeia para escrever que

$$\frac{\partial \varepsilon_{\mathbf{p}}}{\partial x_{ki}} = \sum_{n=0}^{N_{k+1}} \frac{\partial \varepsilon_{\mathbf{p}}}{\partial a_{k+1,n}} \frac{\partial a_{k+1,n}}{\partial x_{ki}}. \quad (4.22)$$

Esta equação considera todos os efeitos dos erros das unidades da camada superior em uma unidade de uma camada que não a de saída. O segundo fator de (4.22) é dado pela equação (4.11), logo,

$$\frac{\partial a_{k+1,n}}{\partial x_{ki}} = \frac{\partial}{\partial x_{ki}} \left( \sum_{j=0}^{N_k} w_{k+1,n,j} x_{kj} \right) = w_{k+1,n,i} \quad (4.23)$$

e o primeiro fator é dado pela própria definição de  $\delta_{ki}$ , (4.15),

$$\frac{\partial \varepsilon_{\mathbf{P}}}{\partial a_{k+1,n}} = -\delta_{k+1,n}. \quad (4.24)$$

Substituindo (4.23) e (4.24) em (4.22) tem-se,

$$\frac{\partial \varepsilon_{\mathbf{P}}}{\partial x_{ki}} = - \sum_{n=0}^{N_{k+1}} \delta_{k+1,n} w_{k+1,n,i}, \quad (4.25)$$

que substitui-se em (4.18) para obter,

$$\delta_{ki} = f'_{ki}(a_{ki}) \sum_{n=0}^{N_{k+1}} \delta_{k+1,n} w_{k+1,n,i}, \quad i = 0, 1, \dots, N_k \quad (4.26)$$

para todos os neurônios não pertencentes à camada de saída, isto é,  $k \neq M$ .

As equações (4.21) e (4.26) fornecem um procedimento recursivo para se obter os valores de  $\delta$  de todos os neurônios da rede. Com os valores de  $\delta$  encontrados pode-se ajustar os pesos utilizando-se da equação (4.17). Costuma-se acrescentar um componente de amortecimento à equação (4.17) para retirar os efeitos de uma variação brusca dos pesos da rede, este componente é na forma de

$$\Delta_{\mathbf{P}} w_{kij}(t + \Delta t) = \eta \delta_{ki} x_{k-1,j} + \alpha \Delta_{\mathbf{P}} w_{kij}(t). \quad (4.27)$$

Os valores de  $\eta$  e  $\alpha$  devem ser escolhidos segundo o critério da seção 4.1, ou seja, ambos devem ser escolhidos para se obter a mais rápida convergência possível sem, contudo, provocar oscilações nos pesos. A escolha destes dois parâmetros é experimental.

Para uma melhor apresentação, um resumo completo deste algoritmo pode ser encontrado na tabela 4.1.

#### 4.4.2 Mínimos Quadrados Recursivo

Este método origina-se a partir da teoria de filtros adaptativos, mais especificamente da teoria do filtro de Kalman [6]. Neste método deve-se separar o neurônio em duas partes distintas, uma não-linear dada pela equação (4.10),

$$x_{ki} = f_{ki}(a_{ki})$$

“Backpropagation”	
1	Inicializar os pesos com valores aleatórios pequenos.
2	Escolher um padrão $\mathbf{p}_q$ do conjunto de padrões de entrada e propagá-lo até a saída usando a equação (4.7).
3	Calcular o erro $\varepsilon_{\mathbf{p}_q}$ pela equação (4.8), se o erro estiver acima do desejado, ajustar os pesos recursivamente usando as equações (4.21), (4.26) e (4.27).
4	Repetir 2 e 3 até que todos os erros $\varepsilon_{\mathbf{p}_q}$ estejam abaixo de um erro $\mathcal{E}$ máximo admissível.

Tabela 4.1: Resumo do algoritmo de aprendizagem.

e em outra linear dada pela equação (4.11),

$$a_{ki} = \sum_{j=0}^{N_{k-1}} w_{kij} x_{k-1,j}$$

Com isso o problema de ajuste de pesos torna-se linear como pode ser visto na equação (4.11). Esta linearização do neurônio é possível pois dada uma saída desejada qualquer  $\mathbf{d}$  encontra-se um valor  $\tilde{a}_{Mi}$ , tal que

$$\tilde{a}_{Mi} = f_{Mi}^{-1}(d_i) \tag{4.28}$$

e deste modo os pesos dos neurônios da camada de saída seriam adaptados em função de  $\tilde{a}_{Mi}$ .

Como ocorreu no desenvolvimento do algoritmo “backpropagation”, o problema restringe-se em obter as estimativas para as saídas dos neurônios das camadas escondidas,  $\tilde{x}_{ki}$ . Mas, pela equação (4.10),  $\tilde{x}_{ki} = f_{ki}(\tilde{a}_{ki})$ , o que torna necessária apenas a estimativa de  $\tilde{a}_{ki}$  [27].

Como é desejado obter uma estimativa  $\tilde{a}_{ki}$  de modo a minimizar o erro provocado por  $a_{ki}$  na saída da rede pode-se dizer que a diferença entre  $\tilde{a}_{ki}$  e  $a_{ki}$  seja proporcional à variação que  $a_{ki}$  provoca no erro, isto é,

$$\tilde{a}_{ki} - a_{ki} \propto -\frac{\partial \varepsilon_{\mathbf{p}}}{\partial a_{ki}},$$

onde  $\varepsilon_{\mathbf{p}}$  é o erro quadrático definido anteriormente no algoritmo “backpropagation”. Pela definição de  $\delta_{ki}$ , equação (4.15),

$$\delta_{ki} = -\frac{\partial \varepsilon_{\mathbf{p}}}{\partial a_{ki}},$$

a estimativa de  $a_{ki}$  poderá ser obtida através do cálculo de  $\delta_{ki}$  para os neurônios das camadas escondidas pela seguinte equação:

$$\bar{a}_{ki} = a_{ki} + \mu_{ki}, \quad (4.29)$$

onde  $\mu_{ki}$  é um ganho positivo qualquer ( $\mu_{ki} \in \mathcal{R}^+$ ) e  $\delta_{ki}$  pode ser obtido recursivamente pelas equações (4.21) e (4.26) já comentadas no desenvolvimento do algoritmo “backpropagation”. É claro que usando uma estimativa para as camadas escondidas, o sistema de equações lineares formado por (4.11) não irá gerar valores exatos para  $w_{kij}$ .

Neste método introduz-se um *fator de esquecimento* na equação (4.8) com a finalidade de reduzir a influência dos padrões mais antigos no cômputo dos pesos [1, 6, 11, 27]. Os padrões serão apresentados à rede em determinados intervalos de tempo, dando uma característica temporal ao algoritmo.

Para facilitar o desenvolvimento do algoritmo, consideraremos apenas um neurônio da camada  $k$ , sendo o erro total para este neurônio dado por

$$\varepsilon_k(t) = \sum_{n=1}^t \lambda_k(t, n) |e_k(t)|^2, \quad (4.30)$$

onde  $e(t)$  é o erro entre a saída  $a$  e a saída desejada  $\bar{a}$ , e é dado por

$$e_k(t) = \bar{a}_k(t) - a_k(t), \quad (4.31)$$

mas pela equação (4.11), para um único neurônio,<sup>2</sup>

$$e_k(t) = \bar{a}_k(t) - \sum_{j=0}^{N_{k-1}} w_{kj}(t) x_{k-1,j}(t). \quad (4.32)$$

O fator de esquecimento  $\lambda_k(t, n)$  na equação (4.30) deve ter a seguinte propriedade

$$0 < \lambda(t, n) \leq 1, \quad n = 1, 2, \dots, t. \quad (4.33)$$

A forma mais usada para  $\lambda$  é o fator exponencial definido por

$$\lambda(t, n) = \beta^{t-n}, \quad (4.34)$$

onde  $\beta$  é uma constante aproximadamente igual a 1 tal que,  $0 < \beta \leq 1$ .  $\beta$  é, a grosso modo, a medida da “memória” do algoritmo. Neste método, deve-se minimizar o seguinte índice de desempenho

$$\varepsilon_k(t) = \sum_{n=1}^t \beta_k^{t-n} |e_k(t)|^2. \quad (4.35)$$

---

<sup>2</sup> $w_{kj}$  corresponde a uma linha da matriz  $\mathbf{W}_k = \{w_{kij}\}$



Os valores ótimos para  $w_{kij}(t)$  que minimizam a equação (4.35) são dados por

$$\frac{\partial \varepsilon_k(t)}{\partial w_{kh}(t)} = 0, \quad k = 1, 2, \dots, M, \quad h = 1, 2, \dots, N_{k-1}. \quad (4.36)$$

Então, derivando-se a equação (4.35) em relação a  $w_{kh}(t)$ , tem-se

$$\frac{\partial}{\partial w_{kh}(t)} \left\{ \sum_{n=1}^t \beta_k^{t-n} |e_k(t)|^2 \right\} = 2 \sum_{n=1}^t \beta_k^{t-n} e_k(t) \frac{\partial e_k(t)}{\partial w_{kh}(t)}, \quad (4.37)$$

a derivada  $\frac{\partial e_k(t)}{\partial w_{kh}(t)}$  pode ser obtida diretamente de (4.32),

$$\frac{\partial \varepsilon_k(t)}{\partial w_{kh}(t)} = -2 \sum_{n=1}^t \beta_k^{t-n} \left( \tilde{a}_k(t) - \sum_{j=0}^{N_{k-1}} w_{kj}(t) x_{k-1,j}(t) \right) x_{k-1,h}. \quad (4.38)$$

Como é desejado  $\frac{\partial \varepsilon_k(t)}{\partial w_{kh}(t)} = 0$ , pode-se arrumar o lado direito de (4.38) da seguinte maneira,

$$\sum_{n=1}^t \beta_k^{t-n} \tilde{a}_k(t) x_{k-1,h}(t) = \underbrace{\sum_{n=1}^t \beta_k^{t-n} \sum_{j=0}^{N_{k-1}} w_{kj}(t) x_{k-1,j}(t) x_{k-1,h}(t)}_{(*)}. \quad (4.39)$$

Reescrevendo o lado direito de (4.39) desta maneira,

$$\sum_{n=1}^t \beta_k^{t-n} \tilde{a}_k(t) x_{k-1,h}(t) = \sum_{n=1}^t \beta_k^{t-n} x_{k-1,h}(t) \underbrace{\sum_{j=0}^{N_{k-1}} w_{kj}(t) x_{k-1,j}(t)}_{(**)}. \quad (4.40)$$

e escrevendo o somatório (\*\*) como um produto de vetores (produto interno), tem-se

$$\sum_{n=1}^t \beta_k^{t-n} x_{k-1,h}(t) \overbrace{\mathbf{w}_k^T(t) \mathbf{x}_{k-1}(t)}^{(**)} = \sum_{n=1}^t \beta_k^{t-n} x_{k-1,h}(t) \mathbf{x}_{k-1}^T(t) \mathbf{w}_k(t), \quad (4.41)$$

$$h = 0, \dots, N_{k-1}, \quad k = 1, \dots, M.$$

Definindo

$$\mathbf{R}_k(t) = \sum_{n=1}^t \beta_k^{t-n} \mathbf{x}_{k-1}(t) \mathbf{x}_{k-1}^T(t) \quad (4.42)$$

e

$$\mathbf{p}_k(t) = \sum_{n=1}^t \beta_k^{t-n} \tilde{a}_k(t) \mathbf{x}_{k-1}(t) \quad (4.43)$$

pode-se, finalmente, escrever a equação (4.39) como

$$\mathbf{p}_k(t) = \mathbf{R}_k(t)\mathbf{w}_k(t). \quad (4.44)$$

$\mathbf{R}_k$  pode ser interpretado como uma matriz de correlação do conjunto de padrões a serem aprendidos e o vetor  $\mathbf{p}_k$  como a correlação entre padrões a serem treinados e as respostas desejadas.

O vetor de pesos  $\mathbf{w}_k(t)$  pode ser encontrado usando técnicas de resolução de sistemas de equações lineares que levariam a

$$\mathbf{w}_k(t) = \mathbf{R}_k^{-1}(t)\mathbf{p}_k(t). \quad (4.45)$$

Deseja-se, no entanto, um método recursivo [6] para a resolução da equação (4.45). Isolando o termo correspondente a  $n = t$  na equação (4.42) do somatório, chega-se a

$$\mathbf{R}_k(t) = \beta_k \left[ \sum_{n=1}^{t-1} \beta^{t-1-n} \mathbf{x}_{k-1}(n)\mathbf{x}_{k-1}^T(n) \right] + \mathbf{x}_{k-1}(t)\mathbf{x}_{k-1}^T(t). \quad (4.46)$$

Entretanto, a expressão entre colchetes na equação (4.46) é, pela definição (4.42), igual a  $\mathbf{R}(t-1)$ . Então, encontra-se a seguinte expressão para o ajuste de  $\mathbf{R}_k$  recursivamente,

$$\mathbf{R}_k(t) = \beta_k \mathbf{R}_k(t-1) + \mathbf{x}_{k-1}(t)\mathbf{x}_{k-1}^T(t). \quad (4.47)$$

Similarmente, a partir da equação (4.43), pode-se obter uma forma recursiva para o cálculo de  $\mathbf{p}_k(t)$ ,

$$\mathbf{p}_k(t) = \beta_k \mathbf{p}_k(t-1) + a_k(t)\mathbf{x}_{k-1}(t). \quad (4.48)$$

Para ajustar os pesos a partir da equação (4.45) deve-se determinar a inversa da matriz de correlação  $\mathbf{R}_k(t)$ . Na prática isso seria inviável, pois torna-se necessário calcular  $\mathbf{R}_k^{-1}(t)$  para cada apresentação de um padrão  $\mathbf{p}_t$ ,  $t = 1, 2, \dots, \infty$ . No entanto pode-se chegar a um método recursivo para o cálculo de  $\mathbf{R}_k^{-1}(t)$  através do *lema de inversão de matrizes*.

**Lema 1** *Sejam  $\mathbf{A}$  e  $\mathbf{B}$  duas matrizes  $M \times M$  definidas-positivas relacionadas por*

$$\mathbf{A} = \mathbf{B}^{-1} + \mathbf{C}\mathbf{D}^{-1}\mathbf{C}^T, \quad (4.49)$$

*onde  $\mathbf{D}$  é uma matriz  $N \times N$  definida-positiva e  $\mathbf{C}$  é uma matriz  $M \times N$  quaisquer. De acordo com o lema de inversão de matrizes, pode-se expressar a inversa da matriz  $\mathbf{A}$  por*

$$\mathbf{A}^{-1} = \mathbf{B} - \mathbf{B}\mathbf{C} \left( \mathbf{D} + \mathbf{C}^T\mathbf{B}\mathbf{C} \right)^{-1} \mathbf{C}^T\mathbf{B}. \quad (4.50)$$

A prova deste lema é obtida facilmente, multiplicando-se as equações (4.49) e (4.50) para chegar à matriz identidade  $\mathbf{I}$ .

Como a matriz  $\mathbf{R}_k(t)$  é definida-positiva e possivelmente não-singular, pode-se aplicar o lema de inversão de matrizes na equação recursiva (4.47), com  $\mathbf{A} = \mathbf{R}_k(t)$ ,  $\mathbf{B}^{-1} = -\beta_k \mathbf{R}_k(t-1)$ ,  $\mathbf{C} = \mathbf{x}_{k-1}(t)$  e  $\mathbf{D} = 1$ , substituindo-os diretamente em (4.50) tem-se a seguinte equação recursiva para a inversa da matriz de correlação,

$$\mathbf{R}_k^{-1}(t) = \beta_k^{-1} \mathbf{R}_k^{-1}(t-1) - \frac{\beta_k^{-2} \mathbf{R}_k^{-1}(t-1) \mathbf{x}_{k-1}(t) \mathbf{x}_{k-1}^T(t) \mathbf{R}_k^{-1}(t-1)}{1 + \beta_k^{-1} \mathbf{x}_{k-1}^T(t) \mathbf{R}_k^{-1}(t) \mathbf{x}_{k-1}(t)}. \quad (4.51)$$

Para facilidade de cálculo, seja

$$\mathbf{Q}_k(t) = \mathbf{R}_k^{-1}(t) \quad (4.52)$$

e

$$\mathbf{k}_k(t) = \frac{\beta_k^{-1} \mathbf{Q}_k(t-1) \mathbf{x}_{k-1}(t)}{1 + \beta_k^{-1} \mathbf{Q}_k(t-1) \mathbf{x}_{k-1}(t)}. \quad (4.53)$$

Usando estas definições, reescreve-se (4.51) deste modo,

$$\mathbf{Q}_k(t) = \beta_k^{-1} \mathbf{Q}_k(t-1) - \beta_k^{-1} \mathbf{k}_k(t) \mathbf{x}_{k-1}^T(t) \mathbf{Q}_k(t-1). \quad (4.54)$$

Nota-se que  $\mathbf{Q}_k(t)$  é uma matriz  $N_{k-1} \times N_{k-1}$  e que  $\mathbf{k}_k(t)$  é um vetor  $N_{k-1} \times 1$  sendo comumente chamado de *vetor de ganho* do filtro de Kalman.

Rearranjando (4.53) tem-se,

$$\begin{aligned} \mathbf{k}_k(t) &= \beta_k^{-1} \mathbf{Q}_k(t-1) \mathbf{x}_{k-1}(t) - \beta_k^{k-1} \mathbf{k}_k(t) \mathbf{x}_{k-1}^T(t) \mathbf{Q}_k(t-1) \mathbf{x}_{k-1}(t) \\ &= \left[ \beta_k^{-1} \mathbf{Q}_k(t-1) - \beta_k^{-1} \mathbf{k}_k(t) \mathbf{x}_{k-1}^T(t) \mathbf{Q}_k(t-1) \right] \mathbf{x}_{k-1}(t). \end{aligned} \quad (4.55)$$

Vê-se de (4.54) que a expressão entre colchetes do lado direito de (4.55) é igual a  $\mathbf{Q}_k(t)$ . Assim, pode-se simplificar (4.55) até

$$\mathbf{k}_k(t) = \mathbf{Q}_k(t) \mathbf{x}_{k-1}(t). \quad (4.56)$$

Obtêm-se uma equação recursiva para o ajuste dos pesos  $\mathbf{w}_k$ , estimado pelo método dos mínimo quadrados. Para tal, usam-se as equações (4.44), (4.48) e (4.52) obtendo

$$\begin{aligned} \mathbf{w}_k(t) &= \mathbf{R}_k^{-1}(t) \mathbf{p}_k(t) \\ &= \mathbf{Q}_k(t) \mathbf{p}_k(t) \\ &= \beta_k \mathbf{Q}_k(t) \mathbf{p}_k(t) + \mathbf{Q}_k(t) \mathbf{x}_{k-1}(t) \tilde{a}_k(t). \end{aligned} \quad (4.57)$$

Substituindo a equação (4.54) somente no primeiro termo  $\mathbf{Q}_k(t)$  do lado direito de (4.57) tem-se

$$\begin{aligned}
 \mathbf{w}_k(t) &= \mathbf{Q}_k(t-1)\mathbf{p}_k(t-1) - \mathbf{k}_k(t)\mathbf{x}_{k-1}^T(t)\mathbf{Q}_k(t-1)\mathbf{p}_k(t-1) + \\
 &\quad \mathbf{Q}_k(t)\mathbf{x}_{k-1}(t)\tilde{a}_k(t) \\
 &= \mathbf{R}_k^{-1}(t-1)\mathbf{p}_k(t-1) - \mathbf{k}_k(t)\mathbf{x}_{k-1}^T(t)\mathbf{Q}_k(t-1)\mathbf{p}_k(t-1) + \\
 &\quad \mathbf{Q}_k(t)\mathbf{x}_{k-1}(t)\tilde{a}_k(t) \\
 &= \mathbf{w}_k(t-1) - \mathbf{k}_k(t)\mathbf{x}_{k-1}^T(t)\mathbf{w}_k(t-1) + \mathbf{Q}_k(t)\mathbf{x}_{k-1}(t)\tilde{a}_k(t). \tag{4.58}
 \end{aligned}$$

Finalmente, usando a equação (4.56), obtêm-se a equação recursiva para o ajuste do vetor de pesos,

$$\mathbf{w}_k(t) = \mathbf{w}_k(t-1) + \mathbf{k}_k(t) \left[ \tilde{a}_k(t) - \mathbf{x}_k^T(t)\mathbf{w}_k(t-1) \right]. \tag{4.59}$$

Vê-se que o termo  $\mathbf{x}_k^T(t)\mathbf{w}_k(t-1)$  é a saída linear do neurônio, antes dos pesos serem modificados, ou a saída corrente do neurônio, portanto, para a camada de saída,

$$\mathbf{w}_M(t) = \mathbf{w}_M(t-1) + \mathbf{k}_k(t) [\tilde{a}_k(t) - a_k(t)], \tag{4.60}$$

e para os neurônios das camadas escondidas,  $\tilde{a}_k(t)$  é obtido diretamente das equações (4.60) e (4.29),

$$\mathbf{w}_k(t) = \mathbf{w}_k(t-1) + \mathbf{k}_k(t) [\mu_k \delta_{ki}] \tag{4.61}$$

sendo  $\delta_{ki}$  obtido recursivamente a partir das equações (4.21) e (4.26), do algoritmo de “backpropagation.”

Para uma melhor apresentação, um resumo completo deste algoritmo pode ser encontrado na tabela 4.2.

## 4.5 Exemplos de Redes

Como exemplo de utilização das redes neurais, implementa-se nessa sessão uma rede capaz de emular a função lógica *XOR*. Esta função lógica binária exige redes neurais de pelo menos duas camadas de neurônios para emulá-la, pois uma rede neural com uma única camada de neurônios não conseguirá criar as regiões de decisões necessárias para a função *XOR*, como pôde ser visto anteriormente. Esta função lógica pode ser considerada

Mínimos Quadrados Recursivo	
1	Inicializar os pesos com valores aleatórios pequenos, inicializar $\mathbf{Q}_k(0)$ . Esta inicialização é feita, geralmente, com $\mathbf{Q}_k(0) = \xi_k \mathbf{I}$ , onde $\xi_k$ é um número grande e $\mathbf{I}$ é a matriz identidade.
2	Escolher um padrão a ser ensinado e apresentá-lo à rede, a entrada é $\mathbf{x}_0(t)$ e a saída desejada $\mathbf{d}(t)$ .
3	Propagar o padrão através da rede, equações (4.10) e (4.11).
4	Calcular $\mathbf{Q}_k(t)$ e $\mathbf{p}_k(t)$ usando as equações (4.53) e (4.54) para as camadas $k = 1, 2, \dots, M$ .
5	Propagar o sinal de erro $\delta_{ki}$ , $k = 1, \dots, M$ , $i = 0, \dots, N_M$ , usando as equações (4.21) e (4.26).
6	Encontrar a saída linear desejada $\tilde{a}_{ki}$ , $k = 1, 2, \dots, M$ , $i = 0, 1, \dots, N_M$ , usando a equação (4.28).
7	Adaptar os pesos usando as equações (4.60) e (4.61).
8	Testar o erro quadrático médio, equação (4.8), se ainda não foi atingido o valor desejado, recomeçar pelo item 2.

Tabela 4.2: Resumo do algoritmo de aprendizagem.

Função Lógica XOR		
Entrada $x_1$	Entrada $x_2$	Saída $n_3$
-1,0	-1,0	-1,0
-1,0	1,0	1,0
1,0	-1,0	1,0
1,0	1,0	-1,0

Tabela 4.3: Tabela de padrões de entrada e saída para a rede emuladora da função XOR.

um exemplo clássico para as redes neurais, principalmente devido ao fato de que uma rede “perceptron” simples é incapaz de emulá-la.

Para a implementação desta rede neural, somente serão necessários três neurônios, um na camada de saída e dois na camada escondida, conforme a figura 4.6. Uma das soluções para este tipo de problema, ajustando-se os pesos diretamente pelo posicionamento das regiões de decisão, é apresentada na figura 4.7, para os valores de entrada e saída dados pela tabela 4.3. Neste caso específico o neurônio  $N1$  emula a função lógica OR,  $N2$  a função NAND e  $N3$  a função AND.

O comportamento da figura 4.7 é dada pelas seguintes equações:

$$y = \text{sinal}(n_1 + n_2 - 1);$$

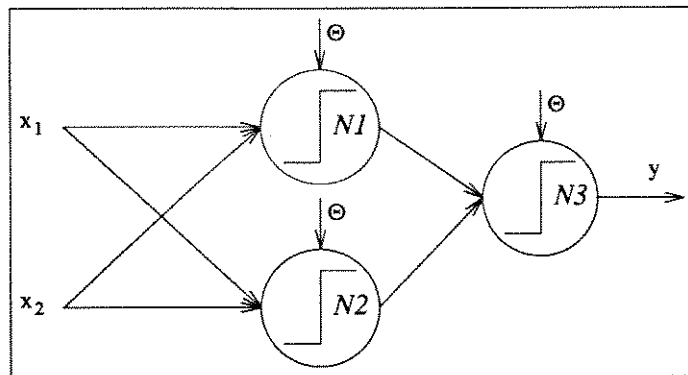


Figura 4.6: Rede Neural para emulação da função lógica XOR.

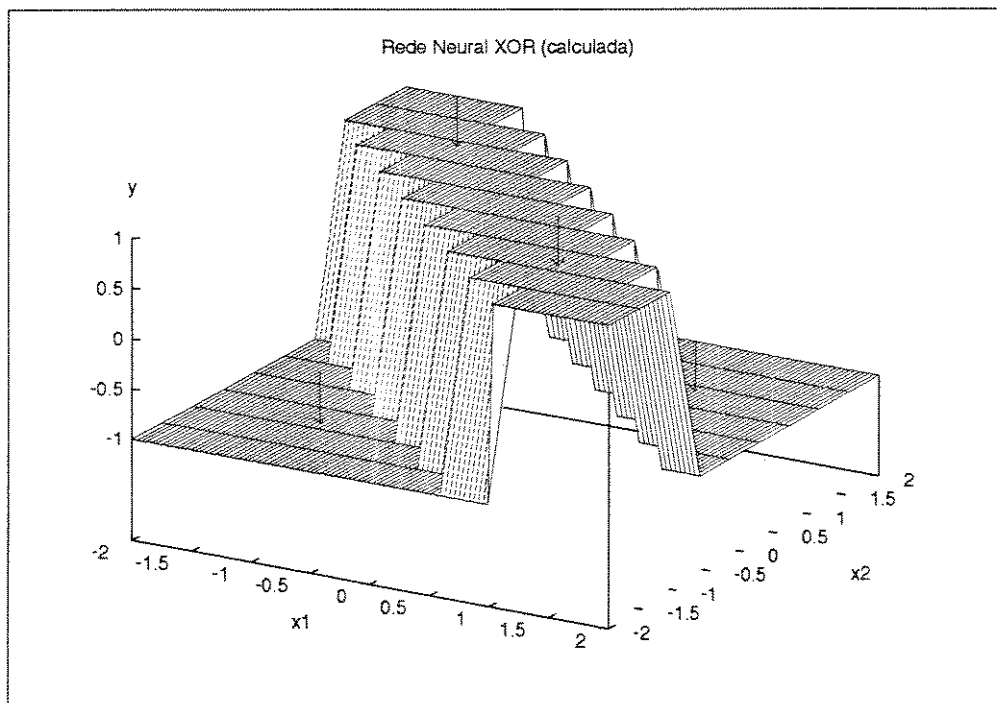


Figura 4.7: Superfície da saída da rede em função das entradas.

$$n_1 = \text{sinal}(x_1 + x_2 + 1);$$

$$n_2 = \text{sinal}(-x_1 - x_2 + 1);$$

onde  $n_1$ ,  $n_2$  e  $y$  são as saídas dos neurônios,  $N1$ ,  $N2$  e  $N3$  (saída), respectivamente, e  $x_1$  e  $x_2$  as entradas da rede, (vide figura 4.6).

Para efeito de comparação, a mesma topologia de rede da figura 4.6 com os mesmos padrões de entrada/saída da tabela 4.3, foi ensinada com os dois métodos descritos na seção 4.4, substituindo-se a função de ativação dos neurônios de  $\text{sinal}(x)$  por  $\tanh(x)$ . Os resultados obtidos serão apresentados a seguir.

Em cada etapa, para ambos os métodos, todos os padrões eram apresentados à rede, verificava-se, então, o erro máximo<sup>3</sup> entre a saída obtida e a desejada e ensinava-se somente os padrões que apresentavam erro máximo maior do que o desejado. Normalmente a cada etapa de apresentação, verificação e aprendizagem dá-se o nome de *época*.

Quando usado o método de ensino “backpropagation”, a melhor convergência foi obtida com  $\mu = 0,1$  e  $\alpha = 0,9$  e com esses valores, conseguiu-se um erro máximo menor que 0,0001 em 43 épocas. O comportamento encontrado por esse método é descrito pelas equações,

$$y = \tanh(-2,25040 \cdot n_1 + 2,31787 \cdot n_2 - 1,675130);$$

$$n_1 = \tanh(-1,23323 \cdot x_1 - 1,12024 \cdot x_2 - 0,623711);$$

$$n_2 = \tanh(-1,42369 \cdot x_1 - 1,17022 \cdot x_2 + 1,224600);$$

onde  $n_1$ ,  $n_2$  e  $y$  são as saídas dos neurônios,  $N1$ ,  $N2$  e  $N3$  (saída), respectivamente, e  $x_1$  e  $x_2$  as entradas da rede. As regiões de saída podem ser visualizadas na figura 4.8.

Quanto ao método de ensino MQR (filtro de kalman), a melhor convergência obtida foi com  $\beta = 0,8$  e  $\mu = 30,0$ . Estes valores levaram à convergência em 13 épocas para o mesmo valor de erro no caso do “backpropagation”. As equações que descrevem a rede obtidas com este método são as seguintes:

$$y = \tanh(1,47171 \cdot n_1 + 1,47179 \cdot n_2 + 1,47178);$$

$$n_1 = \tanh(-10,02490 \cdot x_1 + 12,89710 \cdot x_2 - 11,61500);$$

$$n_2 = \tanh(27,17370 \cdot x_1 - 59,07270 \cdot x_2 - 43,35370);$$

onde  $n_1$ ,  $n_2$  e  $y$  são as saídas dos neurônios,  $N1$ ,  $N2$  e  $N3$  (saída), respectivamente, e  $x_1$  e  $x_2$  as entradas da rede. As regiões de saída podem ser verificadas na figura 4.9.

<sup>3</sup>O erro máximo entre dois vetores  $v = \{v_i\}$  e  $u = \{u_i\}$  de mesma dimensão é dado por:  $\epsilon = \max_i(|v_i - u_i|)$

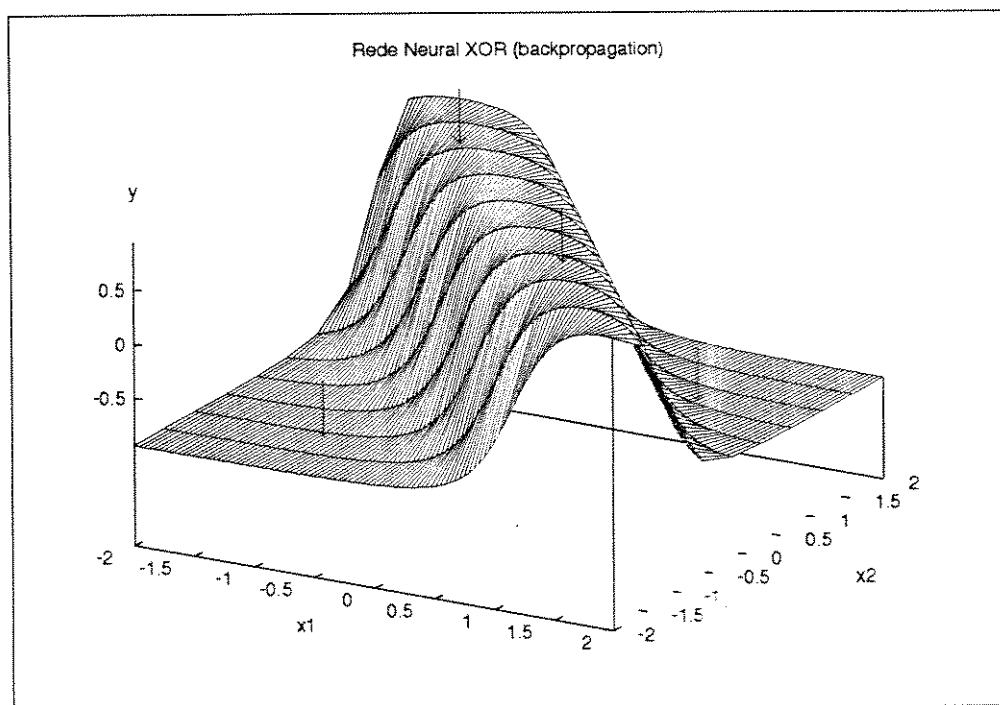


Figura 4.8: Superfície da saída da rede em função das entradas após ensino com o método “backpropagation”.



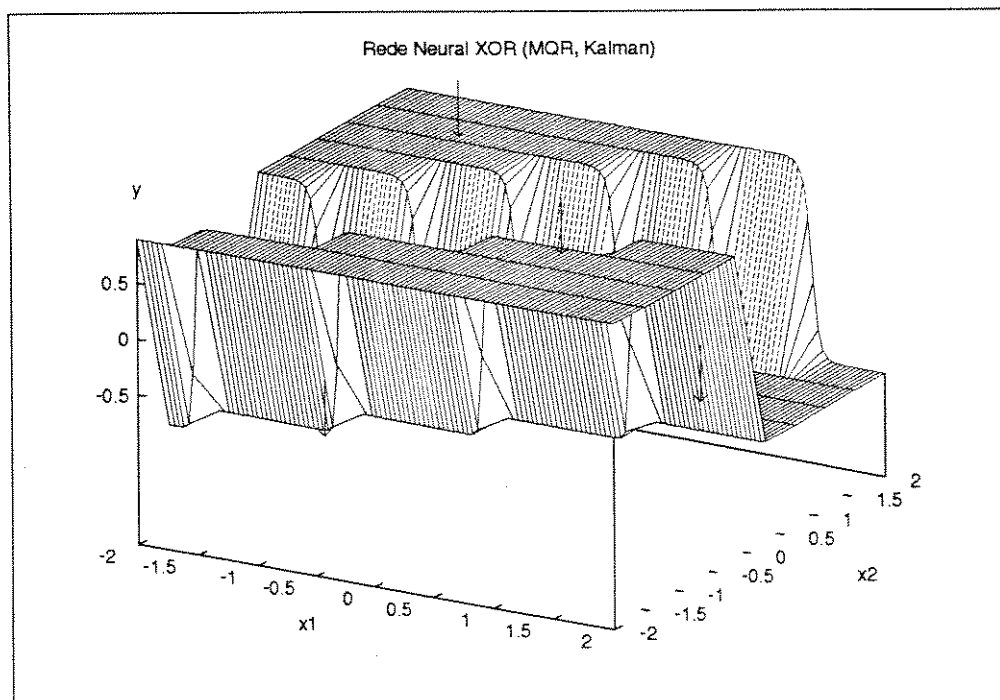


Figura 4.9: Superfície da saída da rede em função das entradas após ensino com o método MQR (filtro de kalman).

Pode-se observar que ambos os métodos conseguem resolver o problema do *ou exclusivo*, sendo que o método “backpropagation” obteve um resultado extremamente semelhante ao calculado através da equação (4.5), enquanto que o método MQR (filtro de kalman) chegou a uma solução bem diferente. Podemos no entanto “forçar” o MQR a obter uma solução parecida inserindo o padrão  $\mathbf{x}^T = [0,0; 0,0]$  com a saída  $y = 1,0$ .

Observa-se que método MQR provoca o crescimento dos pesos da rede, de fato, para um grande número de padrões a serem ensinados, o método provoca uma “sobrecarga” nos pesos provocando instabilidade numérica quando em simulação computacional. Apesar desse inconveniente, para um pequeno número de padrões, este método é o mais eficiente entre os dois apresentados.

## Capítulo 5

# Implementação do Classificador de Fonemas

Neste capítulo será descrito a implementação do *classificador de fonemas* detalhando a metodologia usada na extração dos *padrões de voz* reconhecidos pelas *redes neurais* bem como as topologias usadas para a construção das redes.

A seção 5.1 aborda a extração dos padrões de voz, indicando os principais tipos de fonemas da língua portuguesa. Nesta seção estão indicados os fonemas utilizados como padrões de voz para a aprendizagem e teste das redes neurais. A seção 5.2 mostra as topologias de redes neurais utilizadas neste trabalho.

### 5.1 Extração dos Padrões de Voz

Existe uma grande variedade de fonemas que podem ser produzidos pelo *aparelho fonador humano*, no entanto, estes podem ser classificados em grupos específicos. Estes grupos caracterizam os fonemas por suas similaridades quanto à excitação do aparelho fonador [34].

O aparelho fonador é constituído pelos pulmões, músculo diafragma, cordas vocais, faringe, boca, cavidade nasal, língua, dentes, lábios, úvula<sup>1</sup> e nariz, mostrados esquematicamente na figura 5.1. Com exceção da cavidade nasal, das aberturas do nariz (narinas) e da traquéia, os demais componentes do aparelho fonador são controlados pelo locutor [25, 34]. Este aparelho é objeto de estudos na extração de parâmetros dos sinais de voz e

---

<sup>1</sup>Geralmente conhecida como a “campainha” da boca.

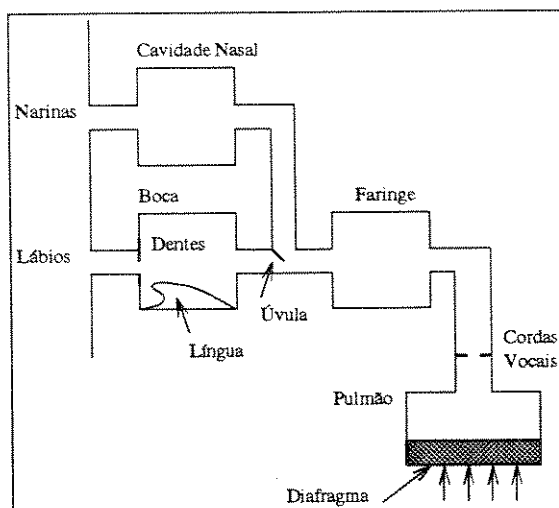


Figura 5.1: Esquemática acústica do trato vocal, [34].

alguns modelos matemáticos foram desenvolvidos para o mesmo com o intuito de facilitar a compreensão dos diferentes tipos de fonemas produzidos [4, 25, 34]. James L. Flanagan cita alguns modelos mecânicos, bastante interessantes, que emulam o sistema fonador humano em [4].

A seguir serão mostrados como são classificados os fonemas da língua portuguesa de acordo com as suas semelhanças quando na produção dos sons e os métodos utilizados neste trabalho para a extração dos padrões de espectro da voz assim como os padrões originados do modelo do ouvido interno.

Antes, no entanto, será explicada a notação simplificada adotada neste trabalho para exprimir os fonemas da língua portuguesa.

### 5.1.1 Comentários sobre Notação utilizada para os Fonemas

Adota-se neste trabalho uma notação simplificada para representar os fonemas do que aquela utilizada pelos profissionais da área. Tal notação consiste em se usar o próprio alfabeto da língua portuguesa mais os acentos gráficos. A representação fonética da palavra será apresentada sempre entre os símbolos /.../. Assim, a palavra *casa* será foneticamente representada por /kaza/.

Neste trabalho, consideram-se as sílabas como sendo fonemas, principalmente devido à dificuldade em se conseguir isolar os fonemas das sílabas, pois não há separação visível entre eles (principalmente em sílabas formadas por consoantes explosivas e vogais).

A separabilidade dos fonemas torna-se mais difícil em trechos da língua falada (sem pausas entre as palavras) devido a aglutinação natural entre as palavras. Como por exemplo em *às vezes* e em *trinta e dois* que tornam-se /ázvezes/ e /trinteidois/.

### 5.1.2 Características dos Fonemas da Língua Portuguesa

Na língua portuguesa há basicamente três formas de excitação do aparelho fonador e portanto três grupos principais de fonemas. Estas excitações resultam nos sons *sonoros* (geralmente vogais), *fricativos* e *explosivos* [34], que serão detalhados nesta seção.

#### Sons Sonoros

O fluxo de ar proveniente dos pulmões é controlado pela abertura e fechamento das *cordas vocais* ou *dobras vocais*, que se assemelham a dois lábios delgados que podem ser aproximados e ter sua rigidez alterada através do controle do locutor. A abertura entre as dobras é denominada *glote*. Estando a glote totalmente fechada, o fluxo ar originário dos pulmões é interrompido, aumentando a pressão sub-glótica até que as cordas vocais sejam separadas, liberando o ar aprisionado e gerando um pulso de ar de curta duração. Com o deslocamento do ar a pressão glótica é reduzida, permitindo a reaproximação das cordas vocais. O processo se repete em uma forma aproximadamente periódica, sendo a frequência média desses pulsos determinante do *período de pitch* [25, 34].

Exemplos típicos de sons sonoros são as vogais, cujo grau de nasalização é determinado pelo abaixamento da úvula. Observa-se nos sinais uma variação “rápida” e quase regular, relativa à excitação e uma envoltória “lenta”, dada pela resposta do trato vocal que modela a excitação. Algumas consoantes, como a lateral /l/ em /lá/ e a nasal /m/ em /má/, são produzidas com a excitação glotal.

Os sons descritos aqui estão agrupados na categoria de sons *vocálicos* ou *ressonantes*, que são caracterizados pela presença da vibração glotal e uma maior concentração de energia na faixa de 200–3 500Hz [34]. As figuras 5.2, 5.3, 5.4, 5.5 e 5.6 mostram o sinal em tempo discreto (amostrado a 8kHz) para as vogais /á/, /é/, /i/, /ó/ e /u/, respectivamente.

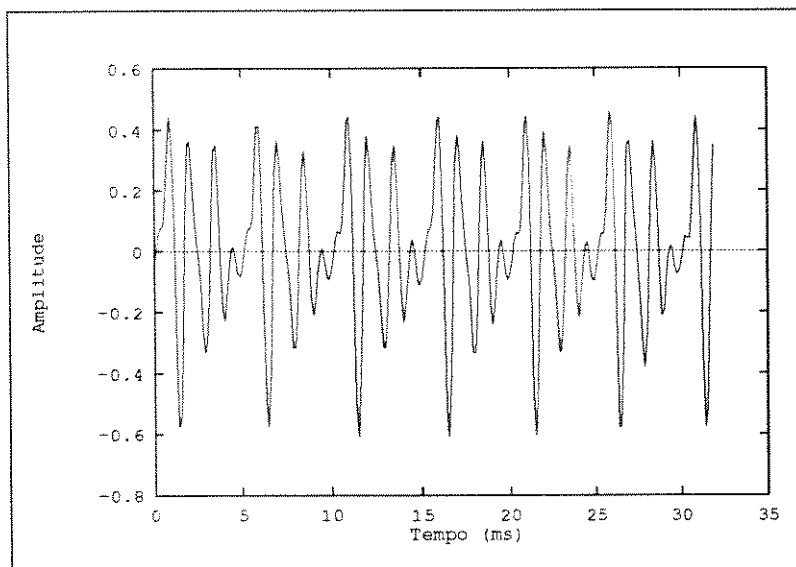


Figura 5.2: Sinal no tempo para um trecho de 32ms da vogal /á/ amostrada a 8kHz.

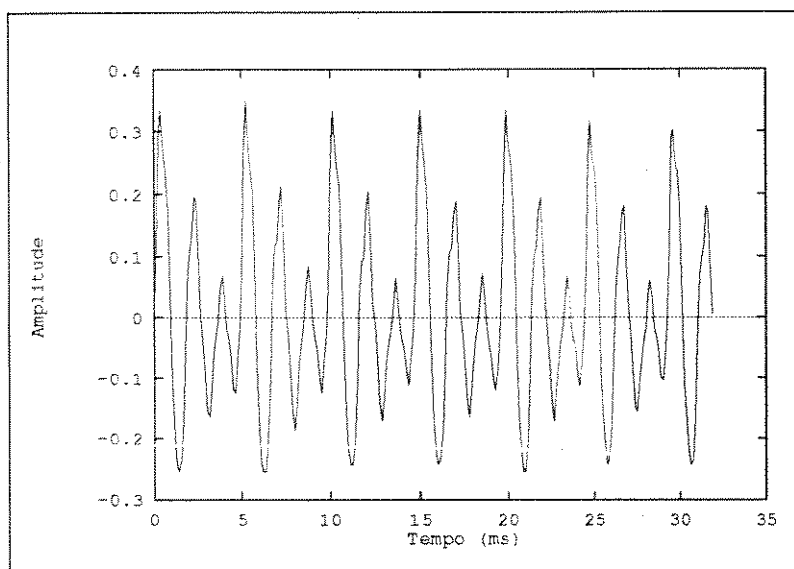


Figura 5.3: Sinal no tempo para um trecho de 32ms da vogal /é/ amostrada a 8kHz.

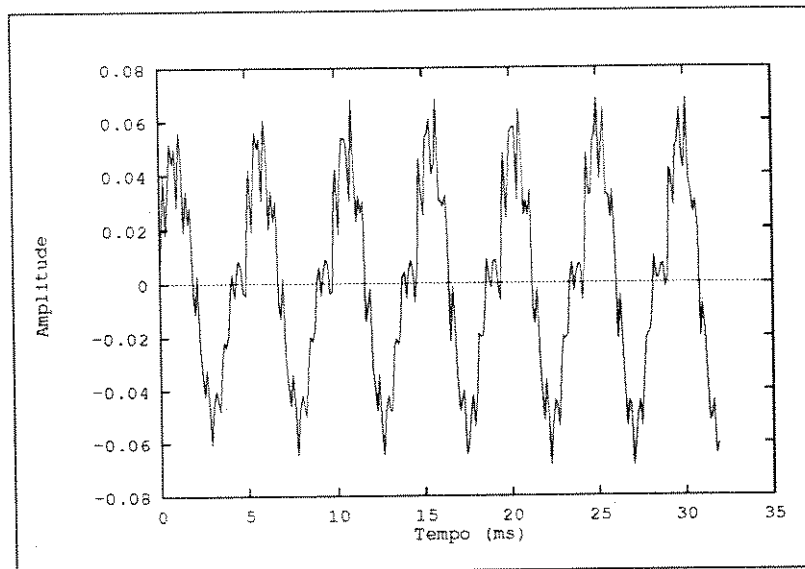


Figura 5.4: Sinal no tempo para um trecho de 32ms da vogal /i/ amostrada a 8kHz.

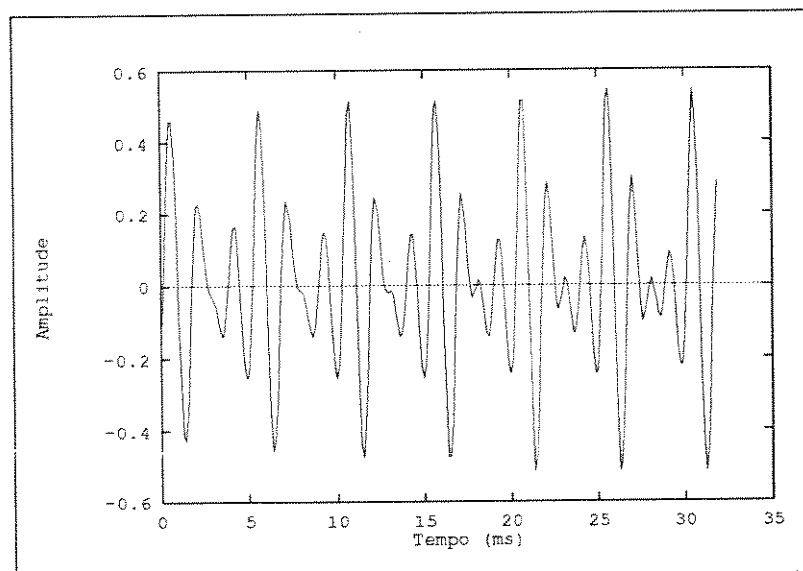


Figura 5.5: Sinal no tempo para um trecho de 32ms da vogal /ó/ amostrada a 8kHz.

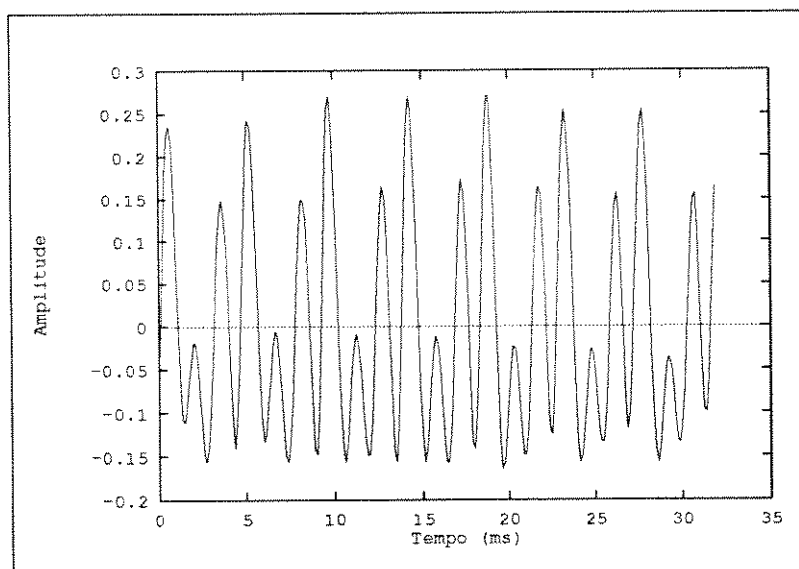


Figura 5.6: Sinal no tempo para um trecho de 32ms da vogal /u/ amostrada a 8kHz.

### Sons Fricativos Surdos

Durante a produção de sons *fricativos surdos*, ou *sibilantes*, a glote permanece aberta, não havendo vibrações nas cordas vocais. Entretanto um estreitamento é realizado em algum ponto do trato vocal. Para o fonema /f/, por exemplo, os lábios e dentes são ligeiramente pressionados resultando em uma passagem estreita para o ar; o fluxo do ar torna-se turbulento nas imediações da constricção, excitando as cavidades do trato vocal. Esta excitação é de baixa intensidade, assemelhando-se ao *ruído branco* [34], geralmente contém maior concentração de energia acima de 2kHz até aproximadamente 10kHz [25]. Outros fonemas pertencentes a esta classe são /s/ e /x/. As figuras 5.7 e 5.8 mostram o sinal em tempo discreto (amostrado a 8kHz) para as sílabas /sá/ e /fá/, respectivamente.

### Sons Explosivos Surdos

A última maneira básica de excitação do sistema fonador é realizada em /p/, /t/ ou /k/, onde o ar é dirigido à boca, que se encontra totalmente fechada. Com o aumento da pressão a oclusão é rompida bruscamente, gerando um pulso que excita o aparelho fonador. A explosão é acompanhada de um movimento rápido dos articuladores em direção à configuração do som seguinte. Esta excitação é denominada, também, de *excitação transitória* ou *sons oclusivos surdos* [34]. As figuras 5.9 e 5.10 mostram o sinal em tempo



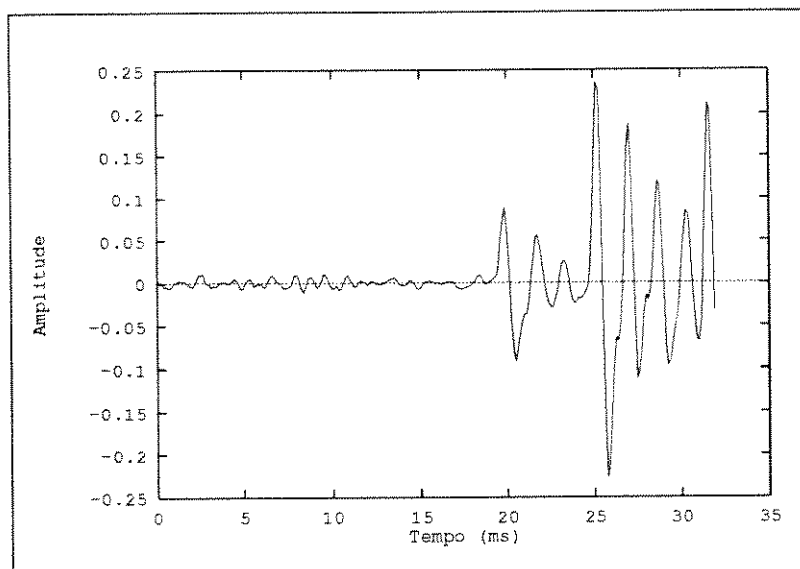


Figura 5.7: Sinal no tempo para o trecho inicial de 32ms da sílaba /sá/ amostrada a 8kHz.

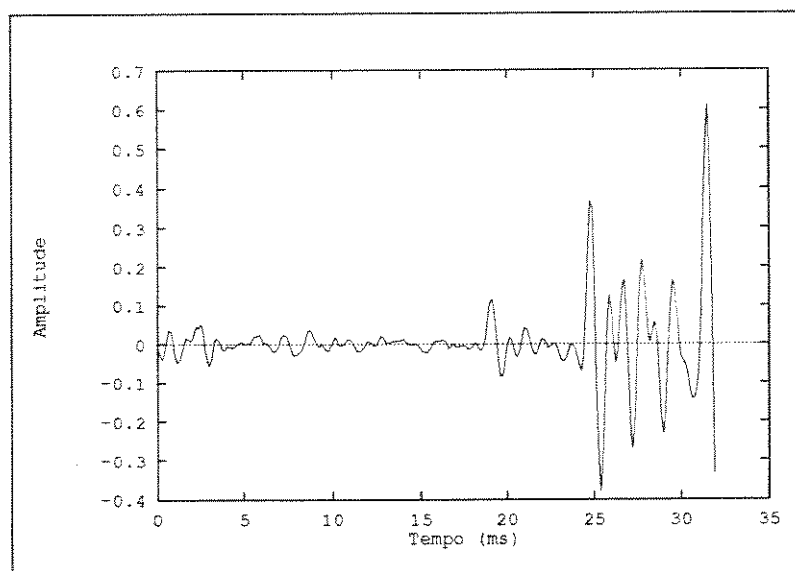


Figura 5.8: Sinal no tempo para o trecho inicial de 32ms da sílaba /fá/ amostrada a 8kHz.

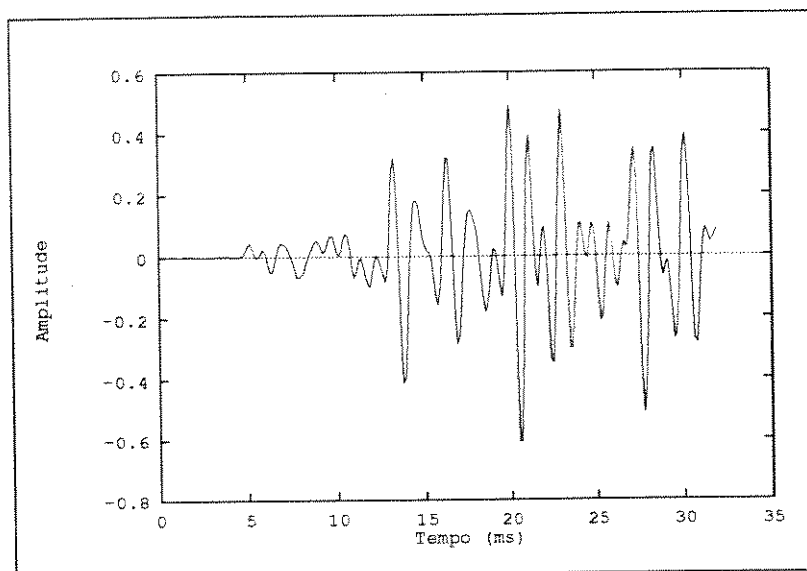


Figura 5.9: Sinal no tempo para o trecho inicial de 32ms da sílaba /pá/ amostrada a 8kHz.

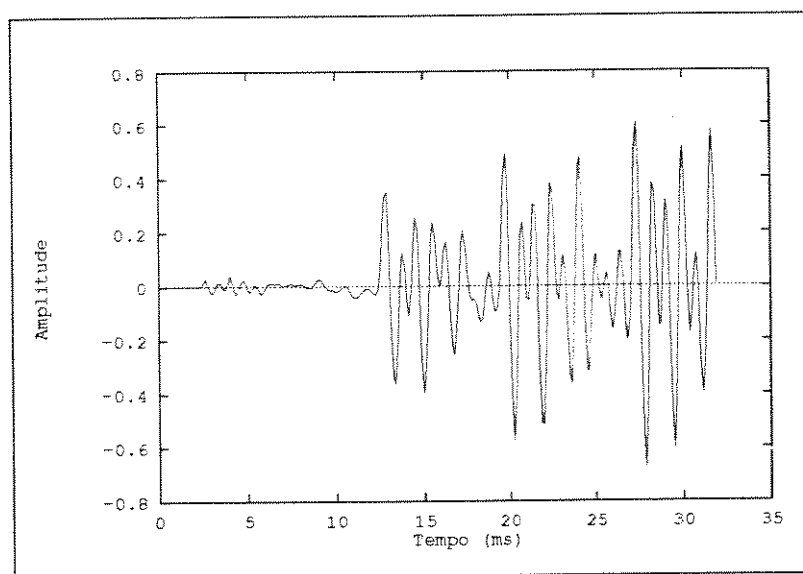


Figura 5.10: Sinal no tempo para o trecho inicial de 32ms da sílaba /tá/ amostrada a 8kHz.

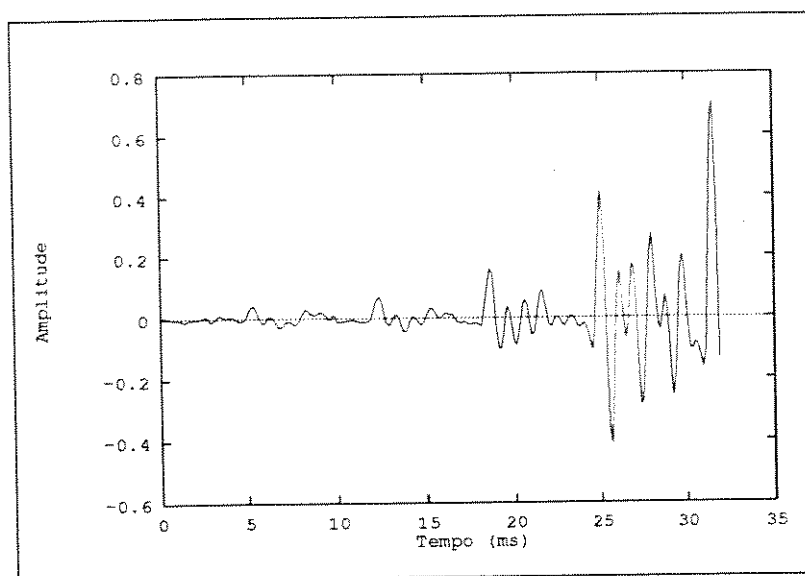


Figura 5.11: Sinal no tempo para o trecho inicial de 32ms da sílaba /vá/ amostrada a 8kHz.

discreto (amostrado a 8kHz) para as sílabas /pá/ e /tá/, respectivamente. As oclusões sonoras diferenciam-se dos fricativos sonoros por atenuar as componentes em frequência dos sons produzidos pelo trato vocal na faixa entre 1 e 3kHz do espectro durante a oclusão [34].

### Sons de Excitação Mista

Os sons *fricativos sonoros*, como /j/, /v/ e /z/, são produzidos combinando-se a vibração das cordas vocais e a excitação turbulenta. Nos períodos de máxima pressão glótica o escoamento pela obstrução torna-se turbulento, gerando caráter fricativo do som. Assim que a pressão glótica cai abaixo de certo valor, extingue-se o escoamento turbulento de ar e as ondas de pressão têm um comportamento mais suave [34].

Os sons *oclusivos sonoros*, /b/, /d/ e /g/,<sup>2</sup> são produzidos de forma semelhante aos correspondentes não-sonoros, /p/, /t/ e /k/, todavia havendo vibração das cordas vocais durante a fase de fechamento da cavidade oral.

As figuras 5.11 e 5.12 mostram o sinal em tempo discreto (amostrado a 8kHz) para as sílabas /vá/ e /bá/, respectivamente.

<sup>2</sup>Como em /gato/ e não /jato/.

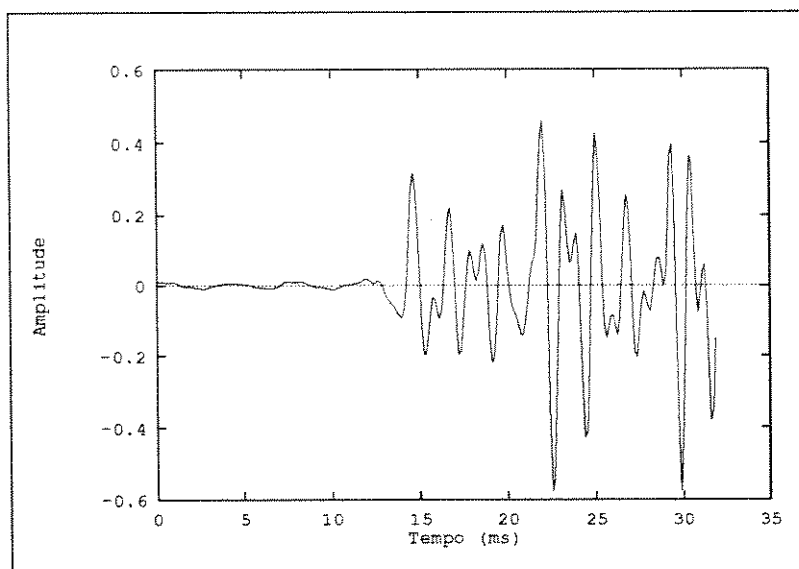


Figura 5.12: Sinal no tempo para o trecho inicial de 32ms da sílaba /bá/ amostrada a 8kHz.

### 5.1.3 Amostragem dos Padrões no Tempo

As amostras dos fonemas foram obtidas utilizando-se o conversor analógico/digital existente nas estações de trabalho *Sparc 1+* da *Sun Microsystems*. Este conversor A/D amostra a uma taxa fixa de 8KHz com 12bits de precisão, o conversor possui ainda um estágio de filtro passa-baixa com frequência de corte de 4kHz, caracterizando um sistema similar ao encontrado em telefonia [33]. As amostras são posteriormente comprimidas para 8bits através do método de compressão  $\mu$ -Law (a nível de hardware) e posteriormente linearizadas por uma biblioteca de funções de manipulação do conversor e dos dados fornecidos por ele, disponível nas estações de trabalho [32].

Este sistema de amostragem é comparativamente pobre com relação aos existentes no mercado, que amostram a 12bits com taxas de aquisição de cerca de 44kHz, porém possui a característica de ser um sistema fácil de ser construído e pouco dispendioso (em recursos de memória e financeiros). Na verdade, em alguns sons amostrados, o sistema de aquisição das estações de trabalho mostrou-se pior que o sistema telefônico. Com essa taxa de amostragem e banda de passagem os sistemas telefônicos podem ser os maiores usuários de um sistema de reconhecimento de fonemas tal qual o descrito neste trabalho ou de um sistema de reconhecimento de pequenas palavras, como os algarismos numéricos, que poderá ser construído utilizando-se as técnicas apresentadas neste trabalho.

Apesar da baixa qualidade do sinal amostrado deve ficar claro que o intuito deste trabalho é analisar o comportamento das redes na classificação de sinais ruidosos e de baixa qualidade. Como existe uma dificuldade natural em se reconhecer a voz e algumas palavras dos usuários de sistemas telefônicos (devido à banda de passagem de 3,4kHz ser muito estreita), espera-se encontrar dificuldades em se reconhecer alguns tipos de fonemas nesta taxa de aquisição (8kHz).

Os fonemas foram divididos nos três principais grupos de fonemas, *vogais* (sons sonoros produzidos pelas vogais), *explosivos* (oclusivos surdos e sonoros) e *fricativos* (surdos e sonoros) sendo que em cada grupo foram amostradas cinquenta *sílabas* (ou fonemas) *isoladas*, de cada um dos fonemas pertencentes ao grupo. Tentou-se manter, durante a amostragem, a pronúncia dos fonemas o mais constante possível para evitar problemas com relação à entonação e ao volume do som produzido. Destas sílabas, metade foi utilizada no ensino da rede neural e o restante durante a fase dos testes. As sílabas foram extraídas de uma amostra de som por uma janela retangular de 64ms (512 amostras) sendo os seguintes fonemas e sílabas utilizados, de cada um dos grupos acima:

**Vogais**— foram amostradas as vogais /á/, /é/, /i/, /ó/ e /ú/;

**Explosivos**— foram amostradas as sílabas /pá/, /tá/, /bá/ e o início da vogal /á/ incluída para teste de discernimento entre vogais e vogais precedidas de explosivos, e;

**Fricativos**— foram amostradas as sílabas /sá/, /fá/, /vá/ e a vogal /á/ incluída pelo mesmo motivo apontado no caso dos explosivos.

Em cada grupo, o modo de se obter o padrão foi alterado de acordo com as características do fonema. Para as vogais, que possuem o mesmo comportamento durante um intervalo de tempo grande, foi retirada a amostra de 64ms no trecho central da locução da mesma, não incluindo nem o início nem o final da pronúncia. No caso dos explosivos procurou-se deixar uma faixa de 8 a 12ms de silêncio no início da locução. Tal faixa é uma garantia de que, realmente, os fonemas /p/, /t/ e /b/ fossem amostrados, bem como uma porção de sinal correspondente à vogal /á/. No grupo dos fricativos deixou-se uma faixa de cerca de 15 a 20ms antes do início da pronúncia da vogal /á/, pois neste caso teremos uma porção de sinal com características similares ao ruído branco antes do início da pronúncia da vogal.

Estas amostras por si só não constituem padrões válidos para as redes neurais.

Os padrões foram pré-processados através da *transformada rápida de Fourier* e do *modelo computacional da cóclea* (do capítulo 3).

#### 5.1.4 Obtenção dos Padrões para as Redes Neurais

O objetivo principal deste trabalho era de se utilizar o *modelo da cóclea* como gerador dos padrões para as redes neurais. No entanto usou-se também a *transformada rápida de Fourier* [23] para comparar o comportamento da rede frente aos padrões originados de espectros de frequência e aqueles originados do modelo da cóclea.

O sinal no tempo não foi considerado pois apresenta um grande número de variáveis dependendo de fatores como o instante exato em que ocorre a extração do padrão no tempo, o que acarretaria problemas com relação à fase. Na verdade, segundo a fisiologia do ouvido humano e os trabalhos relacionados ao processamento, análise, produção e reconhecimento de voz, o uso do domínio da frequência (ou no domínio “posicional” da cóclea) é recomendado [25, 34]. Normalmente supõe-se que a fase do sinal tem importância pouco significativa no reconhecimento da voz [18, 25, 34].

#### Padrões Baseados na Transformada Rápida de Fourier

Para a obtenção dos padrões neurais baseados na *transformada rápida de Fourier* (FFT — Fast Fourier Transform), utilizou-se o espectro em frequência segmentado no tempo, conforme esquema da figura 5.13. Cada segmento do sinal de voz possui duração fixa de 16ms, sendo este intervalo de tempo pequeno o suficiente para que os articuladores (língua, lábios, etc) do trato vocal não variem significativamente (podem ser considerados estacionários em intervalos menores do que 30ms). Mas também devem conter pelo menos dois períodos de pitch o que limita em intervalos maiores que 10ms [34]. O número e a disposição dos segmentos foram baseados nos trabalhos de R. W. Prager e F. Fallside [24] e Jeffrey L. Elman e David Zipser [3].

Cada segmento do sinal de voz obtido foi multiplicado, antes de se calcular as componentes de frequência pela FFT, pela *janela de Hamming* [23, 25, 34] dada pela seguinte equação (para uma taxa de amostragem de 8kHz)

$$h(n) = \begin{cases} 0,54 - 0,46 \cdot \cos\left(\frac{2\pi n}{N-1}\right), & \text{se } 0 \leq n < N \\ 0, & \text{caso contrário,} \end{cases} \quad (5.1)$$

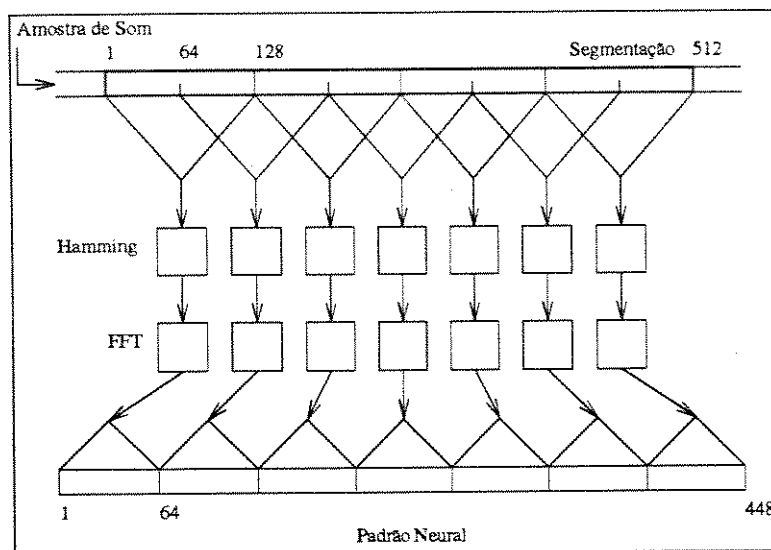


Figura 5.13: Metodologia usada para extração de um padrão de voz através do espectro em frequência segmentado no tempo.

onde  $N$  é o número de amostras do segmento. A janela de Hamming foi utilizada neste trabalho para eliminar as descontinuidades no início e no fim da seqüência de amostras do segmento de tempo que podem provocar problemas na convergência do algoritmo de FFT (fenômeno de Gibbs [23]).

Após o *janelamento* do sinal o espectro em frequência foi obtido através da transformada de Fourier discreta dada pela equação [23, 25]

$$S_i(k) = \sum_{n=0}^{N-1} (s_i(n) \cdot h(n)) \exp\left(-j \frac{2\pi nk}{N}\right), \quad (5.2)$$

onde  $S_i(k)$  é o espectro de frequência do sinal contido no  $i$ -ésimo segmento (no caso sete segmentos),  $s_i(n)$  correspondem aos sinais sonoros de cada segmento e  $h(n)$  à janela de Hamming dada pela equação (5.1).  $N$  é o número de amostras do segmento e  $k$  varia entre  $0 \leq k < N$ . Como a segunda metade de  $S_i(k)$  contém a mesma informação da primeira metade ( $S_i(p) = S_i(N - 1 - p)$  onde  $0 \leq p < N/2$ ) somente foi considerada a primeira metade do espectro, ou seja,  $0 \leq k < N/2$ . Este corte fornece uma economia significativa de memória, reduzindo a quantidade de dados de  $7 \times 128 = 896$  para  $7 \times 64 = 448$ .

A equação (5.2) não corresponde diretamente a FFT, sendo esta apenas um algoritmo computacional. No entanto, neste trabalho preferiu-se utilizar a *transformada rápida de Hartley* por se tratar de uma transformada cujo domínio e imagem são números reais e

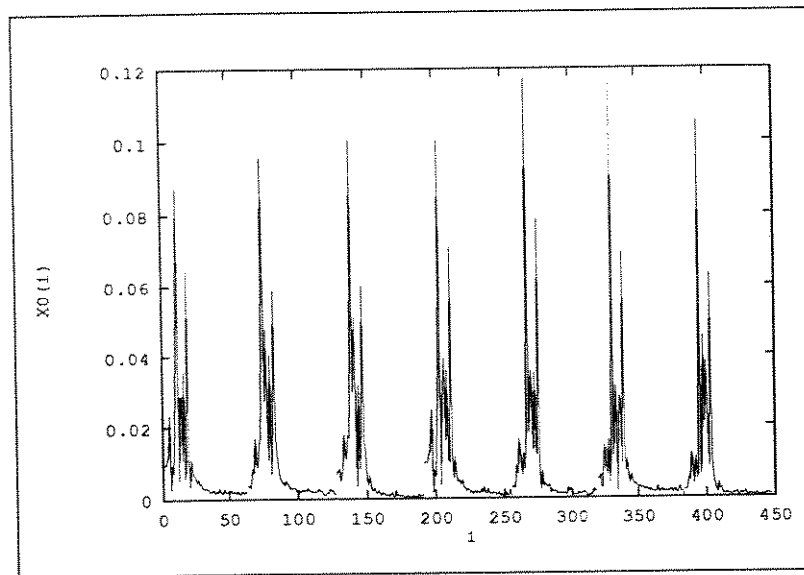


Figura 5.14: Padrão neural para a vogal /á/ obtido através da FFT dos 7 segmentos extraídos conforme a figura 5.13.

que possui relação direta com a transformada rápida de Fourier. A descrição detalhada da *transformada discreta de Hartley* pode ser vista no apêndice C. A principal vantagem desta transformada é economizar tempo computacional diminuindo a quantidade de operações aritméticas necessárias para a obtenção do espectro.

Posteriormente à obtenção dos espectros, estes dados foram organizados na forma de um vetor contendo os segmentos obtidos (de um único padrão). Utilizando-se a notação do capítulo 4 o vetor  $\mathbf{x}_0$  corresponde ao vetor de entrada da rede, assim um dos cinquenta padrões de um dos fonemas de um dos grupos é obtido através da expressão

$$\mathbf{x}_0 = \{1, S_0(0), \dots, S_0(N-1), S_1(0), \dots, S_1(N-1), \dots, S_6(0), \dots, S_6(N-1)\}^T, \quad (5.3)$$

onde  $S_i(\cdot)$  vem da equação (5.2). A figura 5.14 mostra um exemplo típico de um padrão correspondente a uma vogal /á/ sendo a figura correspondente ao gráfico  $x_{0i} \times i$ . A constante 1 no primeiro elemento do vetor não é contada como uma entrada válida ou pertencente ao padrão, pois ela somente é usada para representar o limiar para a camada dos neurônios escondidos que processam os dados deste vetor.



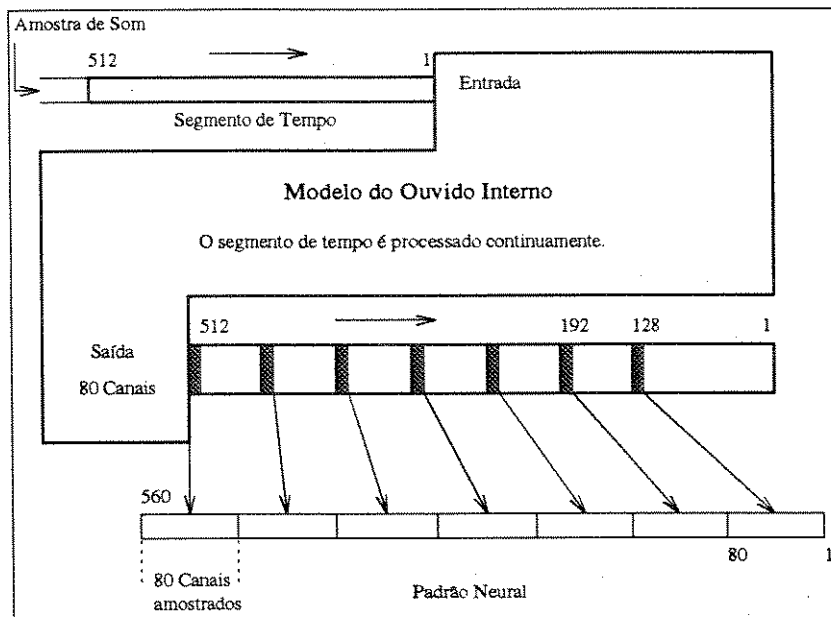


Figura 5.15: Metodologia usada para extração de um padrão de voz através do modelo da cóclea.

### Padrões Baseados no Modelo do Ouvido Interno

A obtenção dos padrões neurais baseados no modelo computacional da cóclea é facilitado pela própria natureza do modelo, visto que o mesmo já fornece representações neurais em sua saída. No entanto o modelo fornece saídas no tempo o que acarreta uma enorme quantidade de dados. Para a taxa de amostragem de 8kHz utilizada neste trabalho e com os parâmetros do capítulo 3 o modelo terá 80 canais de filtros (um vetor de saída com 80 elementos) gerando, para os mesmos 64ms de amostragem, cerca de  $80 \times 512 = 40\,960$  dados.

A redução do número total dos dados foi obtida passando-se os 64ms de sinal de voz amostrado (continuamente) pelo modelo do ouvido e amostrando o vetor de saída do modelo a cada 8ms (a primeira amostra é obtida somente após 16ms, devido à cascata de filtros), exatamente no instante final de cada segmento da figura 5.13 gerando o método de extração mostrado na figura 5.15.

Obteve-se portanto  $7 \times 80 = 560$  dados para a formação do padrão. Do mesmo modo dos padrões baseados em FFT os padrões de entrada da rede neural foram obtidos

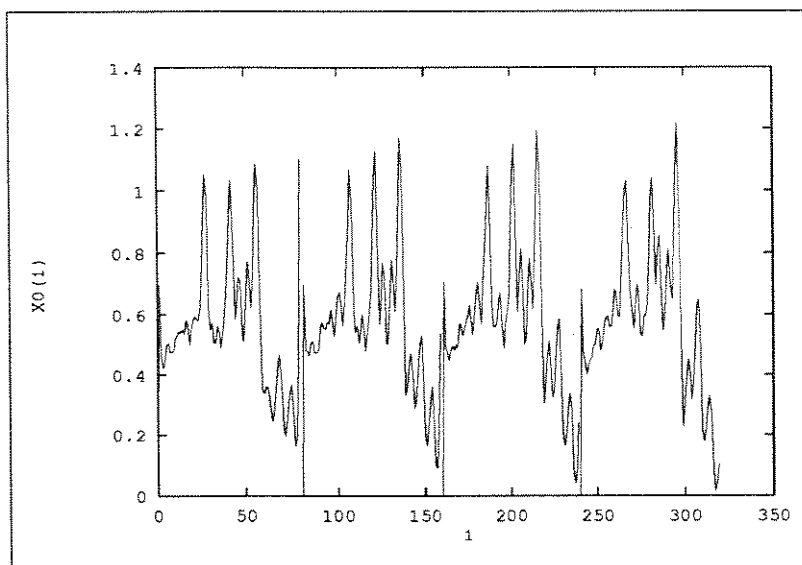


Figura 5.16: Padrão neural para a vogal /á/ obtido através do modelo da cóclea das 4 últimas amostras do vetor de saída extraídos conforme a figura 5.15.

através da construção do vetor de entrada  $\mathbf{x}_0$ ,

$$\mathbf{x}_0 = \{1, C_0(0), \dots, C_0(79), C_1(0), \dots, C_1(79), \dots, C_6(0), \dots, C_6(79)\}^T, \quad (5.4)$$

onde  $C_i(n)$  é a resposta do  $n$ -ésimo canal da cóclea da  $i$ -ésima amostra extraída de acordo com a figura 5.15. A figura 5.16 mostra um padrão típico (gráfico de  $x_{0i} \times i$ ) fornecido pela cóclea artificial, para o mesmo /á/ da figura 5.14. O primeiro elemento do vetor  $\mathbf{x}_0$ , a constante 1, não é contado como uma entrada válida pois somente é usado para fornecer o limiar para a camada de neurônios escondidos que processam os dados da entrada.

Em testes preliminares feitos com os padrões obtidos pelo modelo da cóclea foram eliminados alguns dos vetores amostrados na saída da cóclea. Os conjuntos de dados amostrados originalmente estavam provocando oscilações no algoritmo de “backpropagation” dificultando a aprendizagem da rede. Após uma análise visual dos padrões constatou-se que os elementos correspondentes aos dois primeiros vetores amostrados ( $\mathbf{C}_0$  e  $\mathbf{C}_1$ ) apresentavam valores bem menores que os elementos dos vetores restantes e que, também, não forneciam diferenças significativas entre os vários fonemas amostrados (no caso dos fonemas explosivos e fricativos, no caso das vogais incluiu-se o terceiro vetor,  $\mathbf{C}_2$ ). Assim, para o grupo das vogais optou-se por retirar os 3 primeiros vetores amostrados e para os grupos fricativo e explosivo retirou-se os 2 primeiros e o último ( $\mathbf{C}_6$ ). No caso dos explosivos

observou-se que o sétimo (último) vetor amostrado é semelhante em todos os padrões do grupo, este vetor já correspondia, portanto, à pronúncia da vogal /á/ das sílabas. O sétimo vetor amostrado dos padrões do grupo dos fricativos foi retirada pelo mesmo motivo. A retirada destes vetores dos padrões não comprometeu o desempenho das redes, como poderá ser visto no capítulo 6, além disso a quantidade de dados foi reduzida para  $4 \times 80 = 320$  acelerando o processo da aprendizagem.

## 5.2 Especificação das Redes Neurais

Foram implementadas nove configurações de redes neurais multi-camadas para cada conjunto de padrões extraídos pelos dois métodos acima descritos. Destas nove redes implementadas, quatro foram específicas para os padrões obtidos pela FFT e cinco para o modelo da cóclea. Em todos os casos a função de ativação dos neurônios foi sempre a *sigmóide*, dada pela equação (4.3):

$$f(x) = \text{sigmóide}(x) = \frac{1}{1 + \exp(-x)}.$$

Foram criadas diversas configurações de redes neurais pois as formas e as quantidades de regiões de decisão, necessárias para a correta classificação dos padrões, são desconhecidas. Como o número de neurônios da camada escondida define a complexidade (e o número de regiões) que uma rede multi-camadas pode formar (capítulo 4), foi necessário mudar o número de neurônios entre uma rede e outra para avaliar o desempenho obtido pelas mesmas.

Usa-se aqui representar o número de neurônios das camadas das redes neurais (capítulo 4) por uma seqüência de números entre parênteses, ( $\dots$ ) e separados por “-”, assim uma rede de três camadas contendo três entradas, oito neurônios na primeira camada escondida, cinco na segunda e quatro na de saída fica representada pela seqüência (4-5-8-3). Esta notação visa somente simplificar a identificação da rede no texto.

Os padrões obtidos pelos dois métodos de extração para cada um dos três grupos de fonemas foram apresentados às redes durante a fase de ensino e testes. Sendo que para uma rede ensinada a discernir *vogais* não foram apresentados os padrões obtidos de *explosivos* ou *fricativos*. No caso dos grupos *explosivo* e *fricativo* foi adicionada uma vogal para testar a capacidade da rede em discernir vogais precedidas das consoantes dos grupos e vogais isoladas.

Configurações das Redes para FFT	
Número	Configuração
1	FFT( $x$ -100-448)
2	FFT( $x$ -50-75-448)
3	FFT( $x$ -25-50-448)
4	FFT( $x$ -50-50-448)

Tabela 5.1: Configurações utilizadas para as redes neurais multi-camadas que recebem padrões extraídos com a FFT,  $x$  vale 4 para explosivos e fricativos e 5 para as vogais.

O número de saídas de cada uma das redes depende apenas do número de tipos de padrões diferentes que cada grupo possui, assim teremos cinco neurônios na saída das redes utilizadas no grupo *vogais* e quatro nos demais grupos. Cada neurônio de saída está associado a um único fonema (ou sílaba), assim, por exemplo, para o grupo *vogais* se o neurônio associado a vogal /á/ estiver ativo então considera-se que a entrada seja um padrão neural correspondente a esta vogal.

Durante o ensino dos padrões considera-se um neurônio ativo aquele que tem como saída o valor 0,9 e um inativo o valor 0,1. Na fase dos testes utilizou-se *competição* entre os neurônios da camada de saída para isolar o ativo, ou seja, durante esta fase considera-se o neurônio ativo aquele que tem o maior valor numérico de saída com relação aos outros da mesma camada. Este critério possibilita a eliminação da *confusão* acarretada quando mais de um neurônio pode ser considerado ativo (por exemplo, quando dois deles possuem as saídas 0,90 e 0,89 e os demais em torno de 0,1; por exemplo).

### 5.2.1 Redes para os Padrões da FFT

Para os padrões extraídos pela FFT, foram usadas quatro redes neurais diferentes. Preferiu-se utilizar as mesmas quatro redes em cada um dos grupos *vogais*, *fricativos* e *explosivos* para testar a capacidade das mesmas perante as características diferentes dos fonemas de cada grupo.

Estas redes são identificadas pelo prefixo FFT e foram especificadas de acordo com a tabela 5.1. Assim a rede número 2, para as vogais, será identificada como FFT(5-50-75-448) quando referenciada neste texto.

O número de entradas no caso do conjunto de padrões FFT é constante e igual a 448, já as saídas dependem do número de tipos diferentes de fonemas, assim, para as vogais

Configurações das Redes para Cóclea	
Número	Configuração
1	COC( $x-25-50-320$ )
2	COC( $x-40-80-320$ )
3	COC( $x-50-100-320$ )
4	COC( $x-45-85-320$ )
5	COC( $x-50-90-320$ )

Tabela 5.2: Configurações utilizadas para as redes neurais multi-camadas que recebem padrões extraídos com o modelo da Cóclea,  $x$  vale 4 para explosivos e fricativos e 5 para as vogais.

têm-se 5 saídas, e para os explosivos e fricativos têm-se 4 saídas. Deve-se lembrar que cada neurônio de saída é associado a um único fonema do grupo, assim os 5 neurônios das redes do grupo *vogais* estarão associados aos fonemas /á/, /é/, /i/, /ó/ e /u/, respectivamente.

### 5.2.2 Redes para os Padrões da Cóclea

No caso dos padrões neurais fornecidos pela cóclea foram especificadas cinco redes diferentes para serem usadas nos três grupos de fonemas. Do mesmo modo que das redes FFT(.) estas cinco redes foram utilizadas em cada um dos grupos visando testar suas habilidades em classificar diferentes tipos de fonemas.

As redes usadas para este conjunto de padrões são identificadas pelo prefixo COC e foram especificadas segundo a tabela 5.2. Assim a rede número 4 é representada por COC(5-45-85-320).

De maneira similar ao caso das redes FFT(...) a entrada das redes COC(...) é constante e igual a 320 e as saídas são 5 para as vogais e 4 para os grupos dos explosivos e fricativos. Do mesmo modo cada neurônio de saída é associado a um único fonema do grupo, assim cada um dos 5 neurônios das redes do grupo *vogais* estará associado a um dos fonemas /á/, /é/, /i/, /ó/ ou /u/.

## 5.3 Comentários sobre a Implementação

A implementação deste classificador de fonemas foi realizada em computador digital e através de programas escritos na linguagem C e C++ [31]. Estes programas foram projetados para serem independentes uns dos outros e realizam basicamente as seguintes

tarefas:

- amostrar o sinal já digitalizado com uma janela de 64ms gerando a base das amostras dos fonemas;
- segmentar e gerar os padrões, baseados no método da FFT, para as redes neurais;
- construir e usar o modelo da cóclea sobre as amostras dos fonemas para gerar padrões para as redes neurais e
- construir, ensinar e testar redes perceptron multi-camadas com os padrões obtidos.

Foram criadas classes de objetos na linguagem C++ para facilitar a construção dos programas e de certa forma facilitar desenvolvimentos futuros nesta área. Estas classes de objetos implementam a manipulação de matrizes, de filtros discretos, de redes neurais multi-camadas genéricas e do modelo do ouvido interno. Os códigos fontes para os programas acima e para estas classes totalizam aproximadamente doze mil linhas de código.

Durante a fase de ensino usou-se a seguinte expressão (aqui neste trabalho chamada de *erro máximo*) para o cálculo do erro de cada um dos padrões:

$$\varepsilon_{\mathbf{p}} = \max_i (|d_i - x_{Mi}|), \quad (5.5)$$

onde  $d_i$  é a saída desejada e  $x_{Mi}$  é a saída obtida pela rede para um determinado padrão de entrada  $\mathbf{p}$ . A vantagem do *erro máximo* com relação ao *erro quadrático* (definido no capítulo 4) é que o primeiro é mais simples de ser calculado e garante que o erro provocado por cada elemento  $x_{Mi}$  esteja contido em uma faixa de valores.

O maior valor do *erro máximo* admissível para cada um dos padrões  $\mathbf{p}$  de entrada foi de 0,1 para todas as redes com exceção das redes COC(x-50-100-320) em que o mesmo foi reduzido para 0,01. Em todas as redes, no caso do algoritmo de “backpropagation” (capítulo 4 na seção 4.4.1), os valores da taxa de ensino ( $\mu$ ) e do amortecimento ( $\alpha$ ) foram ajustados para obter a convergência o mais rapidamente possível.

A utilização do algoritmo de aprendizagem foi realizada por épocas que correspondem aos seguintes passos:

1. Apresenta-se um padrão à rede e calcula-se o erro associado a ele;
2. Se o erro máximo associado ao padrão estiver acima do valor máximo admissível utilizar o algoritmo de aprendizagem uma vez;

3. Repetir (1), utilizando-se um novo padrão, até todos os padrões serem testados.

Esses passos (1 época) são repetidos até que todos os padrões do conjunto de aprendizagem produzam erros máximos menores que o erro máximo admissível.

Infelizmente o método de aprendizagem dos *mínimos quadrados recursivo* (seção 4.4.2 do mesmo capítulo) não foi possível de ser utilizado em nenhuma circunstância, pois, com o número elevado de padrões de ensino usado neste trabalho, o método apresentava uma elevada taxa de crescimento dos pesos dos neurônios, acarretando erro computacional (somente foi observado esse tipo de problema com o algoritmo quando é utilizado um conjunto grande de padrões durante a fase de aprendizagem). Este erro ocorre principalmente devido ao uso da inversa da função de ativação do neurônio e devido a estimativa do erro para as camadas escondidas o que provavelmente provoca algum efeito acumulativo quando é utilizado um grande conjunto de padrões durante a fase de aprendizagem. Para contornar o problema dos pesos provocada por este método de aprendizagem tentou-se limitar os valores dos pesos dentro de uma faixa de valores, mas houve persistência do problema. Se este método for ajustado, em trabalhos futuros, para suportar um grande número de padrões durante a fase de ensino, possibilitará uma melhora significativa nos sistemas de reconhecimento de padrões. Este algoritmo, apesar dos problemas, mostrou ser melhor que o "backpropagation" para o caso de redes com um número reduzido de padrões a serem ensinados (capítulo 4).

Deve-se salientar que o número de padrões utilizados para cada fonema de cada grupo durante as fases de ensino e teste foram fixados em cinquenta no total, devido ao limitado espaço em disco das estações de trabalho em que os programas foram executados.

## Capítulo 6

# Resultados Experimentais

Após as redes terem sido ensinadas, foi realizada a fase de testes, que corresponde em se utilizar os 25 padrões restantes, que não foram usados durante a aprendizagem, como entradas para as redes. Os resultados das saídas das redes para cada um destes padrões mostram se a configuração da rede e o algoritmo de aprendizagem foram suficientes para a classificação correta destes padrões.

O tempo de execução dos programas durante a fase de aprendizagem foi de cerca de 20 horas em tempo de CPU (*Sparc 1+* da *Sun Microsystems*), que em tempo real corresponde a cerca de 3 dias até que as redes convergissem abaixo do *erro máximo admissível* (definido no capítulo 5). O número de épocas até a convergência variou de acordo com o tipo de entrada da rede sendo que em média uma rede FFT( $\cdot$ ) convergiu em 2 800 épocas e uma rede COC( $\cdot$ ) em 360 épocas sendo que a rede COC(5-50-100-320), cujo erro máximo admissível foi de 0,01, convergiu em 2 000 épocas (média).

A título de recordação, uma rede especificada por FFT(5-50-75-448) corresponde a uma rede cujos padrões de entrada são aqueles obtidos pela *transformada de Fourier* e cuja configuração corresponde a uma rede de 3 camadas com 448 entradas, 75 na primeira camada escondida, 50 na segunda e com 5 neurônios na camada de saída. O modo de se obter os valores das saídas consiste em se aplicar recursivamente a equação (4.6) a partir da primeira camada de neurônios escondidos até a camada de saída:

$$y_j = f \left( \sum_{i=0}^N w_{ij} x_i \right), \quad x_0 = 1,$$

Cada uma das saídas da rede neural está ligada a um único grupo de padrões da entrada, assim, cada um dos cinco neurônios das saídas nas redes FFT( $\cdot$ ) ou COC( $\cdot$ )



utilizadas com os padrões extraídos das vogais representará apenas um dos 5 fonemas correspondentes aos sons das vogais /á/, /é/, /i/, /ó/ e /u/. Deve-se observar que, em todas as redes, os padrões de saída são considerados mutuamente exclusivos.

O neurônio considerado ativo, ou seja, aquele que classificou a entrada, é encontrado através da *competição* entre os neurônios da camada de saída, conforme mencionado no capítulo 5, isto é, aquele neurônio que possui o maior valor de saída é considerado o neurônio que classificou o padrão de entrada. Este tipo de atitude quanto à escolha do neurônio de saída ativo possibilita a eliminação da *confusão*.

A *confusão* ocorre quando mais de um neurônio da camada de saída respondem com valores muito próximos, como por exemplo, quando dois deles estão com os valores de saída 0,9 e 0,8999. Neste caso gera-se uma dúvida sobre qual dos dois neurônios é realmente o ativo. Como neste trabalho as saídas das redes são mutuamente exclusivas e presumindo-se que não haverá dois ou mais neurônios com exatamente o mesmo valor de saída, a *competição* eliminará esta *confusão*.

A seguir serão mostrados os resultados obtidos com os diferentes grupos de padrões para cada uma das redes utilizadas. Os resultados foram divididos em duas tabelas, uma com a porcentagem dos padrões reconhecidos corretamente e outra com o erro do reconhecimento e o número de ocorrências do erro (em porcentagem). Colocou-se os dados sobre os erros cometidos pelas redes para observar o modo como as redes *erram*.

Todos os valores estão em porcentagem com relação ao número de padrões utilizados no teste (25 padrões), com exceção do *total* que corresponde à porcentagem com relação aos 125 padrões do grupo vogal e aos 100 padrões dos grupos explosivos e fricativos.

## 6.1 Padrões Correspondentes às Vogais

Os resultados obtidos com as redes neurais para os padrões extraídos via *transformada rápida de Fourier* são mostrados nas tabelas 6.1 e 6.3, correspondendo aos acertos e erros das redes, respectivamente.

Pode-se observar que, pela tabela 6.1, houve uma melhora quando utilizadas redes de 3 camadas. Realmente, com a impossibilidade de se obter formatos de regiões complexas que conteriam os padrões de entrada (neste caso  $\mathcal{R}^{448}$ ) e devido principalmente a rede FFT(5-100-448) ter alcançado um rendimento insatisfatório, optou-se por utilizar redes de 3 camadas nas demais configurações. Estas redes conseguem isolar regiões no espaço de

Resultados das Redes—Vogais/FFT						
Rede	/á/	/é/	/i/	/ó/	/u/	Total
FFT(5-100-448)	92%	28%	92%	100%	100%	82,4%
FFT(5-50-75-448)	80%	76%	100%	100%	88%	88,8%
FFT(5-25-50-448)	88%	60%	76%	100%	72%	79,2%
FFT(5-50-50-448)	92%	64%	76%	100%	76%	81,6%

Tabela 6.1: Acertos obtidos durante os testes com as redes ensinadas com os padrões da FFT correspondentes às vogais.

Resultados das Redes—Vogais/Cóclea						
Rede	/á/	/é/	/i/	/ó/	/u/	Total
COC(5-25-50-320)	88%	80%	60%	92%	52%	74,4%
COC(5-40-80-320)	96%	56%	68%	80%	60%	72,0%
COC(5-50-100-320)	92%	48%	76%	88%	72%	75,2%
COC(5-45-85-320)	96%	72%	72%	84%	64%	71,2%
COC(5-50-90-320)	96%	40%	72%	76%	72%	71,2%

Tabela 6.2: Acertos obtidos durante os testes com as redes ensinadas com os padrões da Cóclea correspondentes às vogais.

Erros das Redes—Vogais/FFT					
Rede	/á/	/é/	/i/	/ó/	/u/
FFT(5-100-448)	8%:/ó/	12%:/ó/ 60%:/u/	8%:/u/	—	—
FFT(5-50-75-448)	20%:/ó/	8%:/ó/ 16%:/u/	—	—	8%:/é/ 4%:/á/
FFT(5-25-50-448)	12%:/ó/	8%:/ó/ 28%:/u/	24%:/u/	—	28%:/é/
FFT(5-50-50-448)	8%:/ó/	4%:/ó/ 8%:/u/	24%:/u/	—	24%:/ó/

Tabela 6.3: Erros obtidos durante os testes com as redes ensinadas com os padrões da FFT correspondentes às vogais.

Erros das Redes—Vogais/Cóclea					
Rede	/á/	/é/	/i/	/ó/	/u/
COC(5-25-50-320)	12%:/ó/	20%:/u/	40%:/u/	8%:/u/	24%:/é/ 24%:/ó/
COC(5-40-80-320)	4%:/ó/	44%:/u/	20%:/u/	20%:/u/	40%:/ó/
COC(5-50-100-320)	4%:/ó/ 4%:/u/	52%:/u/	4%:/é/ 20%:/u/	12%:/u/	12%:/é/ 16%:/ó/
COC(5-45-85-320)	4%:/ó/	28%:/u/	28%:/u/	20%:/u/	20%:/é/ 16%:/ó/
COC(5-50-90-320)	4%:/ó/	60%:/u/	28%:/u/	16%:/u/	12%:/é/ 12%:/ó/

Tabela 6.4: Erros obtidos durante os testes com as redes ensinadas com os padrões da Cóclea correspondentes às vogais.

entrada mais complexas que as de menor número de camadas (vide capítulo 4).

As redes que aprenderam os padrões extraídos pela *cóclea* apresentaram resultados parecidos que são mostrados nas tabelas 6.2 e 6.4. O fato das redes FFT(·) e COC(·) apresentarem resultados semelhantes deve-se à proximidade dos métodos de extração, pois a cóclea realiza um mapeamento em frequência bem parecido com a transformada de Fourier, e também devido a baixa qualidade do sinal de áudio.

As tabelas que mostram a maneira como as redes erraram fornecem dados interessantes sobre elas. Pode-se observar pelas tabelas 6.3 e 6.4 que as redes seguem um comportamento específico e vicioso quanto aos erros nos dois métodos de extração de padrões (cóclea e FFT). Estas redes sempre confundem /á/ com /ó/, /é/ e /i/ com /u/ e /u/ com /é/ e /ó/ e ainda as redes COC(·) confundem o padrão /ó/ com /u/.

## 6.2 Padrões Correspondentes aos Fonemas Explosivos

Os testes referentes aos padrões de fonemas explosivos forneceram os resultados apresentados nas tabelas 6.5 e 6.6 para as redes FFT(·) e COC(·), respectivamente.

Nota-se de imediato que as mesmas redes utilizadas com os padrões extraídos das vogais foram relativamente melhores quando usadas na classificação dos fonemas explosivos. Observa-se ainda que as redes COC(·) tiveram um desempenho melhor que as FFT(·).

Os erros obtidos com as redes são mostrados nas tabelas 6.7 e 6.8 respectivamente para as redes FFT(·) e COC(·). Do mesmo modo que no caso das vogais, observa-se um

Resultados das Redes—Explosivos/FFT					
Rede	/pá/	/tá/	/bá/	/á/	Total
FFT(4-100-448)	96%	88%	92%	60%	84%
FFT(4-50-75-448)	88%	92%	88%	56%	81%
FFT(4-25-50-448)	52%	32%	92%	60%	59%
FFT(4-50-50-448)	92%	92%	92%	56%	83%

Tabela 6.5: Acerto obtidos durante os testes com as redes ensinadas com os padrões da FFT correspondentes aos fonemas explosivos.

Resultados das Redes—Explosivos/Cóclea					
Rede	/pá/	/tá/	/bá/	/á/	Total
COC(4-25-50-320)	84%	80%	100%	96%	90%
COC(4-40-80-320)	96%	84%	100%	100%	95%
COC(4-50-100-320)	92%	48%	100%	100%	85%
COC(4-45-85-320)	92%	76%	100%	100%	92%
COC(4-50-90-320)	84%	96%	100%	96%	94%

Tabela 6.6: Acertos obtidos durante os testes com as redes ensinadas com os padrões da Cóclea correspondentes aos fonemas explosivos.

Erros das Redes—Explosivos/FFT				
Rede	/pá/	/tá/	/bá/	/á/
FFT(4-100-448)	4%:/bá/	4%:/bá/ 8%:/á/	8%:/pá/	28%:/pá/ 12%:/tá/
FFT(4-50-75-448)	8%:/bá/ 4%:/tá/	8%:/bá/	12%:/pá/	12%:/pá/ 32%:/tá/
FFT(4-25-50-448)	44%:/bá/ 4%:/tá/	12%:/bá/ 36%:/á/ 20%:/pá/	8%:/pá/	20%:/pá/ 20%:/tá/
FFT(4-50-50-448)	8%:/bá/	8%:/bá/	8%:/pá/	12%:/pá/ 28%:/tá/

Tabela 6.7: Erros obtidos durante os testes com as redes ensinadas com os padrões da FFT correspondentes aos fonemas explosivos.

Erros das Redes—Explosivos/Cóclea				
Rede	/pá/	/tá/	/bá/	/á/
COC(4-25-50-320)	16%:/bá/	12%:/bá/ 4%:/á/	—	4%:/tá/
COC(4-40-80-320)	4%:/bá/	16%:/á/	—	—
COC(4-50-100-320)	8%:/tá/	48%:/á/ 4%:/pá/	—	—
COC(4-45-85-320)	8%:/bá/	24%:/á/	—	—
COC(4-50-90-320)	16%:/bá/	4%:/á/	—	4%:/tá/

Tabela 6.8: Erros obtidos durante os testes com as redes ensinadas com os padrões da Cóclea correspondentes aos fonemas explosivos.

comportamento similar em relação aos erros. O fato mais importante referente aos erros é a classificação dos padrões /pá/ como /bá/, e vice-versa nas redes FFT(.). Deve-se observar que a única diferença na pronúncia dos fonemas /pá/ e /bá/ é que, durante a oclusão (fechamento de alguma parte do trato vocal, no caso os lábios), existe vibração das *cordas vocais* na pronúncia do /b/ da sílaba /bá/.

O padrão /tá/ foi confundido com /bá/, /pá/ e /á/ e o padrão /á/ confundido com /pá/ e /tá/. Os fonemas explosivos se diferenciam apenas nos primeiros instantes da pronúncia das sílabas que os contém, isto parece ser marcante nos erros das classificações dos fonemas /tá/ e /á/. O fato da sílaba /bá/ fazer vibrar as cordas vocais durante a oclusão fez dela um fonema bastante diferenciado em relação aos outros do grupo.

Em geral as redes que foram ensinadas com os padrões extraídos pelo modelo da cóclea se adaptaram melhor a este tipo de fonema.

### 6.3 Padrões Correspondentes aos Fonemas Fricativos

Os resultados obtidos com as redes FFT(.) e COC(.) ensinadas com os padrões retirados de fonemas fricativos estão apresentados nas tabelas 6.9 e 6.10 respectivamente. Os erros nas classificações dos fonemas estão apresentados nas tabelas 6.11 para as redes FFT(.) e 6.12 para as redes COC(.).

Nota-se neste caso que com exceção de um único fonema, /fá/ nas redes FFT(.) e /vâ/ nas redes COC(.), os demais foram bem diferenciados. Os erros tendem a seguir uma mesma característica nos dois tipos de redes, isto é, o fonema /sá/ é confundido quase sempre

Resultados das Redes—Fricativos/FFT					
Rede	/sá/	/fá/	/vá/	/á/	Total
FFT(4-100-448)	96%	48%	92%	100%	84%
FFT(4-50-75-448)	80%	36%	88%	96%	75%
FFT(4-25-50-448)	88%	48%	84%	100%	80%
FFT(4-50-50-448)	92%	40%	88%	100%	80%

Tabela 6.9: Acertos obtidos durante os testes com as redes ensinadas com os padrões da FFT correspondentes aos fonemas fricativos.

Resultados das Redes—Fricativos/Cóclea					
Rede	/sá/	/fá/	/vá/	/á/	Total
COC(4-25-50-320)	80%	100%	60%	100%	85%
COC(4-40-80-320)	84%	100%	60%	100%	86%
COC(4-50-100-320)	76%	100%	60%	100%	84%
COC(4-45-85-320)	76%	100%	64%	100%	85%
COC(4-50-90-320)	80%	100%	56%	96%	83%

Tabela 6.10: Acertos obtidos durante os testes com as redes ensinadas com os padrões da Cóclea correspondentes aos fonemas fricativos.

Erros das Redes—Fricativos/FFT				
Rede	/sá/	/fá/	/vá/	/á/
FFT(4-100-448)	4%:/vá/	52%:/vá/	8%:/fá/	—
FFT(4-50-75-448)	20%:/fá/	64%:/vá/	12%:/fá/	4%:/fá/
FFT(4-25-50-448)	8%:/fá/ 4%:/á/	52%:/vá/	16%:/fá/	—
FFT(4-50-50-448)	8%:/fá/	60%:/vá/	12%:/fá/	—

Tabela 6.11: Erros obtidos durante os testes com as redes ensinadas com os padrões da FFT correspondentes aos fonemas fricativos.

Erros das Redes—Fricativos/Cóclea				
Rede	/sá/	/fá/	/vá/	/á/
COC(4-25-50-320)	20%:/fá/	—	40%:/fá/	—
COC(4-40-80-320)	16%:/fá/	—	40%:/fá/	—
COC(4-50-100-320)	24%:/fá/	—	40%:/fá/	—
COC(4-45-85-320)	24%:/fá/	—	36%:/fá/	—
COC(4-50-90-320)	20%:/fá/	—	44%:/fá/	4%:/fá/

Tabela 6.12: Erros obtidos durante os testes com as redes ensinadas com os padrões da Cóclea correspondentes aos fonemas fricativos.

com /fá/ e praticamente não existem erros na classificação do fonema /á/. O fonema /á/ foi classificado corretamente, pois este não possui as características dos fonemas fricativos, que corresponde a uma faixa de sinal antes da pronúncia da vogal da sílaba com características similares ao ruído branco (vide capítulo 5).

Os fonemas /fá/ e /vá/ foram muito confundidos nos dois casos, pois são produzidos de maneira semelhante, sendo a única diferença entre eles é o fato de que o fonema /vá/ é produzido acompanhado por vibrações das cordas vocais durante a pronúncia do /v/.

Esperava-se uma confusão maior com os fonemas fricativos pois os mesmos possuem componentes nas frequências entre 5 e 10kHz [25] que foram eliminadas pela baixa taxa de amostragem utilizada neste trabalho (8kHz). De fato, quando foram ouvidas as amostras dos sons correspondentes aos fricativos houve confusão entre as sílabas principalmente com a sílaba /sá/ que foi confundida como /fá/ por várias pessoas e algumas vezes o /fá/ com o /vá/, que já era esperado. Pode-se observar pela tabela 6.12 que o modelo da cóclea se enganou da mesma maneira que as pessoas que ouviram as sílabas. Um aumento na taxa de amostragem para 10kHz e melhores condições na amostragem do sinal devem melhorar a classificação destes fonemas.

## 6.4 Testes com Outros Locutores

Com o intuito de se conhecer o comportamento das redes já ensinadas com a voz de uma única pessoa, foram realizados testes nas redes que aprenderam a reconhecer as vogais com as vozes de dois “locutores” diferentes, aqui identificados por A e B. Foram amostrados 10 sons de cada uma das vogais utilizadas no treino da rede de cada um dos locutores e processadas da mesma maneira que as vogais utilizadas durante o ensino e teste com um único locutor, aqui identificado por “locutor principal”.

Os resultados obtidos com as redes que classificam as vogais do locutor principal foram colocados nas tabelas 6.13 e 6.14 para os padrões obtidos via transformada rápida de Fourier, respectivamente para os locutores A e B. E nas tabelas 6.15 e 6.16 para os padrões obtidos via o modelo da cóclea para os locutores A e B, respectivamente. Os valores das tabelas estão em porcentagem com relação às 10 amostras de cada uma das vogais com exceção dos totais que são sobre as 50 amostras.

É curioso observar que as redes FFT(.) obtiveram resultados melhores que as redes COC(.), mesmo com relação aos testes realizados com o locutor principal. Isto deve

Resultados das Redes—Vogais/FFT—Locutor A						
Rede	/á/	/é/	/i/	/ó/	/u/	Total
FFT(5-100-448)	100%	80%	60%	90%	100%	86%
FFT(5-50-75-448)	100%	80%	90%	100%	90%	92%
FFT(5-25-50-448)	100%	90%	60%	90%	100%	94%
FFT(5-50-50-448)	100%	90%	60%	90%	100%	94%

Tabela 6.13: Acertos obtidos durante os testes com as redes ensinadas com os padrões da FFT correspondentes às vogais para o locutor A.

Resultados das Redes—Vogais/FFT—Locutor B						
Rede	/á/	/é/	/i/	/ó/	/u/	Total
FFT(5-100-448)	100%	100%	100%	80%	100%	96%
FFT(5-50-75-448)	100%	100%	100%	60%	90%	84%
FFT(5-25-50-448)	100%	100%	100%	70%	100%	94%
FFT(5-50-50-448)	100%	100%	100%	70%	100%	94%

Tabela 6.14: Acertos obtidos durante os testes com as redes ensinadas com os padrões da FFT correspondentes às vogais para o locutor B.

Resultados das Redes—Vogais/Cóclea—Locutor A						
Rede	/á/	/é/	/i/	/ó/	/u/	Total
COC(5-25-50-320)	60%	90%	40%	90%	100%	76%
COC(5-40-80-320)	30%	80%	30%	80%	90%	62%
COC(5-50-100-320)	40%	80%	70%	80%	100%	74%
COC(5-45-85-320)	20%	80%	40%	80%	100%	64%
COC(5-50-90-320)	20%	80%	40%	80%	100%	64%

Tabela 6.15: Acertos obtidos durante os testes com as redes ensinadas com os padrões da Cóclea correspondentes às vogais para o locutor A.

Resultados das Redes—Vogais/Cóclea—Locutor B						
Rede	/á/	/é/	/i/	/ó/	/u/	Total
COC(5-25-50-320)	70%	100%	10%	90%	60%	66%
COC(5-40-80-320)	60%	90%	10%	80%	50%	58%
COC(5-50-100-320)	80%	80%	30%	70%	60%	64%
COC(5-45-85-320)	50%	90%	30%	70%	60%	60%
COC(5-50-90-320)	80%	90%	10%	90%	50%	64%

Tabela 6.16: Acertos obtidos durante os testes com as redes ensinadas com os padrões da Cóclea correspondentes às vogais para o locutor B.



ter ocorrido devido ao fato dos padrões do locutor principal terem formado as fronteiras de decisão no espaço de entrada da rede e, como os padrões de teste foram obtidos nas mesmas condições dos padrões utilizados no treino, os padrões de testes permaneceram localizados nas imediações das fronteiras. Os padrões dos locutores A e B, no caso das redes FFT( $\cdot$ ), mantiveram-se dentro das regiões de decisão porém longe da fronteira, conseguindo assim, as redes FFT( $\cdot$ ) generalizar melhor na classificação das vogais.

No caso das redes COC( $\cdot$ ) provavelmente as regiões de decisão assumiram uma forma tal que a especialização destas redes em classificar as vogais do locutor principal provocou um rendimento baixo com relação aos padrões de voz dos locutores A e B. Os erros obtidos com as redes que aprenderam os padrões vindos do modelo da cóclea não seguiram nenhum padrão como o encontrado no caso do locutor principal.

Apesar dos resultado promissores obtidos com as redes neurais para os locutores A e B, não se pode afirmar que estas redes possam trabalhar como sistemas multilocutores, uma vez que elas somente foram treinadas com a voz do locutor principal. Tampouco os testes realizados podem servir de base para a validação destas redes como um sistema multilocutor, para tanto seria necessário treinar as redes com vozes de vários locutores diferentes e depois testar com um conjunto de novos locutores.

## 6.5 Testes com Palavras Inteiras

O sistema de reconhecimento de fonemas proposto neste trabalho pode ser utilizado no reconhecimento de um conjunto pequeno de palavras (vocabulário) sem qualquer modificação em relação ao processamento do sinal sonoro. Um conjunto interessante de palavras com utilização imediata em sistemas de telefonia conteria as palavras correspondentes aos numerais de “zero” (0) a “nove” (9).

Para se testar o classificador de fonemas como um sistema de reconhecimento de palavras inteiras foram amostrados 24 sinais sonoros de cada uma das palavras correspondentes aos numerais de “zero” a “nove”, sendo 12 utilizados para a formação dos padrões de ensino e 12 para os padrões de teste. O procedimento para a extração dos parâmetros do sinal sonoro através do modelo da Cóclea foi o mesmo adotado nas seções anteriores (figura 5.15). No entanto, foram processados 384ms (3072 amostras) do sinal sonoro, no qual estavam incluídas cada uma das palavras. Para cada palavra, os 80 canais de saída da Cóclea foram re-amostrados a cada 64ms (a cada 512 amostras), sendo desprezada a amos-

Resultados das Redes—Números/Cóclea		
Palavra	Rede	Rede
	10-100-200-480	10-10-20-480
/zero/	67%	83%
/um/	100%	100%
/dois/	100%	100%
/três/	92%	83%
/quatro/	92%	92%
/cinco/	92%	100%
/seis/	83%	92%
/sete/	100%	92%
/oito/	83%	75%
/nove/	100%	100%
Total	91%	92%

Tabela 6.17: Acertos obtidos durante os testes com as redes ensinadas com os padrões da Cóclea correspondentes aos números de 0 a 9.

tra no instante inicial ( $t=0s$ ). Deste modo construiu-se vetores de  $6 \times 80 = 480$  elementos utilizados como padrões de ensino e teste.

Neste experimento foram utilizadas duas redes neurais, uma com 200 e 100 neurônios, respectivamente na primeira e na segunda camadas escondidas (Rede 10-100-200-480) e outra menor com 20 neurônios na primeira camada escondida e 10 neurônios na segunda (Rede 10-10-20-480). Os resultados dos testes (acertos) das duas redes neurais estão apresentados na tabela 6.17.

Pode-se observar diretamente da tabela 6.17 que as redes neurais obtiveram êxito no reconhecimento dos números com uma margem de acerto muito semelhante àquelas encontradas nas seções anteriores. Observa-se também que a rede neural com o menor número de neurônios obteve resultados melhores. Isto se deve, principalmente, ao fato de que quanto menor o número de neurônios nas camadas escondidas maior a capacidade da mesma em generalizar sobre um conjunto de padrões. Deve-se observar que a re-amostragem dos canais da Cóclea, que foi realizada a uma taxa fixa, pode ter retirado alguma informação do sinal sonoro, como por exemplo, as variações de amplitude presentes nos fonemas explosivos.

De uma maneira geral o sistema de reconhecimento de fonemas proposto neste trabalho obteve bons índices de acerto quando utilizado no reconhecimento de palavras onde a variedade de sons (fonemas) é muito maior.

## 6.6 Comentários

Apesar da baixa qualidade do som amostrado pelo conversor A/D disponível nas estações de trabalho *Sun Sparc 1+* as redes mostraram um alto grau de discernimento na classificação dos padrões de entrada. Como pôde ser visto no pior caso, o dos sons fricativos, onde houve confusão mesmo quando ouvida as amostras dos fonemas, as redes COC(·) foram relativamente melhores que as redes FFT(·) com um índice de acerto de pelo menos 84%.

No caso das vogais as redes FFT(·) obtiveram índices de acerto um pouco melhores que as redes COC(·). Como as vogais são fonemas de natureza sonora, elas podem ser identificadas através de picos em determinadas posições do espectro de frequência, o fato da transformada rápida de Fourier conseguir obter um espectro de frequência mais detalhado do que o obtido pelo modelo do ouvido interno (que é baseado na filtragem das componentes dos sinais), caracterizado por ser um sistema posicional, explica a boa classificação dos fonemas por parte das redes FFT(·). Uma melhora significativa na classificação das redes COC(·) pode ser obtida se se isolar os canais onde certamente existirão as informações capazes de identificar as vogais e usá-los como entrada das redes neurais.

Devido à facilidade de manejo do modelo da cóclea, este oferece possibilidades de uso em reconhecimento de voz melhores que as oferecidas pela transformada rápida de Fourier. Outra vantagem do modelo da cóclea é que este fornece continuamente as informações dos canais, sincronizados com a entrada, podendo haver uma amostragem dos valores dos canais a qualquer instante de tempo. Este fato permite uma concentração de amostras em determinadas regiões dos fonemas onde há grande variação das componentes do sinal sonoro, como por exemplo, no início das sílabas explosivas como a sílaba /pá/.

Com relação aos tempos de execução, em uma estação de trabalho *Sun Sparc 1+*, os tempos gastos nos processos que calculam os padrões utilizados pelas redes são em média de 28ms para os sete espectros de frequência da transformada rápida de Fourier (FFT) e de 251ms para as 512 amostras dos canais da cóclea (que posteriormente foram re-amostrados). O tempo de execução do processo que obtém a transformada rápida de Fourier, portanto, é de cerca de 4ms contra os 0,5ms gastos para se calcular cada passo dos filtros e CAGs do modelo da cóclea. Os dois métodos podem ser implementados em circuitos integrados, existindo componentes digitais capazes de calcular a FFT e componentes capazes de realizar a função de filtros discretos com grande rapidez.

## Capítulo 7

# Conclusões

É muito difícil avaliar o desempenho do algoritmo de aprendizagem das redes neurais, principalmente por eles serem sensíveis às condições iniciais e aos padrões de entrada. Ainda mais, estes algoritmos de ensino são sensíveis à ordem em que os padrões são apresentados, promovendo comportamentos diferentes para uma mesma rede. Vê-se no capítulo 6 que foi atingido um índice de pelo menos 60% de acerto no reconhecimento dos padrões, compatíveis com a baixa taxa de amostragem utilizada na aquisição dos sinais de voz. Esta baixa taxa de aquisição (8kHz) aliada à baixa qualidade do sinal amostrado não comprometeram de fato os desempenhos das redes, que comprovaram possuir um nível razoável de robustez.

Um dos principais resultados diz respeito à complexidade em se criar uma estrutura neural capaz de classificar com razoável precisão todos os fonemas utilizados na língua portuguesa falada. Uma rede desta natureza requeria uma quantidade muito grande de recursos computacionais durante a fase de aprendizagem, inviabilizando sua construção.

A melhor solução, atualmente, tem como direção a criação de redes com topologias e configurações mais simples e altamente especializadas em um conjunto fixo de padrões, que poderiam dividir-se conforme os grupos utilizados neste trabalho. Tais redes, isoladamente, reconheceriam com precisão os padrões de seus grupos e, a um nível superior, uma outra rede promoveria algum tipo de competição para obter um resultado válido.

O motivo das redes com padrões vindos do modelo da cóclea e da FFT segmentada obterem resultados parecidos é devido principalmente a uma certa similaridade com que os dois processos tratam o sinal sonoro. Entretanto, o modelo da cóclea associado às redes neurais torna-se uma combinação extremamente atrativa quando se leva em conta que ambos

podem ser implementados a nível de *hardware*. Por exemplo, Richard F. Lyon implementou em [19] uma cóclea artificial cobrindo as frequências entre 20Hz e 20kHz com 480 canais em um circuito integrado cuja corrente de alimentação é de alguns microampéres. Associados aos modelos de redes neurais existem circuitos digitais neurais como, por exemplo, o desenvolvido por T. Watanabe [35] et al, capaz de processar mais de 1 bilhão de conexões por segundo com até  $10^6$  sinapses, operando a 1,5V.

Outra possibilidade oferecida pelo modelo do ouvido interno é a capacidade de mudar a região de atuação dos filtros, podendo concentrá-los em uma região de frequências onde ocorrem a maioria dos eventos relacionados à fala. Tal característica do modelo não foi abordada neste trabalho, mas parece ser uma solução para um aumento nos índices de acerto na classificação dos fonemas, uma vez que eles possuem componentes em regiões diferentes do espectro.

Diferente da *transformada discreta de Fourier*, o *modelo da cóclea* possui características que não foram exploradas aqui em sua totalidade, assim, a partir deste modelo pelo menos três frentes de pesquisas poderiam ser abordadas:

1. Otimização do posicionamento dos filtros do modelo da cóclea para a gama de frequências dos sons característicos dos fonemas da língua portuguesa, a nível de um ou mais locutores;
2. Otimização da amostragem realizada nos canais da cóclea para a geração dos padrões neurais, como, por exemplo, retirando-se mais amostras no início de sílabas de vogais precedidas por consoantes explosivas (/b/, /p/, /g/, etc) e
3. Estudos de viabilidade do uso do modelo da cóclea em reconhecimento de fonemas de mais de um locutor (sistemas multilocutores).

A própria topologia da *rede neural* usada neste trabalho poderia ser repensada, uma vez que redes “perceptron” multi-camadas possuem características de processamento temporal muito limitadas. Assim, com relação às topologias de redes poderiam ser realizados trabalhos com o intuito de substituir as redes “perceptron” multi-camadas por:

1. Redes multi-camadas recorrentes<sup>1</sup> capazes de aprender características temporais;
2. Redes auto-associativas capazes de aprender sem supervisão ou;

---

<sup>1</sup>Redes neurais caracterizadas pela realimentação das saídas.

3. Arquiteturas híbridas, contendo mais de um tipo ou topologia de redes neurais.

Com os avanços recentes nas teorias sobre as redes neurais e devido principalmente ao crescimento das atividades relacionadas à elas, abriu-se um grande campo de pesquisa e desenvolvimento, tanto a nível de estudos sobre as redes quanto sobre as aplicações das mesmas. Também, o fato das redes neurais serem modelos de sistemas nervosos biológicos, capazes de aprender com grande desenvoltura, contribui para torná-las uma matéria de estudos no mínimo estimulante. Ainda, aliando as redes neurais às descobertas recentes em reconhecimento de voz criam-se áreas de pesquisas diversificadas e tentadoras.

# Bibliografia

- [1] Mahmood R. Azimi-Sadjadi and Ren-Jean Liou. Fast learning process of multilayer neural networks using recursive least squares method. *IEEE Trans. on Signal Processing*, 40(2):446–450, Fevereiro de 1992.
- [2] K. H. et al Davis. Automatic recognition of spoken digits. *Journal of Acoustical Society of America*, 24(6):637–642, 1952.
- [3] Jeffrey L. Elman and David Zipser. Learning the hidden structure of speech. *Journal of Acoustical Society of America*, 83(4):1615–1626, Abril de 1988.
- [4] James L. Flanagan. *Speech Analysis Synthesis and Perception*. Spinger-Verlag—Berlin, 2<sup>a</sup> edition, 1983. 3rd Printing 1983.
- [5] Arthur C. Guyton. *Fisiologia Humana e Mecanismos das Doenças*, chapter 41, pages 408–417. Guanabara, 1986.
- [6] S. Haykin. *Adaptive Filter Theory*, chapter 8. Prentice-Hall—Englewood Cliffs, 1986.
- [7] A. L. Hodgkin and Huxley L. F. A quantitative description of membrane current and its applications to conduction and excitation in nerve. *Journal of Physiology*, 117:500–544, 1952.
- [8] J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. Em *Proceedings of the National Academy of Sciences*, volume 79, pages 2554–2558, 1982.
- [9] Hsieh S. Hou. The fast hartley transform algorithm. *IEEE Trans. Acoust. Speech and Signal Processing*, C-36:147–1156, Fevereiro de 1987.

- [10] S. R. Hyde. Automatic speech recognition: A critical survey and discussion of the literature. Em McGraw Hill, editor, *Human Communication: A Unified View*, NY, 1972. E. E. David Jr. & P. B. Denes.
- [11] Youji Iiguni, Hideaki Sakai, and Hideratsu Tokumaru. A real-time learning algorithm for a multilayered neural networks based on the extended kalman filter. *IEEE Trans. on Signal Processing*, 40(4):959–966, Abril de 1992.
- [12] W. Koenig, H. K. Dunn, and L. Y. Lacy. The sound spectrograph. *Journal of Acoustical Society of America*, 17:19–49, 1946.
- [13] Teuvo Kohonen. *Self-Organization and Associative Memory*. Spring-Verlag, 3 edition, 1989.
- [14] Sir James Lighthill. Biomechanics of hearing sensitivity. *Journal of Vibration and Acoustics*, 113:1–13, Janeiro de 1991.
- [15] Richard P. Lippmann. An introduction to computing with neural networks. *IEEE ASSP Magazine*, pages 4–22, Abril de 1987.
- [16] Richard F. Lyon. A computational model of filtering, detection, and compression in the cochlea. Em *Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 1282–1285, Paris, França, Maio de 1982. IEEE.
- [17] Richard F. Lyon and Lounette Dyer. Experiments with a computational model of the cochlea. Em *Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 1975–1978, Tokyo, Japão, Abril de 1986. IEEE.
- [18] Richard F. Lyon and Niels Lauritzen. Processing speech with the multi-serial signal processor. Em *Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 981–984, Tampa, FL, Março de 1985. IEEE.
- [19] Richard F. Lyon and Carver Mead. An analog electronic cochlea. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 36(7):1119–1134, Julho de 1988.
- [20] W. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.



- [21] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, 1969.
- [22] Vernon B. Mountcastle. *Fisiologia Médica*, volume I, chapter 12, pages 382–409. Guanabara Koogan, 1978.
- [23] Alan V. Oppenheim and Ronald W. Schaffer. *Digital Signal Processing*. Prentice/Hall International, Inc., 1975.
- [24] R.W. Prager and F. Fallside. The modified kanerva model for automatic speech recognition. *Computer Speech and Language*, 3:61–81, 1989.
- [25] Lawrence R. Rabiner and Ronald W. Schaffer. *Digital Processing of Speech Signals*. Prentice-Hall, Inc.—Englewood Cliffs, 1978.
- [26] David E. Rumelhart and James L. et al McClelland. *Parallel Distributed Processing*, volume I e II. The MIT Press, 1989.
- [27] Robert S. Scalero and Nazif Tepedelenlioglu. A fast new algorithm for training feed-forward neural networks. *IEEE Trans. on Signal Processing*, 40(1):202–210, Janeiro de 1992.
- [28] M. R. Schroeder. An integrable model for the basilar membrane. *Journal of the Acoustical Society of America*, 53(2):429–434, 1973.
- [29] Malcolm Slaney. Lyon's cochlear model. Apple Technical Report 13, Apple Computer, Inc, 1988.
- [30] H. V. Sorensen, Jones D. L., C. S. Burrus, and M.T. Heidman. On computing the discrete hartley transform. *IEEE Trans. Acoust. Speech and Signal Processing*, ASSP-33:1231–1238, Outubro de 1985.
- [31] Sun Microsystems S/A. *Sun C++ Programmer's Guide*, 1989.
- [32] Sun Microsystems S/A. *SunOS 4.1 Release Manual*, 1990.
- [33] Sun Microsystems S/A. *SunOS Reference Manual*, 1990. Volume II, Seção 4, *Audio*.

- [34] Maurício Nunes Vieira. Modulação frontal para um sistema de reconhecimento automático de voz. Master's thesis, Faculdade de Engenharia Elétrica, Departamento de Comunicações, UNICAMP, Janeiro de 1990.
- [35] Takao Watanabe, Katsutaka Kimura, Masakazu Aoki, Takeshi Sakata, and Kiyoo Ito. A single 1.5-v digital chip for a  $10^6$  synapse neural network. *IEEE Trans. on Neural Networks*, 4(3):387–393, Maio de 1993.
- [36] G. Zweig, R. Lipes, and J. R. Pierce. The cochlear compromise. *Journal of the Acoustical Society of America*, 59(4):975–982, Abril de 1976.

## Apêndice A

# Análise LPC

Demonstra-se que na produção de sons não nasalados o trato vocal pode ser razoavelmente modelado por um filtro contendo apenas pólos, cuja ordem,  $M$ , é diretamente proporcional à frequência de amostragem [25, 34]. Formalmente ele pode ser escrito como:

$$V(z) = U(z) \cdot H(z) = U(z) \frac{G}{1 - \sum_{i=1}^M a_i \cdot z^{-i}}, \quad (\text{A.1})$$

onde  $H(z)$  engloba os efeitos do trato vocal e da irradiação do som,  $U(z)$  ora corresponde a uma seqüência de impulsos (trem de impulsos), ora representa uma seqüência aleatória (ruído branco). A seqüência de impulsos modela os pulsos produzidos pelas cordas vocais e a seqüência aleatória o ruído branco provocado pela constrição em alguma seção do trato vocal quando na produção dos sons fricativos surdos [25, 34]. Este modelo não considera os efeitos dos zeros do aparelho fonador, na produção de sons fricativos, consoantes nasais ou vogais nasaladas. Porém se o número de pólos for suficientemente elevado,  $H(z)$  pode simular razoavelmente o efeito destes zeros [34].

O método de *predição linear* (LPC - Linear Prediction) consiste em se obter os parâmetros da equação (A.1), o ganho ( $G$ ) e os coeficientes do polinômio do denominador ( $a_i$ ) [25, 34]. Este tipo de metodologia pode ser utilizada quando se usa o modelo simplificado acima e se considera os parâmetros do trato vocal estacionários, o que ocorre somente em intervalos curtos de tempo (cerca de 10 a 30ms) devido principalmente à inércia dos articuladores (língua, lábios, etc) [34].

Para se obter os coeficientes do polinômio do denominador deve-se extrair um segmento de comprimento finito,  $N$ , e de curta duração do sinal de voz. Este segmento,  $v(n)$ , deverá ser nulo para  $n < 0$  e  $n \geq N$ . A equação (A.1) será escrita no domínio do

tempo como

$$v(n) = G \cdot u(n) + \sum_{i=1}^M \alpha_i \cdot v(n-i) \quad (\text{A.2})$$

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

A idéia fundamental da Predição Linear consiste em aproximar cada amostra do sinal de voz pela combinação linear das amostras anteriores do sinal. Sendo  $M$  o número de amostras passadas utilizadas na combinação linear, pode-se formalizar a aproximação da amostra genérica  $v(n)$  pela relação

$$\tilde{v}(n) = \sum_{i=1}^M \alpha_i \cdot v(n-i), \quad (\text{A.3})$$

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

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

$$e(n) = v(n) - \tilde{v}(n) = v(n) - \sum_{i=1}^M \alpha_i \cdot v(n-i) \quad (\text{A.4})$$

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

$$E(n) = \sum_{n=-\infty}^{\infty} e(n)^2. \quad (\text{A.5})$$

Como o segmento de voz é nulo para  $n < 0$  e  $n \geq N$ , o erro de predição, equação (A.5) é conseqüentemente nulo para  $n < 0$  e  $n > N + M - 1$ . Com esta consideração e substituindo-se a equação (A.4) na equação (A.5), obtém-se:

$$E(n) = \sum_{n=0}^{N+M-1} \left[ v(n) - \sum_{i=1}^M \alpha_i \cdot v(n-i) \right]^2. \quad (\text{A.6})$$

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

$$\frac{\partial[E(n)]}{\partial[\alpha_i]} = 0, \quad 1 \leq i \leq M. \quad (\text{A.7})$$

Com a substituição da equação (A.6) em (A.7) e a utilização das  $M$  derivadas parciais chega-se ao seguinte sistema de equações lineares:

$$\sum_{k=1}^M \alpha_k \cdot R(|i-k|) = R(i), \quad 1 \leq i \leq M, \quad (\text{A.8})$$

onde

$$R(k) = \sum_{n=0}^{N-k-1} v(n) \cdot v(n+k) \quad (\text{A.9})$$

é a função de autocorrelação a curto prazo. As equações (A.8) e (A.9) são conhecidas por equações de Yule-Walker [25, 34], que podem ser melhor visualizadas se apresentadas na forma matricial:

$$\begin{pmatrix} R(0) & R(1) & R(2) & \dots & R(M-1) \\ R(1) & R(0) & R(1) & \dots & R(M-2) \\ R(2) & R(1) & R(0) & \dots & R(M-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(M-1) & R(M-2) & R(M-3) & \dots & R(0) \end{pmatrix} \cdot \begin{Bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_M \end{Bmatrix} = \begin{Bmatrix} R(0) \\ R(1) \\ R(2) \\ \vdots \\ R(M) \end{Bmatrix}. \quad (\text{A.10})$$

Os coeficientes  $\alpha_i$  do preditor são determinados através da solução da equação (A.10) e utilizados como estimativas dos coeficientes  $a_i$  do filtro  $H(z)$ .

Uma vez determinados os coeficientes do  $\alpha_i$  resta determinar o ganho  $G$ , que pode ser obtido por [34]:

$$G = \sqrt{R(0) - \sum_{i=0}^M \alpha_k \cdot R(k)}.x \quad (\text{A.11})$$

A exploração das simetrias da matriz de autocorrelações ( $R(|i-k|)$ ) permite a elaboração de algoritmos recursivos muito eficientes para a solução do sistema. Não é objetivo deste apêndice entrar em maiores detalhes sobre o cálculo dos coeficientes LPC, maiores informações poderão ser encontradas em [25] e [34], por exemplo.

Após o cômputo dos coeficientes  $\alpha_i$  e do ganho  $G$  é geralmente obtida a resposta em frequência do filtro  $H(z)$  que gera um espectro de frequência “suavizado” com relação a aquele obtido diretamente do sinal de voz. Tanto os coeficientes  $\alpha_i$  quanto o espectro suavizado são utilizados em sistemas de reconhecimento de voz clássicos e nos sistemas de reconhecimento com redes neurais.

## Apêndice B

# Filtragem Homomórfica de Sinais

Este apêndice não entrará em detalhes sobre a teoria de processamento homomórfico de sinais, concentrando-se apenas na filtragem homomórfica dos sinais de voz. As informações sobre o processamento homomórfico de sinais podem ser encontradas em [23] onde detalha-se bem os aspectos matemáticos, com exemplos de casos em processamento de imagens e sinais de voz, ou em [25].

Um sinal de voz pode ser caracterizado como a convolução de dois sinais principais, a excitação provocada pelas cordas vocais,  $x_1(n)$ , e os efeitos da resposta impulsiva do trato vocal e da irradiação,  $x_2(n)$ , [23, 25, 34]. Formalmente,

$$x(n) = x_1(n) * x_2(n) = \sum_{k=-\infty}^{\infty} x_1(n) \cdot x_2(n - k). \quad (\text{B.1})$$

Deseja-se uma transformação,  $D_*[\cdot]$ , cuja entrada corresponda a uma operação de convolução e a saída a uma operação de adição. Isto pode ser obtido em duas partes, uma correspondendo à transformação da convolução em multiplicação, através da transformada  $\mathcal{Z}$ , e outra como uma transformação da multiplicação em adição, que pode ser obtida através do uso da função logaritma complexa, [23]. Formalmente tem-se

$$D_*[x_1(n) * x_2(n)] = \hat{x}_1(n) + \hat{x}_2(n) = \mathcal{Z}^{-1}[\log(\mathcal{Z}[x_1(n) * x_2(n)])]. \quad (\text{B.2})$$

Pode-se, ainda, definir a transformada inversa de  $D_*[\cdot]$  como

$$D_*^{-1}[x(n)] = \mathcal{Z}^{-1}[\exp(\mathcal{Z}[x(n)])]. \quad (\text{B.3})$$

Considerando-se os dois sinais,  $x_i(n)$ , convolvidos na entrada, teremos uma transformação  $D_*[\cdot]$  cuja saída será

$$\hat{x}(n) = \hat{x}_1(n) + \hat{x}_2(n) \quad (\text{B.4})$$

assim, através de um filtro linear no domínio do tempo, geralmente uma janela retangular, consegue-se atenuar os efeitos da excitação dada pelas cordas vocais isolando-se a resposta do trato vocal.

Para sinais de voz, os primeiros instantes de tempo (geralmente algo em torno de 4ms, [23, 34]) de  $\hat{x}(n)$  corresponderão quase que totalmente à resposta impulsiva do trato vocal [23, 25]. Assim um filtro do tipo

$$l(n) = \begin{cases} 1 & \text{se } |n| < \tau \\ 0 & \text{caso contrário} \end{cases} \quad (\text{B.5})$$

aplicado a  $\hat{x}(n)$  poderá isolar a componente  $x_2(n)$  do sinal de voz.

Geralmente o cálculo da  $D_*[\cdot]$  é simplificado para

$$c(n) = D_*[x(n)] = \mathcal{F}^{-1}[\log(|\mathcal{F}[x(n)]|)], \quad (\text{B.6})$$

onde  $\log$  é a função logarítmica real e  $\mathcal{F}[\cdot]$  é a transformada discreta de Fourier. Esta simplificação pode ser realizada se se considerar a seqüência  $x(n)$  de fase mínima<sup>1</sup> (ou fase máxima<sup>2</sup>) o que torna a seqüência  $\hat{x}(n)$  causal.

Após o cômputo de  $c(n) = D_*[x_1(n) * x_2(n)]$  geralmente obtém-se o espectro de frequência filtrado por  $l(n)$  através de  $\mathcal{F}[c(n) \cdot l(n)]$ . Este espectro corresponde a uma “suavização” do espectro obtido diretamente de  $x(n)$  devido a filtragem da componente de excitação dada pelas cordas vocais. Tanto o cepstrum,  $c(n)$ , quanto o espectro de frequência obtido através da seqüência  $c(n) \cdot l(n)$  são utilizados em sistemas de reconhecimento de voz clássico ou com redes neurais.

<sup>1</sup>Sistemas de fase mínima possuem pólos e zeros dentro do círculo unitário.

<sup>2</sup>Sistemas de fase máxima possuem pólos e zeros fora do círculo unitário.

## Apêndice C

# Transformada Discreta de Hartley

A transformada discreta de Hartley—TDH—[30] é definida para uma seqüência de números reais  $x(n)$  de comprimento  $N$ ,  $0 \leq n \leq N - 1$ , pelas seguintes equações:

$$H(k) = \sum_{n=0}^{N-1} x(n) \operatorname{cas} \left( \frac{2\pi}{N} kn \right), \quad 0 \leq k \leq N - 1$$
$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} H(k) \operatorname{cas} \left( \frac{2\pi}{N} kn \right), \quad 0 \leq n \leq N - 1$$

onde  $\operatorname{cas}(x) = \cos(x) + \operatorname{sen}(x)$ . A simetria deste par de transformadas é uma característica apreciável da TDH. Pode ser observado que esta transformada difere da transformada discreta de Fourier—TDF—somente pela presença do  $-j$  multiplicando o termo  $\operatorname{sen}$ , o que implica que a TDH é equivalente a simplesmente subtrair a parte imaginária da parte real da TDF, isto é,

$$\operatorname{TDH}[x(n)] = \Re\{\operatorname{TDF}[x(n)]\} - \Im\{\operatorname{TDF}[x(n)]\},$$

onde  $\Re\{\cdot\}$  é a parte real e  $\Im\{\cdot\}$  é a parte imaginária.

Sabendo-se que a parte real da TDF de uma seqüência real é par e a parte imaginária é ímpar, a TDF pode ser facilmente calculada a partir da TDH pelas equações:

$$\Re\{\operatorname{TDF}[x(n)]\} = \frac{1}{2} \{ \operatorname{TDH}[x(N - n)] + \operatorname{TDH}[x(n)] \}$$
$$\Im\{\operatorname{TDF}[x(n)]\} = \frac{1}{2} \{ \operatorname{TDH}[x(N - n)] - \operatorname{TDH}[x(n)] \}$$



onde  $x(N) = x(0)$ . As propriedades da TDH são facilmente provadas pelo uso das propriedades da TDF aplicadas a estas relações. Muitas propriedades são bastante similares aos correspondentes teoremas da TDF. Uma das propriedades mais úteis é o teorema do deslocamento no tempo

$$\text{TDH}[x(n + M)](k) = H(k) \cos\left(\frac{2\pi}{N}Mk\right) - H(N - k) \text{sen}\left(\frac{2\pi}{N}Mk\right).$$

A TDH possui um teorema para a propriedade da convolução, também, de acordo com a seguinte equação:

$$\begin{aligned} \text{TDH}[x_1(n) * x_2(n)](k) = \\ \frac{1}{2}[H_1(k)H_2(k) + H_1(k)H_2(N - k) + H_1(N - k)H_2(k) - H_1(N - k)H_2(N - k)]. \end{aligned}$$

Não é intuito deste trabalho entrar em detalhes da construção de algoritmos para a *transformada rápida de Hartley*—*TRH*—estes detalhes podem ser encontrados em [9] ou [30]. O algoritmo computacional, escrito em *FORTRAN*, para o cálculo da transformada com Radix-2 e dizimação no tempo, foi extraído diretamente de [30] e é transcrito a seguir.

## ALGORITMO PARA CÁLCULO DA TRH

```

CC-----CC
CC
CC      SUBROUTINE FHT2T
CC      -----
CC      Radix-2 decimation in time fast Hartley transform
CC
CC      Input   X      Sequence to be transformed
CC              N,M   Length of sequence N=2**M
CC      Output  X      Hartley transform
CC
CC      Authors: D.L. Jones and H.V. Sorensen
CC              Rice University, August 5, 1984
CC
CC-----CC
      SUBROUTINE FHT2T(X,N,M)
C
      REAL X(1)
C
C-----Digit reverse counter
C
100   J = 1
      N1 = N - 1
      DO 104 I = 1,N1
          IF (I .GE. J) GOTO 101
          XT = X(J)
          X(J) = X(I)
          X(I) = XT
101   K = N/2
102   IF (K .GE. J) GOTO 103
          J = J - K
          K = K/2
          GOTO 102
103   J = J + K
104   CONTINUE
C
C-----Main FHT loops
C
      DO 10 I = 1,N,2
          XT = X(I)
          X(I) = XT + X(I+1)
          X(I+1) = XT - X(I+1)

```

```
10    CONTINUE
C
      N2 = 1
      DO 20 K = 2,M
          N4 = N2
          N2 = N4 + N4
          N1 = N2 + N2
          E = 6.283185307179586/N1
C
      DO 30 J = 1,N,N1
          L2 = J + N2
          L3 = J + N4
          L4 = L2 + N4
          XT = X(J)
          X(J) = XT + X(L2)
          X(L2) = XT - X(L2)
          XT = X(L3)
          X(L3) = XT + X(L4)
          X(L4) = XT - X(L4)
          A = E
C
      DO 40 I = 2,N4
          L1 = J + I - 1
          L2 = J - I + 1 + N2
          L3 = L1 + N2
          L4 = L2 + N2
          CC1 = COS(A)
          SS1 = SIN(A)
          T1 = X(L3)*CC1 + X(L4)*SS1
          T2 = X(L3)*SS1 - X(L4)*CC1
          A = I * E
          XT = X(L1)
          X(L1) = XT + T1
          X(L3) = XT - T1
          XT = X(L2)
          X(L2) = XT + T2
          X(L3) = XT - T2
40    CONTINUE
30    CONTINUE
20    CONTINUE
C
      RETURN
      END
```